Factor Modelling for Tensor Time Series

Weilin Chen

Department of Statistics
The London School of Economics and Political Science

A thesis submitted for the degree of
Doctor of Philosophy

April 2024

To my loving family.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 3, 4, 5 and 6 were jointly co-authored with my supervisor, Professor Clifford Lam. A combined version of Chapter 3 and 4 has been published in *The Annals of Statistics*.

I declare that my thesis consists of 29816 words.

Weilin Chen
April 2024

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Clifford Lam, for his invaluable guidance, unwavering support, and immense encouragement throughout my PhD study. I am deeply grateful to have such a caring and supportive supervisor who not only guides me in academic expertise and research directions, but also provides valuable insights into research methodologies, presentation skills, career paths, and life itself. It was both extremely fortunate and a great privilege to have worked under his supervision and to learn from him. I would also like to express my gratitude to my second supervisor, Professor Wicher Bergsma, for his encouragement and guidance in my pursuit of PhD study.

I extend my profound appreciation to all the staff and colleagues in the Department of Statistics at the London School of Economics and Political Science for fostering my wonderful PhD journey over the last three years. My research would not have been possible without the sponsorship, resources, and facilities provided by the department. Special thanks to Penny Montague, Muhammed Sabih Iqbal, and Imelda Noble for their outstanding administrative work and assistance. Additionally, I am grateful to Dr. Chengchun Shi for offering me excellent teaching opportunities. My sincere thanks go to my PhD cohorts and friends: Sixing Hao, August Shen, Pingfan Su, Zetai Cen, Tao Ma, Xinyi Liu, Xuzhi Yang, and many others. Thank you all for your continuous support and companionship during these years.

Finally, I would like to express my heartfelt gratitude to my family for their enduring love and steadfast support over the years. To my parents, Yumin Chen and Gao Lin, and my grandparents, whose everlasting love has not only sustained me but also instilled in me the values of goodness and morality. My deepest appreciation goes to my beloved family members, Wenjuan Huang, Zhi Chen, and Yi Chen, for their unwavering love, support, and encouragement. They have consistently been my source of strength, both in the past and undoubtedly will continue to be in the future.

Abstract

High dimensional tensor time series data is increasingly prevalent across various fields. In the analysis of such data, factor modelling plays a crucial role as a dimension reduction tool. While traditional factor models primarily handle vector time series, the exploration of matrix or tensor factor models under various assumptions is still in its early stages and has attracted increasing interest in recent years.

In this thesis, we develop a tensor factor model under the presence of both serial and cross-correlations in the idiosyncratic components, assuming only bounded fourth order moments for the time series variables. Moreover, we incorporate a spectrum of different factor strengths into the model, in contrast to the prevalent assumption in many literature that considers only pervasive factors. The inclusion of serial dependence noise and weak factors makes our model more compatible with real data, especially in economics and finance.

With the relaxed assumptions in our model, we propose a pre-averaging procedure to initially estimate the factor loading spaces, which achieves signal accumulation through the random projection of tensor fibres. Furthermore, we develop an iterative projection algorithm to improve the re-estimation of factor loadings by projecting the data onto the strongest estimated factor directions. To estimate the number of factors, we introduce a new core tensor rank estimation method through correlation analysis on the projected data. Theoretical guarantees are provided for all estimators, and extensive simulations, as well as analyses of real datasets, are conducted to compare our methods with other state-of-the-art or traditional alternatives. Finally, we present a new method for estimating factor strengths with empirical results provided and introduce a novel matrix convergence criterion for specific covariance matrix estimators, offering valuable insights into directions for future research.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Thanks to the advancement of the internet and general computing power, the collection and analysis of massive data have become increasingly easier over the past few decades. The richness of data also means that, more often than not, the data we obtain are high dimensional in nature, in the sense that the number of variables is comparable to or even larger than the sample size. To extract useful information from high dimensional time series data, factor modelling is a major dimension reduction tool that allows insights into the common dynamics of different observed time series (Bai, 2003; Bai and Ng, 2002; Chamberlain and Rothschild, 1983; Fan et al., 2013; Forni et al., 2000; Lam and Yao, 2012; Lam et al., 2011; Stock and Watson, 2002). In practice, a few common factors can often capture a large amount of variation and dynamics among a large pool of variables and time series. For instance, when considering many macroeconomic time series for forecasting (Stock and Watson, 2002), the estimation and forecasting through the common factors can give more accurate results overall, and allowing for the interpretation of the factors (e.g., potential grouping of macroeconomic time series as factors) at the same time.

While traditional factor models only deal with time series in vector form, there is a growing trend of encountering time series observed in matrix or tensor forms across various fields, including economics, finance, engineering, and others. Examples of such data formats include Fama-French portfolio return series based on different sizes and operating profitabilities (Wang et al., 2019), multi-category international trading volume series (Chen and Chen, 2022), a collection of economic indicator series across various countries (Chen et al., 2021), taxi traffic transport data (Chen et al., 2022), sequences of gray-scale face recognition images (Chen and Fan, 2021), and neuroimaging data (Zhang, 2019).

To achieve dimension reduction for high dimensional matrix-valued or tensor-valued time series, although it's possible to stack all observed time series into a single vector time series for factor modelling analysis, the problem in doing this is that we are ignoring the natural structure of the data. Moreover, stacking all time series into a long vector can lead to the curse of dimensionality (e.g., when the stacked length is much larger than the sample size), increasing the number of parameters to estimate and, consequently, resulting in inaccurate estimation and predictions.

A more natural and effective approach is to consider factor modelling on matrix-valued or tensor-valued time series directly. Wang et al. (2019) describe a factor model for matrix-valued time series and provide estimation methods along with theoretical results. Their work is extended to a general order-K tensor time series in Chen et al. (2022), Han et al. (2020), and Han et al. (2022), where the tensor factor model structure is constructed in a form similar to Tucker tensor decomposition (Tucker, 1963, 1964, 1966), and iterative approaches to estimate the factor loadings and the number of factors are proposed. Recently, a growing body of research has emerged in this field, introducing various factor model assumptions and estimation methods for the analysis of matrix-valued or tensor-valued time series (Barigozzi et al., 2023a,b; Chen and Fan, 2021; He et al., 2023b, 2022; Yu et al., 2022). Nevertheless, despite these efforts, there are still limitations and challenges in these models that need to be addressed. For example, Chen et al. (2022); Han et al. (2020, 2022); Wang et al. (2019) assume independent noise series, while Barigozzi et al. (2023a,b); Chen and Fan (2021); He et al. (2023b, 2022); Yu et al. (2022) only deal with pervasive factors, which can be restrictive in many real-world applications, especially in economics and finance, where data are usually autocorrelated, heavy-tailed, with weak or "local" factors presence (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986; Uematsu and Yamagata, 2022).

In this thesis, we make several important contributions to the field of factor modelling for high dimensional tensor time series data of order two or above. The first one is to propose a model that allows for a spectrum of different factor strengths, which is a generalisation to Lam et al. (2011) when static vector time series factor model is concerned. To the best of our knowledge, our model is the first one in tensor factor modelling to consider weak factors when both serial and cross-correlations in the noise series are presented, which can be more flexible for wider applications. For instance, when analyzing economic and financial data, empirical studies indicate the common occurrence of weak factors (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986), which makes the classical methods based solely on pervasive factors less effective. In such scenarios, our model

provides more appropriate estimation tools by more effectively capturing signals in the presence of weak factors, as introduced in the next paragraph.

With relaxed assumptions of weak factors in the model, our second contribution is to provide a "pre-averaging" initial estimator and an iterative projection estimator for the factor loading matrices, with theoretical analyzes provided and rate of convergence spelt out. The pre-averaging procedure can be seen as a random projection method by randomly summing tensor fibres, and we provide a method to control for the quality of the random projection. Iterative projection estimators are introduced with idea similar to the projection method in Yu et al. (2022), except that we only project on the direction aligning to the strongest estimated factor, unlike other literature (Barigozzi et al., 2023b; He et al., 2022; Yu et al., 2022) which uses the entire estimated factor loading matrix. With weak factors in the model, numerical experiments show that our estimator outperforms methods for estimating matrix or tensor factor models under the sole assumption of pervasive factors. This is because we only utilize the information that captures the most accurate estimations so far, while projections including estimated weaker factors lead to worse performance.

Our third contribution is to develop an estimator of the core tensor rank through correlation analysis. Core tensor rank is similar to the number of factors, and will be explained in Chapter 2. While many literature estimates the number of factors by examining the eigenvalues of the sample covariance matrix or its variations (Barigozzi et al., 2023b; Chen and Fan, 2021; Han et al., 2022; He et al., 2023a, 2022; Yu et al., 2022), our method utilizes correlation information on the projected data, which is inspired by the correlation thresholding technique introduced by Fan et al. (2022) for the vector factor model, and we further introduce a bootstrap method for tuning parameter selection. We provide theoretical analysis of consistency, and empirical experiments demonstrate that our method can effectively detect weak factors when present.

In addition to estimating the factor loading spaces and the number of factors, which are the two most commonly studied aspects of factor modelling, we also contribute to factor strength estimations—a new and challenging topic under the assumption of weak factors. While literature on factor strength estimation is limited, existing studies (Bailey et al., 2021; Uematsu and Yamagata, 2022) predominantly rely on the sparsity assumption of the factor loading matrices, focusing solely on vector factor models. In contrast, we introduce a novel approach to estimate factor strengths by leveraging covariance information, which proves effective in more general settings where factor loading matrices are not necessarily sparse. We also extend our method to be applicable in matrix factor models. The performance of

our approach is demonstrated through numerical experiments, providing valuable insights for future research in this direction.

Finally, our last contribution is in covariance matrix estimations. In high dimensional settings, it is well-known that the sample covariance matrix is ill-conditioned, and various regularization methods have been studied to improve the performance of the estimators. Among them, eigenvalue-shrinkage estimators, such as the Nonparametric Eigenvalue-Regularized Cvariance Matrix Estimator (NERCOME) proposed by Lam (2016), demonstrate good empirical performance but lack theoretical guarantees regarding consistency. We propose a new type of matrix convergence, called "normalized consistency", and prove it for certain covariance and precision matrix estimators, providing more theoretical support for these estimators with potential applications. It is also worth mentioning that the estimation of factor strengths and covariance matrices can be related and useful for future research on inference problems in factor modelling.

The rest of this thesis is organized as follows. Chapter 2 reviews some basic knowledge on tensor operations and factor modelling in general. Chapter 3 introduces important assumptions on our tensor factor model and presents the pre-averaging and iterative projection algorithm to estimate the factor loading spaces. Subsequently, Chapter 4 presents the method to find the rank of the core tensor through correlation thresholding. Chapter 5 discusses a novel method for factor strength estimations. Finally, Chapter 6 investigates a new form of consistency of covariance matrix estimators.

Chapter 2

Review on Tensor Operations and Factor Models

2.1 Notations and Basic Tensor Manipulations

In this thesis, we use $a \asymp b$ to denote $a = O(b)$ and $b = O(a)$ (also $a \asymp_P b$ for $a = O_P(b)$ and $b = O_P(a)$), while $a \succeq b$ is equivalent to $b = O(a)$, and $a \succ b$ is equivalent to $b = o(a)$. We also use $\|\cdot\|$ to denote the L_2 norm (of a vector or a matrix), and $\|\cdot\|_F$ to denote the Frobenius norm, while $\|\cdot\|_{\max}$ represents the maximum element (of a vector or a matrix). We also use $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$ and $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ to denote the L_∞ and L_1 norm of a matrix $\mathbf{A}$ respectively. The notation $\text{vec}(\cdot)$ represents the vectorisation of a matrix, stacking columns of the matrix from left to right. We use $\mathbf{1}_m$ to represent a vector of ones with length m, $\mathbf{1}_{m,S}$ to represent a vector of ones and zeros with length m, with ones on positions belonging to the set S and zeros otherwise. The identity matrix with size m is denoted by $\mathbf{I}_m$. The notation $\text{diag}(\mathbf{A})$ of a square matrix $\mathbf{A}$ is the diagonal matrix with only the diagonal elements of $\mathbf{A}$ remain, and everything else set to 0. This notation is also used to represent a block diagonal matrix. For instance, $\text{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ is the block diagonal matrix with diagonal block matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$. We use $\lambda_j(\mathbf{A})$ to denote the j-th largest eigenvalue of a square matrix $\mathbf{A}$, and $\text{tr}(\mathbf{A})$ the trace of $\mathbf{A}$. For a positive integer m, we define $[m] := \{1, \ldots, m\}$. The cardinality of a set S is denoted by $|S|$.

We briefly introduce the notations and review on tensor manipulations in this section just enough for the presentation of this thesis. For more information, please refer to Kolda and Bader (2009).

A tensor is a multidimensional array, a generalization of a matrix. Let $\mathscr{X} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be an order-K tensor. Here K represents the number of dimensions in $\mathscr{X}$, also called the number of *modes*. For instance, a vector has $K = 1$ while a matrix has $K = 2$. If we write $\mathscr{X} = (x_{i_1 \cdots i_K})$, then we define a *mode-k fibre* of $\mathscr{X}$ to be a column vector (of length d_k)

$$(x_{i_1 \cdots i_{k-1}, j, i_{k+1} \cdots i_K})_{j \in [d_k]}, \quad i_\ell \in [d_\ell] \text{ with } \ell \in [K],$$

where we define, for any positive integer n, $[n] := \{1, \ldots, n\}$. Hence there are in total $d_{-k} := \prod_{\ell=1; \ell \neq k}^{K} d_\ell$ number of mode-k fibres for the tensor $\mathscr{X}$. For example, for an order-1 tensor (a vector) $\mathbf{x}$, the mode-1 fibre is $\mathbf{x}$ itself; for an order-2 tensor (a matrix) $\mathbf{X}$, the mode-1 fibres are its columns, and the mode-2 fibres are its rows.

The *mode-k matricization/unfolding matrix* $\mathrm{mat}_k(\mathscr{X}) \in \mathbb{R}^{d_k \times d_{-k}}$ (also denoted as $\mathbf{X}_{(k)}$ sometimes) is then defined to be the matrix containing (in order) all the mode-k fibres of $\mathscr{X}$. For instance, for an order-1 tensor (a vector) $\mathbf{x}$, $\mathrm{mat}_1(\mathbf{x}) = \mathbf{x}$; for an order-2 tensor (a matrix) $\mathbf{X}$, $\mathrm{mat}_1(\mathbf{X}) = \mathbf{X}$ and $\mathrm{mat}_2(\mathbf{X}) = \mathbf{X}^{\mathrm{T}}$. See Figure 2.1 for a demonstration of a mode-3 tensor (figure from Tao et al. (2019), where I is the same as our d).

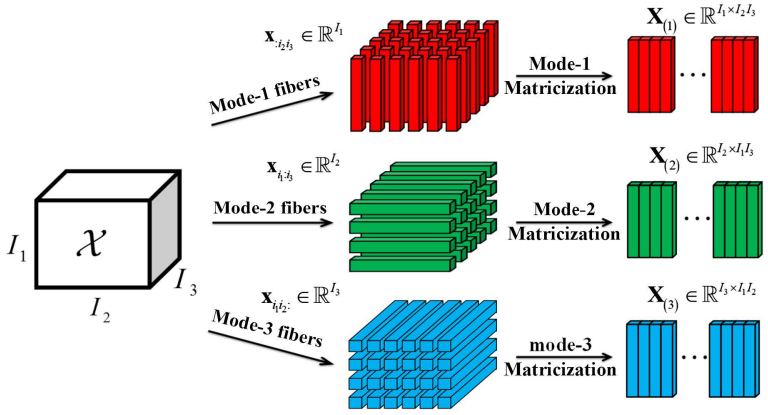


Fig. 2.1 Illustration of the mode-k fibers and its corresponding unfolding matrix.

If there is a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d_k \times r_k}$, $k \in [K]$, and $\mathscr{F} = (f_{i_1 \cdots i_K}) \in \mathbb{R}^{r_1 \times \cdots r_K}$ is an order-K tensor, then the *k-mode product* of $\mathscr{F}$ and $\mathbf{A}$, denoted by $\mathscr{F} \times_k \mathbf{A} \in \mathbb{R}^{r_1 \times \cdots \times r_{k-1} \times d_k \times r_{k+1} \times \cdots \times r_K}$, is defined as

$$(\mathscr{F} \times_k \mathbf{A})_{i_1 \cdots i_{k-1}, j, i_{k+1} \cdots i_K} := \sum_{i_k=1}^{r_k} f_{i_1 \cdots i_k \cdots i_K} a_{j i_k},$$

such that

$$\mathrm{mat}_k(\mathscr{F} \times_k \mathbf{A}) = \mathbf{A} \mathrm{mat}_k(\mathscr{F}).$$

As an example, consider $K = 2$, so that $\mathcal{F}$ is a matrix. Let $\mathbf{A}_1 \in \mathbb{R}^{d_1 \times r_1}$ and $\mathbf{A}_2 \in \mathbb{R}^{d_2 \times r_2}$. Then the mode-1 fibres of $\mathcal{F}$ are in fact the columns of $\mathcal{F}$, while the mode-2 fibres of $\mathcal{F}$ are the rows of $\mathcal{F}$ (made column vectors). We also have

$$\mathcal{F} \times_1 \mathbf{A}_1 = \mathbf{A}_1 \mathcal{F}, \quad \mathcal{F} \times_2 \mathbf{A}_2 = \mathcal{F} \mathbf{A}_2^{\mathrm{T}} = (\mathbf{A}_2 \mathcal{F}^{\mathrm{T}})^{\mathrm{T}}.$$

Note that the two unfolded matrices (mode-1 and mode-2 respectively) are $\mathrm{mat}_1(\mathcal{F}) = \mathcal{F}$ and $\mathrm{mat}_2(\mathcal{F}) = \mathcal{F}^{\mathrm{T}}$, hence the above also means that

$$\mathcal{F} \times_1 \mathbf{A}_1 = \mathbf{A}_1 \mathrm{mat}_1(\mathcal{F}), \quad \mathcal{F} \times_2 \mathbf{A}_2 = (\mathbf{A}_2 \mathrm{mat}_2(\mathcal{F}))^{\mathrm{T}}.$$

In general, to calculate $\mathcal{F} \times_k \mathbf{A}$, we find the mode-$k$ unfolding matrix $\mathrm{mat}_k(\mathcal{F})$ first and then calculate $\mathbf{A}\mathrm{mat}_k(\mathcal{F})$, which contains all mode-k fibres of $\mathcal{F}$ being pre-multiplied by $\mathbf{A}$. Then we put the columns in $\mathbf{A}\mathrm{mat}_k(\mathcal{F})$ back into the original shape of the tensor $\mathcal{F}$. Hence for $K = 2$, to put the columns in $\mathbf{A}\mathrm{mat}_2(\mathcal{F})$ back into the original orientation of the tensor (rows), we take transpose of it, so that $\mathcal{F} \times_2 \mathbf{A} = (\mathbf{A}\mathrm{mat}_2(\mathcal{F}))^{\mathrm{T}}$. The order of distinct mode products does not matter, in the sense that for $i \neq j$,

$$\mathcal{F} \times_i \mathbf{A}_i \times_j \mathbf{A}_j = \mathcal{F} \times_j \mathbf{A}_j \times_i \mathbf{A}_i.$$

Tucker decomposition (Tucker, 1963, 1964, 1966) is a major extension of the matrix singular value decomposition (SVD) to tensors of higher order. Recall that the SVD of a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ can be written as $\mathbf{X} = \mathbf{U}_1 \Lambda \mathbf{U}_2^{\mathrm{T}}$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthonormal matrices of size $d_1 \times r$ and $d_2 \times r$ spanning the column and row spaces of $\mathbf{X}$ respectively, and Λ is an $r \times r$ diagonal matrix with r positive singular values on its diagonal. In parallel, Tucker decomposition decomposes an order-K tensor $\mathcal{X}$ into K orthonormal matrices $\mathbf{U}_k \in \mathbb{R}^{d_k \times r_k}$ containing basis vectors spanning k-mode fibers of the tensor, a potentially much smaller 'core' tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ and the relationship

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_K \mathbf{U}_K. \tag{2.1}$$

See Figure 2.2 for a demonstration. Note that the core tensor $\mathcal{G}$ is similar to the Λ in the middle of matrix SVD, but now it is not necessarily diagonal.

Finally, if $\mathcal{C} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K$, then we have the formula

$$\mathrm{mat}_k(\mathcal{C}) = \mathbf{A}_k \mathrm{mat}_k(\mathcal{F}) \mathbf{A}_{-k}^{\mathrm{T}}, \tag{2.2}$$

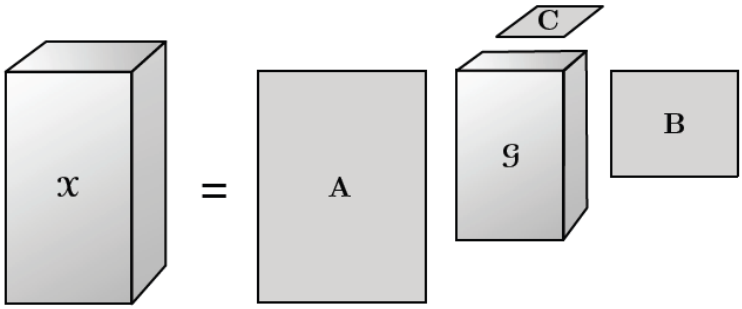


Fig. 2.2 Tucker decomposition of an order-3 tensor.

where $\otimes$ is the Kronecker product, $\mathbf{A}_{-k} := \mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_{k+1} \otimes \mathbf{A}_{k-1} \otimes \cdots \otimes \mathbf{A}_1$, and $\text{mat}_k(\mathscr{C})$ is the mode-k unfolding of $\mathscr{C}$.

2.2 Factor Models

2.2.1 Vector Factor Models

In this section, we briefly review the factor model approach to panel time series data. The earliest research on factor analysis can be traced back more than a century (Spearman, 1904). Over the past few decades, vector factor models have been extensively studied in the statistics and economics literatures. Let $\mathbf{x}_t \in \mathbb{R}^d$, $t \in [T]$ be a vector time series. Factor model assumes

$$\mathbf{x}_t = \mathbf{c}_t + \mathbf{e}_t = \mathbf{A}\mathbf{f}_t + \mathbf{e}_t, \quad t \in [T], \tag{2.3}$$

where $\mathbf{x}_t$ can be decomposed into a signal part $\mathbf{c}_t$ and a noise part $\mathbf{e}_t$, which is the idiosyncratic noise series with a mean of 0. The signal part $\mathbf{c}_t$ can be further decomposed into $\mathbf{A}\mathbf{f}_t$. Here, $\mathbf{f}_t \in \mathbb{R}^r$, $t \in [T]$, represents a set of unobserved latent factor time series with dimension $r \leq d$, and r is the number of factors. The matrix $\mathbf{A} \in \mathbb{R}^{d \times r}$ is an unknown constant factor loading matrix. Note that the decomposition (2.3) always holds. However, it is only useful when $r \ll d$, as then the dimension reduction is achieved in the sense that the dynamics of $\mathbf{x}_t$ is driven by a much lower dimensional process $\mathbf{f}_t$.

The advantage of the factor model is that it achieves a significant reduction in model complexity, quantified by the number of parameters. This reduction occurs because the covariance or autocovariance matrices are now determined by the factor loading matrix $\mathbf{A}$ and the much smaller covariance or autocovariance matrix of the factor process $\mathbf{f}_t$.

Additionally, the hidden dynamics, i.e., the co-movements, become transparent, providing a clearer and more insightful understanding. This transparency is especially valuable when the co-movement of the time series is complex and challenging to uncover without proper modelling of the full panel. Factor models contribute to increasing the explanatory and predictive power of various time series models. For instance, macroeconomic time series can be modeled using common factors, which not only yield more accurate forecasting results by extracting signals from the dynamics of factors, but also facilitate easier interpretability of the factors in real-world applications. (Forni et al., 2003; Stock and Watson, 2002). Other examples include revealing anomalous empirical findings in monetary policy (Forni and Gambetti, 2010) and assisting in covariance matrix estimation for portfolio optimization (Fan et al., 2013), among many others.

In the factor model (2.3), both $\mathbf{A}$ and $\mathbf{f}_t$ are unobserved, and the number of factors, r, is typically unknown. The focus of factor model studies usually lies in estimating (i) the factor loading space of $\mathbf{A}$ and (ii) determining the number of factors, r. It is important to note that the model (2.3) remains unchanged if we replace the pair $(\mathbf{A};\mathbf{f}_t)$ on the right-hand side with $(\mathbf{AH};\mathbf{H}^{-1}\mathbf{f}_t)$ for any invertible $\mathbf{H}$. However, the linear space spanned by the columns of $\mathbf{A}$, referred to as the factor loading space, is uniquely defined. Therefore, the actual challenge lies in estimating the factor loading space of $\mathbf{A}$ or, equivalently, determining $\mathbf{A}$ up to rotations.

In the model (2.3), if there is no temporal dependence in either the factors $\mathbf{f}_t$ or the idiosyncratic component $\mathbf{e}_t$, and no cross-correlation among $\mathbf{e}_t$, then it is called an *Exact Static Factor Model* (Spearman, 1904). Estimation of such model can be conducted using principal components analysis (PCA) (Tipping and Bishop, 1999) or maximum likelihood (Lawley and Maxwell, 1962). However, the overly restrictive assumption of such model limits its application to modern economic and financical data. When considering potential temporal and cross-sectional dependence of the idiosyncratic components, additional and more relaxed model assumptions are necessary. For the purpose of estimation and inference based on (2.3), the literature commonly adopts two different types of model assumptions for these purposes.

One type of model assumes that a common factor must have an impact on 'most' (defined asymptotically) of the time series but allows the idiosyncratic noise to have weak cross-correlations and weak autocorrelations (see Bai (2003); Bai and Ng (2002, 2007, 2023); Bartholomew et al. (2011); Chamberlain and Rothschild (1983); Fan et al. (2013, 2019); Forni et al. (2000); Hallin and Liška (2007); Stock and Watson (2005, 2002) for examples). In other words, the effect of a factor aggregates along the cross-sectional

dimension, thus leaving potentially serial dependence in the idiosyncratic components. This is known as the "*Approximate Factor Model*" by Bai and Ng (2002), where the rigorous definition of the common factors and the idiosyncratic noise can only be established asymptotically when d (i.e., the number of component series) goes to infinity. In many real datasets, such as those in finance, economics, genetics, and imaging, it is common for (weak) serial correlations to exist in $\mathbf{e}_t$, representing any serial correlations in $\mathbf{x}_t$ not captured by the common components (Chen and Fan, 2021; Stock and Watson, 2002). Thus, an idiosyncratic noise series is not necessarily white noise, and each idiosyncratic noise component may, at most, affect the dynamics of a few original time series. Under these assumptions, principal component analysis (PCA) of the sample covariance matrix is typically used to estimate the factor loading space, with various extensions employed (Bai, 2003; Bai and Ng, 2002). The number of factors can be estimated by utilizing the behavior of the eigenvalues of the sample covariance matrix, such as maximizing the ratio of consecutive eigenvalues (Ahn and Horenstein, 2013) or separating diverging eigenvalues from the rest using threshold functions in the form of an information criterion (Bai and Ng, 2002).

Another type of model assumes that the common factors $\mathbf{f}_t$ account for all dynamics of the series $\mathbf{x}_t$, making the idiosyncratic noise $\mathbf{e}_t$ 'white' with no autocorrelation but allowing substantial contemporary cross-correlation among the error processes (see Lam and Yao (2012); Lam et al. (2011); Pan and Yao (2008) for examples). In such a model, the d-dimensional time series is decomposed into two parts: the dynamic part driven by r factors and the static part, which is a vector of white noise. The assumption of independence and identically distributed (i.i.d.) elements in $\mathbf{e}_t$ is considered a standard statistical analysis assumption. Under these sets of assumptions, the estimation of the factor loading space is done through eigenanalysis based on the non-zero lag autocovariance matrices (Lam et al., 2011). The number of factors can be estimated by studying the ratio of eigenvalues of the corresponding sample autocovariance matrices (Lam and Yao, 2012).

It is also worth noting that if there is no temporal dependence in the model (2.3), then the two types of models discussed above coincide. Additionally, all models discussed so far assume static loadings, which are an extension of the exact static factor model. To accommodate broader assumptions, Forni et al. (2000) and Forni and Lippi (2001) proposed the *Generalized Dynamic Factor Model*, which incorporates dynamic loadings to capture the lagged impacts of the common factors. The assumptions of the dynamic factor model are comprehensive and encompass all previously discussed models. However, its more general assumptions also make it more challenging to estimate, particularly when

extending to matrix or tensor settings. As a result, for the rest of this thesis, we focus solely on the static factor model. For a more detailed survey of dynamic factor models, please refer to Barigozzi and Hallin (2024).

Another important assumption in factor models is factor strength. In standard factor models, such as those proposed by Bai (2003); Bai and Ng (2002); Stock and Watson (2002), it is typically assumed that all r factors are strong, also referred to as *pervasive*. This assumption implies that all r eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ diverge proportionally to d, specifically $\lambda_j(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \asymp d$ for $j \in [r]$. This results in a clear partition of the eigenvalues of the observed covariance matrix into two sets: large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation.

However, such a phenomenon is frequently not observed in practice. Empirical studies in economics and finance indicate that eigenvalues often diverge at varying rates (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986). To address this, models introducing 'weak factors' have been proposed. For the j-th column $\mathbf{a}_j$ of $\mathbf{A}$, its factor strength α_j, with a range of $[0,1]$, is defined such that

$$\|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}, \quad j \in [K], \tag{2.4}$$

ensuring that

$$\lambda_j(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \asymp d^{\alpha_j}, \quad j \in [K]. \tag{2.5}$$

A strong (or pervasive) factor has $\alpha_j = 1$, while a weak factor has $\alpha_j < 1$. A weak factor can result from two scenarios: (i) the factor has a weak effect on some or all observables, or (ii) it affects only a subset of observables, referred to as a 'local' factor by Freyaldenhoven (2022). Hallin and Liška (2011) discussed the local factor structure due to the presence of blocks. The literature has developed studies focusing on the estimation of the factor loading space and the number of factors when weak factors are present in the model (Bai and Ng, 2023; Freyaldenhoven, 2022; Lam et al., 2011; Onatski, 2012; Uematsu and Yamagata, 2022). These studies demonstrate that weaker factors are generally harder to detect, and the estimation accuracy of factor loadings could be influenced by the strengths of the factors as well. Thus, estimation and forecasting performances may suffer in the presence of weak factors compared to the classical setting where all factors are pervasive.

There is limited research on estimating factor strengths. With sparsity assumptions applied to the factor loadings, Uematsu and Yamagata (2022) employ techniques akin to adaptive LASSO to recover the local factors and estimate their strengths, while Bailey

et al. (2021) propose estimating factor strengths based on the proportion of statistically significant factor loadings for observed factors. Nevertheless, the sparsity assumption in these literature specifically address scenario (ii) mentioned earlier, i.e., when factors are weak due to being "local". On the other hand, if a factor is weak because it has a 'weak' impact on some or all observables, then it may not be detectable by sparsity assumptions. Connor and Korajczyk (2022) consider such a structure with observed factors, and demonstrate its application when modelling U.S. equity return series. It remains an open topic to estimate the factor strengths in such scenario without sparsity assumption.

2.2.2 Extensions to Matrix and Tensor Factors

The advancement in data collection capabilities has given rise to an extensive volume of time series. The observation of variables organized in a well-defined matrix or tensor structure over time has become prevalent across diverse fields (Chen and Chen, 2022; Chen and Fan, 2021; Chen et al., 2021, 2022; Wang et al., 2019; Zhang, 2019). For example, when forecasting many macroeconomic time series from different countries (Stock and Watson, 2002), a natural approach is to consider the country-categorized macroeconomic time series as matrix-valued (i.e., an order-2 tensor), with different countries represented by rows and various macroeconomic time series represented by columns.

While considerable efforts have been directed towards the development of methodologies and theories for vector factor models in the past few decades, literature addressing matrix-valued or tensor-valued time series has been lacking until very recently. Wang et al. (2019) for the first time describes a matrix factor model for matrix-valued time series $\mathbf{X}_t \in \mathbb{R}^{d_1 \times d_2}$, which takes the form:

$$\mathbf{X}_t = \mathbf{C}_t + \mathbf{E}_t = \mathbf{A}_1 \mathbf{F}_t \mathbf{A}_2^{\mathrm{T}} + \mathbf{E}_t, \quad t \in [T], \tag{2.6}$$

where the signal $\mathbf{C}_t$ and error $\mathbf{E}_t$ are defined similarly to $\mathbf{c}_t$ and $\mathbf{e}_t$ in the vector factor model (2.3), except that they are now matrices of dimension $d_1 \times d_2$. The signal $\mathbf{C}_t$ is then further decomposed into the matrix-valued common factor $\mathbf{F}_t \in \mathbb{R}^{r_1 \times r_2}$ and two factor loading matrices $\mathbf{A}_1 \in \mathbb{R}^{d_1 \times r_1}$ and $\mathbf{A}_2 \in \mathbb{R}^{d_2 \times r_2}$. Similar to vector factor models, it is assumed that the common factors $\mathbf{F}_t$ drive all dynamics and co-movement of the matrix-valued time series $\mathbf{X}_t$. The two loading matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ capture the dependency between each individual time series in the matrix observations $\mathbf{X}_t$ and the matrix factors $\mathbf{F}_t$. More specifically, $\mathbf{A}_2$ reflects how each column of $\mathbf{X}_t$ depends on the columns of $\mathbf{F}_t$ and is thus called the column loading matrix. Similarly, $\mathbf{A}_1$, the row loading matrix, reflects how each

row of $\mathbf{X}_t$ depends on the rows of $\mathbf{F}_t$. Wang et al. (2019) provides estimation methods together with theoretical results for the model (2.6), based on the extension of assumptions proposed by Lam et al. (2011) such that $\mathbf{E}_t$ is white noise.

The work of Wang et al. (2019) is extended to a general order-K tensor $\mathcal{X}_t$ in Chen et al. (2022), where the factor model for each $\mathcal{X}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is

$$\mathcal{X}_t = \mathcal{C}_t + \mathcal{E}_t = \mathcal{F}_t \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K + \mathcal{E}_t, \quad t \in [T], \tag{2.7}$$

with $\mathcal{C}_t$ as the common component, $\mathcal{E}_t$ as the noise tensor, $\mathcal{F}_t \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ as the core tensor, and $\mathbf{A}_k \in \mathbb{R}^{d_k \times r_k}$ as the mode-k factor loading matrix. The product $\times_k$ is the tensor k-mode product. Tensor time series of order 3 or higher are becoming more prevalent in various applications. For instance, consider international import-export volume time series of different categories of products among countries (Chen et al., 2022). These data can be organized into an order-3 tensor time series, where the first two modes represent import-export patterns between countries, and the third mode represents different categories of products. Other examples include taxi traffic transport data collected with different pick-up, drop-off locations (Chen et al., 2022), and neuroimaging 3-dimensional data (Zhang, 2019).

Note that the decomposition of the common component $\mathcal{C}_t$ is based on the Tucker decomposition (2.1), which is general and can be applied to tensors of any order-K. Specifically, when $K = 2$, the model (2.7) reduces to the matrix factor model (2.6), and when $K = 1$, the model (2.7) reduces to the traditional vector factor model (2.3). Similar to the vector factor models, the Tucker tensor factor model (2.7) can only be identified up to rotations. For tensor factor modelling, in addition to Tucker, another possible decomposition is the CP-decomposition (Chang et al., 2023). CP decomposes a tensor into a sum of component rank-one tensors, which is uniquely defined but harder to estimate. One reason is that a CP representation is not guaranteed to exist for all K, which leads to possible numerical difficulties. In contrast, a Tucker decomposition always exists but is not unique and can be rotated. In this paper, we choose to use the Tucker decomposition for its flexibility. In fact, CP can be viewed as a special case of Tucker where the core tensor is superdiagonal and $r_1 = r_2 = \cdots r_k$. Please refer to Kolda and Bader (2009) and Lettau (2022) for more discussions and comparisons between Tucker and CP.

Similar to the vector factor model (2.3) and the matrix factor model (2.6), in the tensor factor model (2.7), the goal is to achieve two objectives: (i) estimate the factor loading space of $\mathbf{A}_k$ for each $k \in [K]$, and (ii) determine the number of factors, r_k, $k \in [K]$, also

known as the rank of the core tensor. To achieve this, Chen et al. (2022) assume that the elements in each $\mathcal{E}_t$ are sub-Gaussian, with each $\mathcal{E}_t$ independent of each other, as an extension of the assumptions from Lam et al. (2011) and Wang et al. (2019). Based on these assumptions, Chen et al. (2022) propose two approaches, named TIPUP (Time series Inner-Product Unfolding Procedure) and TOPUP (Time series Outer-Product Unfolding Procedure), to estimate the factor loading spaces of $\mathbf{A}_k$. These approaches involve a combination of tensor unfolding and the use of lagged cross-product, which is the tensor version of autocovariance. Han et al. (2020) further analyze the iterative projection procedures iTOPUP and iTIPUP, providing improved rates for re-estimating $\mathbf{A}_k$, while Han et al. (2022) propose core rank estimators of $\mathcal{C}_t$ based on information criterion and eigenvalue ratio criterion that are intertwined with iTIPUP and iTOPUP.

In other recent developments, Zhang and Xia (2018) proposes a similar model for an order-3 tensor, with the tensor noise elements being i.i.d. normal having a common variance, and develops minimax theoretical guarantees for their estimators. With the same tensor noise assumption, Yokota et al. (2017) proposes a core rank estimator for $\mathcal{C}_t$ for a general order-K tensor $\mathcal{X}_t$ based on a BIC-like criterion, while Liu et al. (2022) proposes a tensor SVD method for estimation under a CP decomposition of $\mathcal{C}_t$. Chen et al. (2020) proposes a semiparametric model with $\mathcal{C}_t$ taking covariates under the assumption of i.i.d. sub-Gaussian elements in $\mathcal{E}_t$, which are themselves independent of each other.

All the tensor factor modelling works mentioned above assume at least independent noise tensor series $\mathcal{E}_t$ with sub-Gaussian elements. The i.i.d. assumption for the elements in $\mathcal{E}_t$ in many of them is an extension of the assumptions of Lam et al. (2011) from the vector factor model. This assumption is also considered standard for statistical analysis. However, if we have applications in economics and finance for instance, it is very easy that (weak) serial correlations exist in $\{\mathcal{E}_t\}$, representing any serial correlations in $\mathcal{X}_t$ not captured by the common components $\mathcal{C}_t$ (some time series in $\mathcal{X}_t$ have "unique" company or macroeconomic characteristics, for example). As introduced in Section 2.2.1, the approximate factor model of Bai and Ng (2002) allows for such weak serial correlations (as well as weak cross-correlations) in the idiosyncratic noise series $\{\mathcal{E}_t\}$. When $\mathcal{E}_t$ has a higher order tensor structure, allowing for weak-serial and cross-correlations becomes even more essential as there could be even more potentially intricate serial and cross-correlations in $\{\mathcal{E}_t\}$.

In the remainder of this thesis, we adopt such a more flexible approach by allowing for both weak serial correlations and cross-correlations in $\{\mathcal{E}_t\}$. As an extension of Bai (2003), our methods utilize covariance information of the tensor time series by analyzing

the contemporary covariance matrix, which are more natural to apply to financial return data for example as opposed to methods that utilize only autocovariance information (see Wang et al. (2019) or Chen et al. (2022) for example). Due to market efficiency, population autocovariances of the data can be close to zero and methods that only utilize autocovariance information can have low signal-to-noise ratio.

Simultaneously, as we develop this thesis, there has also been other literature developed very recently to deal with matrix or tensor factor models with weak-serial and cross-correlations in $\mathscr{E}_t$ (Barigozzi et al., 2023b; Chen and Fan, 2021; He et al., 2023a, 2022; Yu et al., 2022). In these literature, the factor loading spaces are estimated by performing PCA on the sample covariance matrix or its variations under the matrix or tensor settings, and the number of factors is estimated by studying the behaviors of the eigenvalues of the corresponding covariance information. For matrix factor models (i.e., an order-2 tensor), Chen and Fan (2021) proposes an α-PCA method that aggregates both first and second moment information to estimate the factor loadings, assuming α-mixing of noise series. He et al. (2022) proposes using matrix Kendall's tau instead of the sample covariance matrix by assuming a matrix elliptical distribution of the noise. With the α-mixing assumption, Yu et al. (2022) develops a projection estimation (PE) method to estimate factor loadings for matrix factor models by projecting the observation matrix onto the row or column factor space and performing eigenanalysis on the covariance information after projection. The number of row and column factors is also estimated by the corresponding eigenvalue-ratio statistics. He et al. (2023a) provides the least squares interpretation of PE and proposes a robust method by substituting the least squares loss function with the Huber Loss function (see also He et al. (2023b)). As an extension, Barigozzi et al. (2023b) and Barigozzi et al. (2023a) further generalize PE and the robust method to estimate tensor factor models for a general K.

However, one limitation of all these recent developments is that they assume all factors are pervasive in every mode of the matrix or tensor, which can be restrictive in many real applications when weak factors are present. As far as we are concerned, there has been little literature dealing with tensor factor models assuming the presence of weak factors. The only exception is Han et al. (2020), who propose two parameters, δ_0 and δ_1, to control factor strengths for tensor factor models with independent $\mathscr{E}_t$. However, these parameters are not easily interpretable as an extension of the factor strength defined in (2.4) and (2.5).

For time series tensor factor models with weak-serial and cross-correlations in $\mathscr{E}_t$, the presence of weak factors can diminish the effectiveness of proposed estimation methods developed for pervasive factors only (Barigozzi et al., 2023b; Chen and Fan, 2021; He et al.,

2023a, 2022; Yu et al., 2022). The issue of weak factors becomes particularly pronounced in tensor factor models compared to vector factor models. Since many proposed methods in the current literature rely on iterated estimations of the factor matrices $\mathbf{A}_k$ for different tensor modes (Barigozzi et al., 2023a,b; Han et al., 2020; Yu et al., 2022), inaccuracies in each iteration of estimation are likely to accumulate, resulting in poor performance of the final estimator. Dealing with tensor factor models with weak factors remains a challenge, and this is the focus of our development in this thesis.

Chapter 3

Factor Loadings Estimation in Time Series Tensor Factor Models by Pre-averaging and Iterative Projection

3.1 Introduction

The occurrence of high-dimensional time series, observed in tensor format, is becoming more prevalent across diverse fields, including neuroimaging (Zhang, 2019), face recognition (Chen and Fan, 2021), traffic transport data (Chen et al., 2020), international trading (Chen and Chen, 2022), among many others. An effective approach for reducing the dimensionality of such high dimensional tensor time series involves adopting a factor model structure, similar to the Tucker decomposition for tensors.

As explored in Chapter 2, the application of factor models to matrix and tensor time series has garnered significant interest in recent years. Building upon the framework established for vector time series, researchers have developed two main types of assumptions to formulate tensor factor models. The first type assumes the independence of the noise tensor series $\mathcal{E}_t$ (Chen et al., 2018, 2022; Han et al., 2020, 2022; Liu et al., 2022; Wang et al., 2019; Yokota et al., 2017; Zhang and Xia, 2018), extending the factor model assumptions introduced by Lam et al. (2011). Under such assumptions, the estimation of the factor loading space is performed through analysing the lagged cross-product, the tensor version of non-zero lag autocovariance matrices. The second type allows for weak serial and cross-correlations in $\mathcal{E}_t$ (Barigozzi et al., 2023a,b; Chen and Fan, 2021; He et al., 2023a, 2022; Yu et al., 2022), building upon the approximate factor model assumptions proposed

by Bai and Ng (2002). Under these assumptions, principal component analysis (PCA) of the sample covariance matrix, or its variations under matrix or tensor settings, is typically employed to estimate the factor loading space. We adopt the second approach due to its flexibility, making it more applicable to a range of fields, such as economics and finance.

Though significant efforts have been devoted to the study of various models and estimation methods for tensor time series, there has been limited literature addressing tensor factor models with appropriate assumptions regarding the presence of weak factors. In practical scenarios, it is common for the existence of weak or "local" factors to be observed, particularly in applications related to finance and economics (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986; Uematsu and Yamagata, 2022). In such situations, tensor factor model estimation methods based solely on pervasive factors tend to exhibit poor performance. The exploration of weak factor models for tensor time series remains a challenging research topic.

To bridge the gap described above, we propose our time series tensor factor model (with tensor order two or above) with both serial dependence of $\{\mathscr{E}_t\}$ and the presence of weak factors. In this model, we allow for a spectrum of different factor strengths, which is a generalisation to Lam et al. (2011) when static vector time series factor model is concerned. To the best of our knowledge, our model is the first one in tensor factor modelling to consider weak factors when both serial and cross-correlations in $\{\mathscr{E}_t\}$ are presented. For tensor factor models with independent $\{\mathscr{E}_t\}$, while Han et al. (2020) has two parameters δ_0 and δ_1 controlling the factor strengths, these parameters are intricately linked to the definitions of their TOPUP and TIPUP estimators, making their interpretation within the broader context of the model, particularly in relation to $\mathbf{A}_k$, less straightforward. On the other hand, our factor strengths $\alpha_{k,j}$, $j \in [r_k]$, which has the j-th diagonal entry of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k \asymp d_k^{\alpha_{k,j}}$ (see Assumption (L1) in Section 3.2.1.3 for more details), can be more easily interpretable as a direct extension of the factor strengths definition by Lam et al. (2011) in vector factor model. Hence if the j-th column of $\mathbf{A}_k$ is dense (a pervasive factor), then $\alpha_{k,j} = 1$. If there are only finitely many non-zeros in the j-th column of $\mathbf{A}_k$, then it is a very weak factor, and $\alpha_{k,j} = 0$. Freyaldenhoven (2022) allows for these weaker factors in its vector time series factor model, and called them "local factors".

With relaxed assumptions for wider applications, and allowing for a spectrum of factors with different strengths, we establish a procedure for estimating the factor loading space of $\mathbf{A}_k$ for $k \in [K]$. We introduce a "pre-averaging" initial estimator and an iterative projection estimator for our model, with theoretical analyzes provided and rate of convergence spelt out. The pre-averaging procedure is presented in Section 3.2, which can be seen as a

random projection method by randomly summing tensor fibres, and we provide a method to control for the quality of the random projection in Section 3.2.3. Section 3.2.6 also shows that our pre-averaging estimator is minimax optimal under certain scenarios on a certain localized set. Iterative projection estimators of the factor loading matrices (see Section 3.3) are provided with idea similar to the projection method in Yu et al. (2022), except that we only project on the direction aligning to the strongest estimated factor. This is because we assume there are weak factors which may not be estimated with enough accuracy. With weak factors in the model, numerical experiments show that our estimator outperforms other competitors since we only utilize the information which captures the most accurate estimations so far. Note that all methods discussed in this chapter are proposed to be applied to tensor time series with order two or above. This is because the pre-averaging procedure and iterative algorithm rely on projecting information from different tensor modes, which cannot be applied when $K = 1$ (i.e., for vector time series).

The rest of this chapter is organized as follows. Section 3.2 presents important assumptions of our model, together with the idea of pre-averaging. Discussions and theory on choosing the "best" samples for aggregating results are presented, together with rate of convergence for our pre-averaging estimator for the strongest factors spelt out. Section 3.3 utilizes the pre-averaging estimator as the ideal initial estimator for re-estimating the projection direction by iterations, and presents the key theoretical results on the iterative projection estimators. Section 3.4 presents our simulation studies on a number of different settings and compare to other benchmarks or state-of-the-art estimators. All the proofs are presented in Section 3.5.

3.2 Initial Estimation of Strongest Factors by Pre-averaging

We define the tensor factor model for each $\mathcal{X}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t \in [T]$, as

$$\mathcal{X}_t = \mu + \mathcal{C}_t + \mathcal{E}_t = \mu + \mathcal{F}_t \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K + \mathcal{E}_t, \tag{3.1}$$

where we include a non-zero mean tensor $\mu \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ as compared to (2.7) introduced by Chen et al. (2022), which makes our model more flexible.

Before officially introducing the pre-averaging estimator, we first present some technical assumptions needed for the tensor factor model (3.1).

3.2.1 Assumptions

3.2.1.1 Assumptions on the errors

We present assumptions (E1) - (E2) below with explanations.

(E1) (Decomposition of error) *We assume that K is a constant, and*

$$\mathscr{E}_t = \mathscr{F}_{e,t} \times_1 \mathbf{A}_{e,1} \times \cdots \times \mathbf{A}_{e,K} + \boldsymbol{\varepsilon}_t, \tag{3.2}$$

where $\mathscr{F}_{e,t}$ is an order-K tensor with dimension $r_{e,1} \times \cdots \times r_{e,K}$, containing independent elements with mean 0 and variance 1. The order-K tensor $\boldsymbol{\varepsilon}_k \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ contains independent mean zero elements each with finite variance, with the two time series $\{\boldsymbol{\varepsilon}_t\}$ and $\{\mathscr{F}_{e,t}\}$ being independent.

Moreover, for each $k \in [K]$, $\mathbf{A}_{e,k} \in \mathbb{R}^{d_k \times r_{e,k}}$ is such that $\left\|\mathbf{A}_{e,k}\right\|_1 = O(1)$. That is, $\mathbf{A}_{e,k}$ is (approximately) sparse.

Hence with (E1), we have $\text{mat}_k(\mathscr{E}_t) = \mathbf{A}_{e,k}\text{mat}_k(\mathscr{F}_{e,t})\mathbf{A}_{e,\text{-}k}^{\mathrm{T}} + \text{mat}_k(\boldsymbol{\varepsilon}_t)$, where $\mathbf{A}_{e,\text{-}k} := \mathbf{A}_{e,K} \otimes \cdots \otimes \mathbf{A}_{e,k+1} \otimes \mathbf{A}_{e,k-1} \otimes \cdots \otimes \mathbf{A}_{e,1}$. Each mode-$k$ noise fibre $\mathbf{e}_{t,\text{-}k,\ell}$ for $\ell \in [d_{\text{-}k}]$ can then be decomposed as

$$\mathbf{e}_{t,\text{-}k,\ell} := \mathbf{A}_{e,k}\text{mat}_k(\mathscr{F}_{e,t})\mathbf{a}_{e,\text{-}k,\ell} + (\boldsymbol{\Sigma}_{\varepsilon,\ell}^{(k)})^{1/2}\boldsymbol{\varepsilon}_{t,\ell}^{(k)}, \tag{3.3}$$

where $\mathbf{a}_{e,\text{-}k,\ell}^{\mathrm{T}}$ is the ℓ-th row of $\mathbf{A}_{e,\text{-}k}$, $\boldsymbol{\Sigma}_{\varepsilon,\ell}^{(k)}$ is diagonal and $\boldsymbol{\varepsilon}_{t,\ell}^{(k)}$ contains independent elements each with mean 0 and variance 1. The above decomposition means that each noise fibre is now a sum of two parts. The first part is similar to a common component with a factor loading matrix $\mathbf{A}_{e,k}$, while the second part contains independent noise (but can still exhibit serial correlations; see Assumption (E2)). However, $\mathbf{A}_{e,k}$ is (approximately) sparse here and contains at most a very weak factor with factor strength 0 (see Assumption (L1) in Section 3.2.1.3). This part facilitates cross-correlations of noise fibres, with the covariance of any two mode-k fibres of the noise tensor taking the form

$$\text{cov}(\mathbf{e}_{t,\text{-}k,\ell}, \mathbf{e}_{t,\text{-}k,m}) = \mathbf{a}_{e,\text{-}k,\ell}^{\mathrm{T}}\mathbf{a}_{e,\text{-}k,m}\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}},$$

where $\mathbf{e}_{t,\text{-}k,\ell}$ and $\mathbf{e}_{t,\text{-}k,m}$ are the mode-k noise fibres with $l, m \in [d_{\text{-}k}]$. This error structure satisfies the assumptions in Chen and Fan (2021) when $r_{e,k} = O(d_k)$. In fact, if $\mathbf{A}_{e,k}$ is not (approximately) sparse, it should be counted as a factor loading matrix rather than a noise component in our model.

(E2) (Time series) *There is* $\mathscr{Z}_{e,t}$ *the same dimension as* $\mathscr{F}_{e,t}$*, and* $\mathscr{Z}_{\varepsilon,t}$ *the same dimension as* $\boldsymbol{\varepsilon}_t$*, such that* $\mathscr{F}_{e,t} = \sum_{q\geq 0} a_{e,q}\mathscr{Z}_{e,t-q}$ *and* $\boldsymbol{\varepsilon}_t = \sum_{q\geq 0} a_{\varepsilon,q}\mathscr{Z}_{\varepsilon,t-q}$*, with* $\{\mathscr{Z}_{e,t}\}$ *and* $\{\mathscr{Z}_{\varepsilon,t}\}$ *independent of each other, and each time series have i.i.d. elements with mean 0 and variance 1. The coefficients* $a_{e,q}$ *and* $a_{\varepsilon,q}$ *are so that* $\sum_{q\geq 0} a_{e,q}^2 = \sum_{q\geq 0} a_{\varepsilon,q}^2 = 1$ *and* $\sum_{q\geq 0} |a_{e,q}| \leq C$*,* $\sum_{q\geq 0} |a_{\varepsilon,q}| \leq C$ *for some constant* C*.*

With this assumption, the error variables in $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$ are serially correlated in general. Together with (E1), (weak) serial and cross-sectional dependence within and among fibres are allowed for the errors. (E1) and (E2) together satisfy Assumption 3 of Barigozzi et al. (2023b), who provide a more general expression to quantify the serial and cross-correalation of the noise by bounding the sums of cross and autocovariances of the idiosyncratic terms.

3.2.1.2 Assumptions on the factors

Similar to (E2), the factors in $\mathscr{F}_t$ are assumed to follow general linear processes.

(F1) *There is* $\mathscr{Z}_{f,t}$ *the same dimension as* $\mathscr{F}_t$*, such that* $\mathscr{F}_t = \sum_{q\geq 0} a_{f,q}\mathscr{Z}_{f,t-q}$*. The time series* $\{\mathscr{Z}_{f,q}\}$ *has i.i.d. elements with mean 0 and variance 1. The coefficients* $a_{f,q}$ *are so that* $\sum_{q\geq 0} a_{f,q}^2 = 1$ *and* $\sum_{q\geq 0} |a_{f,q}| \leq C$ *for some constant* C*.*

Note the series of coefficients $\{a_{e,q}\}$, $\{a_{\varepsilon,q}\}$ and $\{a_{f,q}\}$ are not necessarily equal.

3.2.1.3 Assumptions on the model parameters

We present the assumptions needed for the factor loading matrices $\mathbf{A}_k$, $k \in [K]$, and other model parameters.

(L1) (Factor Strength) *We assume that, for* $k \in [K]$*,* $\mathbf{A}_k$ *is of full rank,* $r_k = o(T^{1/3})$*, and as* $d_k \to \infty$*,*

$$\mathbf{D}_k^{-1/2}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-1/2} \to \boldsymbol{\Sigma}_{\mathbf{A},k}, \tag{3.4}$$

where $\mathbf{D}_k = \mathrm{diag}(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k)$ *and* $\boldsymbol{\Sigma}_{\mathbf{A},k}$ *is positive definite with all eigenvalues bounded away from 0 and infinity. We assume* $(\mathbf{D}_k)_{jj} \asymp d_k^{\alpha_{k,j}}$ *for* $j \in [r_k]$*, and* $0 < \alpha_{k,r_k} \leq \cdots \leq \alpha_{k,2} \leq \alpha_{k,1} \leq 1$*.*

Assumption (L1) states that the factors can have different strengths. When $K = 1$ and $\alpha_{1,j} = \alpha$ for $j \in [r_1]$, (3.4) reduces to the assumption of (approximate) vector factor model with the same strength, which is discussed in Bai and Ng (2023). Hence, our assumption is a generalisation of Bai and Ng (2023) to a tensor setting with mixed strengths of factors, which is more flexible to apply on many real datasets. In addition, we do not assume the orthogonality of $\mathbf{A}_k$ as Freyaldenhoven (2022) did, since this would be incompatible with the expression of factor strength and signal accumulation in terms of the norm and row sum of $\mathbf{A}_k$. The concept of a pervasive factor, for instance, depends on a column of $\mathbf{A}_k$ being dense. However, such an interpretation can be lost completely under the assumption of orthogonal columns in $\mathbf{A}_k$.

Note that in Assumption (L1), we allow r_k to diverge as well, facilitating the possibility that the product $r = \prod_{k=1}^{K} r_k$ becomes large with an increasing K. However, we still assume that K itself is finite because in practice, it is hard to observe and interpret a tensor with an extremely large number of modes (K). Our theorems allow for the flexibility by taking the potentially increasing rate of r into consideration.

(R1) *The time series* $\{\mathscr{Z}_{f,t}\}$ *from Assumption (F1),* $\{\mathscr{Z}_{e,t}\}$ *and* $\{\mathscr{Z}_{\varepsilon,t}\}$ *from Assumption (E2) are mutually independent of each other. Define* $z_{\varepsilon,t,i_1,\cdots,i_K}$, *to be an element of* $\mathscr{Z}_{\varepsilon,t}$ *for* $t \in [T], i_k \in [d_k]$. *Similarly define* $z_{f,t,i_1,\cdots,i_K}$ *and* $z_{e,t,i_1,\cdots,i_K}$. *We assume* $\mathbb{E}|z_{\varepsilon,t,i_1,\cdots,i_K}|^4 \le c$, $\mathbb{E}|z_{f,t,i_1,\cdots,i_K}|^4 \le c$, $\mathbb{E}|z_{e,t,i_1,\cdots,i_K}|^4 \le c$ *for some* $c < \infty$ *independent of* t *and* i_k.

(R2) *We assume* $\lambda_{d_k}(\mathbf{\Sigma}_{\varepsilon,\ell}^{(k)})$ *is uniformly bounded below from 0 for* $\ell \in [d_{-k}]$, *where* $\mathbf{\Sigma}_{\varepsilon,\ell}^{(k)}$ *is defined in (3.3). Let* $\mathbf{A}_{\varepsilon,T}$ *be the* $T \times T$ *matrix with its* (t,s) *element to be* $(\mathbf{A}_{\varepsilon,T})_{t,s} = \sum_{q\ge 0} a_{\varepsilon,q} a_{\varepsilon,q+|t-s|}$. *Denote* $0 < y := \lim_{d_k,T\to\infty} \frac{\min(d_k,T)}{\max(d_k,T)} \le 1$ *and* $y^* = \min(y,1)$, *then we assume there exists* $c_1 \in (1-y^*,1]$ *such that* $\lambda_{\lfloor c_1 T\rfloor}(\mathbf{A}_{\varepsilon,T}) > c_2 > 0$ *for large* T, *where* c_2 *is a positive constant.*

Assumption (R1) relaxes the need for Gaussian or sub-Gaussian random variables (see Zhang and Xia (2018) and Chen et al. (2022) for example), with only bounded fourth order moments required. This allows for substantially more types of data to be analyzed. For instance, financial returns data typically displays behavior of heavy-tailed distribution, where we do not usually want to assume moments beyond order four exist. We demonstrate in Section 3.4 that in practice our proposed estimators remain robust even when this assumption is violated and bounded fourth-order moments do not exist. Note that Assumption (E2) and (R1) together imply that we are assuming linear process

for factors and idiosyncratic components with finite fourth order moments. To quantify serial dependence, other types of assumptions are also possible, such as assuming an α-mixing process (Chen and Fan, 2021; Yu et al., 2022) or directly bounding the sums of autocovariances of the idiosyncratic terms (Bai, 2003; Bai and Ng, 2002; Barigozzi et al., 2023b).

Finally, together with Assumption (R1), Assumption (R2) enables us to bound the eigenvalues of various sample covariance matrices from below (see (3.28), (3.29) and (3.36) in Lemma 3.1 and Lemma 3.2). The first part of Assumption (R2) assumes the positive definiteness of the covariance matrix of independent noise, and the second part presents the technical requirement similar to Assumption D of Ahn and Horenstein (2013) which enables us to utilize random matrix theory. As long as the serial correlations of the $\varepsilon_{t,\ell,j}^{(k)}$'s are not too strong, Assumption (R2) will be satisfied.

For convenience of further theoretical analysis, we define $\mathbf{Q}_k = \mathbf{A}_k \mathbf{D}_k^{-1/2}$. Since $\mathbf{Q}_k^{\mathrm{T}}\mathbf{Q}_k \to \mathbf{\Sigma}_{A,k}$, $\mathbf{Q}_k$ is a re-normalized version of $\mathbf{A}_k$. In addition, we apply the singular value decomposition of $\mathbf{A}_k$ as

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{G}_k^{1/2} \mathbf{V}_k^{\mathrm{T}}, \tag{3.5}$$

where $\mathbf{U}_k \in \mathbb{R}^{d_k \times r_k}$ has orthogonal columns such that $\mathbf{U}_k^{\mathrm{T}}\mathbf{U}_k = \mathbf{I}_{r_k}$, $\mathbf{G}_k \in \mathbb{R}^{r_k \times r_k}$ is diagonal and consists of the eigenvalues of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k$ in decreasing order, and $\mathbf{V}_k \in \mathbb{R}^{r_k \times r_k}$ is an orthogonal matrix. The subspaces spanned by the columns of $\mathbf{U}_k$, $\mathbf{Q}_k$ and $\mathbf{A}_k$ are the same, and hence it is equivalent to estimate $\mathbf{U}_k$ (or $\mathbf{Q}_k$) and $\mathbf{A}_k$, and the columns of $\mathbf{U}_k$ form an orthonormal basis for the column space spanned by $\mathbf{Q}_k$ (or $\mathbf{A}_k$). We will estimate $\mathbf{U}_k$ (or $\mathbf{Q}_k$) instead of $\mathbf{A}_k$ in the sections that follow. We need another regularity condition on the singular values on $\mathbf{G}_k$. This can be relaxed at the expense of lengthier explanations involving factor loading spaces in all subsequent theorems.

(L1') The singular values on $\mathbf{G}_k$ are distinct.

3.2.2 Potential advantages of pre-averaging

To estimate $\mathbf{A}_k$, using (2.2), the mode-k unfolding of (3.1) can be written as

$$\mathrm{mat}_k(\mathscr{X}_t) = \mathrm{mat}_k(\boldsymbol{\mu}) + \mathbf{A}_k \mathrm{mat}_k(\mathscr{F}_t)\mathbf{A}_{\text{-}k}^{\mathrm{T}} + \mathrm{mat}_k(\mathscr{E}_t).$$

If we define $S_j \subseteq [d_j]$ for $j \in [K]$, then we can always define the Cartesian product

$$S_{-k} := S_K \times \cdots \times S_{k+1} \times S_{k-1} \times \cdots \times S_1, \text{ such that}$$
$$\mathbf{1}_{d_{-k},S_{-k}} = \mathbf{1}_{d_K,S_K} \otimes \cdots \otimes \mathbf{1}_{d_{k+1},S_{k+1}} \otimes \mathbf{1}_{d_{k-1},S_{k-1}} \otimes \cdots \otimes \mathbf{1}_{d_1,S_1}.$$

Projecting on $\mathbf{1}_{d_{-k},S_{-k}}$, equivalent to summing the fibres in $\mathrm{mat}_k(\mathscr{X}_t)$ over the set S_{-k}, is then

$$\begin{aligned}\mathrm{mat}_k(\mathscr{X}_t)\mathbf{1}_{d_{-k},S_{-k}} &= \mathrm{mat}_k(\mu)\mathbf{1}_{d_{-k},S_{-k}} + \mathbf{A}_k \mathrm{mat}_k(\mathscr{F}_t)\mathbf{A}_{-k}^{\mathrm{T}}\mathbf{1}_{d_{-k},S_{-k}} + \mathrm{mat}_k(\mathscr{E}_t)\mathbf{1}_{d_{-k},S_{-k}}, \text{ where} \\ \mathbf{A}_{-k}^{\mathrm{T}}\mathbf{1}_{d_{-k},S_{-k}} &= \mathbf{A}_K^{\mathrm{T}}\mathbf{1}_{d_K,S_K} \otimes \cdots \otimes \mathbf{A}_{k+1}^{\mathrm{T}}\mathbf{1}_{d_{k+1},S_{k+1}} \otimes \mathbf{A}_{k-1}^{\mathrm{T}}\mathbf{1}_{d_{k-1},S_{k-1}} \otimes \cdots \otimes \mathbf{A}_1^{\mathrm{T}}\mathbf{1}_{d_1,S_1},\end{aligned} \tag{3.6}$$

with $\mathbf{A}_{-k} := \mathbf{A}_K \otimes \cdots \mathbf{A}_{k+1} \otimes \mathbf{A}_{k-1} \otimes \cdots \otimes \mathbf{A}_1$. Hence projection of the data using $\mathbf{1}_{d_{-k},S_{-k}}$ can be seen as pre-averaging the rows of each $\mathbf{A}_j$ using S_j for $j \in [K] \setminus \{k\}$.

It is important to note that the pre-averaging procedure for estimating $\mathbf{A}_k$ is based on summing the rows of $\mathbf{A}_j$ for $j \neq k$. For example, when $K = 2$, pre-averaging the rows of $\mathbf{A}_2$ is for estimating $\mathbf{A}_1$, while pre-averaging the rows of $\mathbf{A}_1$ is for estimating $\mathbf{A}_2$. For ease of understanding, consider the case when $K = 2$ and the model becomes

$$\mathbf{X}_t = \mu + \mathbf{A}_1\mathbf{F}_t\mathbf{A}_2^{\mathrm{T}} + \mathbf{E}_t. \tag{3.7}$$

To estimate $\mathbf{A}_1$, we only have one mode (mode-2) for pre-averaging, and thus $\mathbf{A}_{-1} = \mathbf{A}_2$, $S_{-1} = S_2$ and $\mathbf{1}_{d_{-1},S_{-1}} = \mathbf{1}_{d_2,S_2}$. Then projecting on $\mathbf{1}_{d_2,S_2}$ is equivalent to summing the columns (mode-1 fibres) in $\mathbf{X}_t$ over the set S_2, which leads to

$$\mathbf{X}_t\mathbf{1}_{d_2,S_2} = \mu\mathbf{1}_{d_2,S_2} + \mathbf{A}_1\mathbf{F}_t\mathbf{A}_2^{\mathrm{T}}\mathbf{1}_{d_2,S_2} + \mathbf{E}_t\mathbf{1}_{d_2,S_2}. \tag{3.8}$$

Hence, we are projecting of the data using $\mathbf{1}_{d_2,S_2}$, which can be seen as pre-averaging the rows of $\mathbf{A}_2$ using $S_2 \subseteq [d_2]$.

While we re-estimate by projection in Section 3.3, and papers like Yu et al. (2022) does projection estimation as well, the aim of this section is to provide an initial estimator of projection direction with quality that can be *controlled* by careful selection of randomly generated S_j. The method to select S_j among multiple random samples is introduced in Section 3.2.3, which leads to the pre-averaging estimator in Section 3.2.5.

The potential advantages of pre-averaging can be illustrated as follows. Consider just calculating the second order moments

$$\sum_{t=1}^{T} \mathrm{mat}_k(\mathcal{X}_t - \bar{\mathcal{X}})\mathrm{mat}_k^{\mathrm{T}}(\mathcal{X}_t - \bar{\mathcal{X}}) =: S_0 + N_1 + N_1^{\mathrm{T}} + N_2, \text{ where}$$

$$S_0 := \mathbf{A}_k \sum_{t=1}^{T} \Big(\mathrm{mat}_k(\mathcal{F}_t - \bar{\mathcal{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathrm{mat}_k^{\mathrm{T}}(\mathcal{F}_t - \bar{\mathcal{F}})\Big)\mathbf{A}_k^{\mathrm{T}}, \tag{3.9}$$

$$N_1 := \mathbf{A}_k \sum_{t=1}^{T} \Big(\mathrm{mat}_k(\mathcal{F}_t - \bar{\mathcal{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathrm{mat}_k^{\mathrm{T}}(\mathcal{E}_t - \bar{\mathcal{E}})\Big), \quad N_2 := \sum_{t=1}^{T} \mathrm{mat}_k(\mathcal{E}_t - \bar{\mathcal{E}})\mathrm{mat}_k^{\mathrm{T}}(\mathcal{E}_t - \bar{\mathcal{E}}),$$

and extracting an estimator of $\mathbf{A}_k$ through PCA (e.g., see Bai (2003)). Our proposed pre-averaging estimator, like a projected estimator, can accumulate significantly more signals before doing the PCA step for extracting an estimator of $\mathbf{A}_k$. This is because the signal term $\mathbf{A}_k \sum_{t=1}^{T} \mathrm{mat}_k(\mathcal{F}_t - \bar{\mathcal{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathrm{mat}_k^{\mathrm{T}}(\mathcal{F}_t - \bar{\mathcal{F}})\mathbf{A}_k^{\mathrm{T}}$ (from using $\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}$ as the projection direction of the data) can be significantly larger than S_0 in (3.9), since the diagonal elements of $\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}^{\mathrm{T}}\mathbf{A}_{\text{-}k}$ can be much larger than those in $\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{A}_{\text{-}k}$. For instance, when $\mathbf{A}_{\text{-}k}$ has a column with all positive or negative elements (e.g., factor loading entries for the market factors in finance), we have diagonal elements of $\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}\mathbf{1}_{d_{\text{-}k},S_{\text{-}k}}^{\mathrm{T}}\mathbf{A}_{\text{-}k}$ of order $d_{\text{-}k}^2$, while those in $\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{A}_{\text{-}k}$ are only of order $d_{\text{-}k}$.

For ease of understanding, consider the matrix factor model (3.7) when $K = 2$. To estimate $\mathbf{A}_1$, the signal term of the projected data (3.8) obtained by pre-averaging becomes $\mathbf{A}_1 \left(\sum_{t=1}^{T} \mathbf{F}_t\mathbf{A}_2^{\mathrm{T}}\mathbf{1}_{d_2,S_2}\mathbf{1}_{d_2,S_2}^{\mathrm{T}}\mathbf{A}_2\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{A}_1^{\mathrm{T}}$, while the signal (3.9) for directly using the second order moments becomes $\mathbf{A}_1 \left(\sum_{t=1}^{T} \mathbf{F}_t\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{A}_1^{\mathrm{T}}$. If $\mathbf{A}_2$ has a column with the majority of elements having the same sign and $|S_2| \asymp d_2$, then the diagonal elements of $\mathbf{A}_2^{\mathrm{T}}\mathbf{1}_{d_2,S_2}\mathbf{1}_{d_2,S_2}^{\mathrm{T}}\mathbf{A}_2$ have order d_2^2, while those in $\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2$ are only of order d_2. Thus, pre-averaging the rows of $\mathbf{A}_2$ can potentially accumulate significantly more signals in estimating $\mathbf{A}_1$. For example, in practice, the first PC is usually the mean of data, which has majority of elements of the same sign. If the first factor in $\mathbf{A}_2$ coincides with the first PC, then the diagonal elements of $\mathbf{A}_2^{\mathrm{T}}\mathbf{1}_{d_2,S_2}\mathbf{1}_{d_2,S_2}^{\mathrm{T}}\mathbf{A}_2$ can easily achieve order d_2^2, which leads to significant signal accumulation that helps us obtain a more accurate estimator of $\mathbf{A}_1$.

The following assumption provides a more technical definition of signal accumulation through pre-averaging.

(L2) (Signal accumulation from summing) *For $k \in [K]$, let $M_{k,0} > 0$ be the number of different sums of rows of $\mathbf{A}_k$ considered, and for $m \in [M_{k,0}]$, denote $S_{k,m} \subseteq [d_k]$ to be the m-th index set for summing the rows of $\mathbf{A}_k$. With the choice of $|S_{k,m}| = \lfloor d_k/2 \rfloor$,*

define

$$s_{k,m} := \left\|\mathbf{A}_k^{\mathrm{T}}\mathbf{1}_{d_k,S_{k,m}}\right\|^2, \quad s_{k,\max} := \max_{m\in[M_{k,0}]} s_{k,m}, \quad s_{\text{-}k,\max} := \prod_{j=1;j\neq k}^{K} s_{j,\max}. \tag{3.10}$$

We assume for some $z_k \leq r_k$,

$$\frac{d_{\text{-}k}}{s_{\text{-}k,\max}}\left(1+\frac{d_k}{T}\right) = o\left(d_k^{\alpha_{k,z_k}}\right). \tag{3.11}$$

In Assumption (L2), $s_{k,m}$ can be seen as a measure of accumulation of signals for a specific sample $m \in [M_{k,0}]$, and $s_{k,max}$ is the "largest" accumulation of signal we can attain over the $M_{k,0}$ samples. In Section 3.2.3, the method to provide a carefully selection of randomly generated $S_{j,m}$ is introduced, and Section 3.2.4 gives a more thorough discussion on the number of samples needed to secure enough signal accumulation with a large probability.

Note that we choose $S_{k,m}$ with size $|S_{k,m}| \asymp d_k$ (e.g., $|S_{k,m}| = \lfloor d_k/2 \rfloor$ in Assumption (L2)) for each $m \in [M_{k,0}]$. This choice allows for the sum of rows of $\mathbf{A}_k$ to be potentially large with a large probability (see also Section 3.2.4).

We also remark that the signal accumulation in Assumption (L2) does not require each d_k to be diverging. In Assumption (L2), some d_k's can be finite as long as $d_{\text{-}k}/s_{\text{-}k,\max} = o(1)$. This can be achieved when, for example, there is an $\mathbf{A}_j$ for some $j \neq k$ such that the majority of the elements in a column are of the same sign, so that $s_{j,m} \succ d_j$, resulting in $d_{\text{-}k}/s_{\text{-}k,\max} = o(1)$. Thus, to estimate $\mathbf{A}_k$ for a specific mode k, we can potentially allow d_j's for $j \neq k$ to be finite if the above situation holds, though d_k itself should still be diveging by Assumption (L1). However, we do not know this beforehand, and we suggest to use the method when all d_k's are diverging to be on the safe side. In addition, as in Assumption (L1), having $d_k \to \infty$ is a common assumption to make sure 'spiked' signal eigenvalues are larger (in rate) than those from the noise.

3.2.3 Choosing samples of tensor fibres

We first present an algorithm for choosing the "best" sample of tensor fibres to sum. Recall from Assumption (L2) that we define $s_{k,m}$ to be a measure of signal accumulation for a specific sample $m \in [M_{k,0}]$, and let $s_{\text{-}k,m} := \prod_{j=1;j\neq k}^{K} s_{j,m}$. To estimate $\mathbf{A}_k$, we aim to find the sample with the largest $\mathbf{s}_{\text{-}k,m}$, i.e., the largest signal accumulation by pre-averaging the rows of each $\mathbf{A}_j$ for $j \neq k$, as presented in (3.6). The following algorithm is designed to

produce the sample with the largest signal accumulation over all random samples generated.

Algorithm for choosing the "best" sample of tensor fibres

1. Initialize $M_{k,0}$ for each $k \in [K]$.
2. Generate a sequence of independent sets $\{S_{k,m}\}_{k\in[K],m\in[M_{k,0}]}$. Each $S_{k,m}$ chooses uniformly over $[d_k]$, with $|S_{k,m}| = \lfloor d_k/2 \rfloor$.
3. Fix $k \in [K]$. Define $M_0 := \prod_{j\in[K]\setminus\{k\}} M_{j,0}$. For each $m \in [M_0]$, define $S_{\text{-}k,m} := \times_{j\in[K]\setminus\{k\}} S_{j,m_j}$ and $\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,m}} := \otimes_{j\in[K]\setminus\{k\}} \mathbf{1}_{d_j,S_{j,m_j}}$ for some $m_j \in [M_{j,0}]$.
4. For the same fixed k from step 3, define for each $m \in [M_0]$,

$$\widetilde{\mathbf{X}}_{k,m} := (\mathrm{mat}_k(\mathscr{X}_1)\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,m}}, \ldots, \mathrm{mat}_k(\mathscr{X}_T)\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,m}})^{\mathrm{T}}, \tag{3.12}$$

   and for an integer l satisfying $r_k + 1 \le l \le \lfloor c\min(T, d_k)\rfloor - r_k$ for some $c > 0$, construct

$$\mathrm{ER}_{l,m} := \frac{\lambda_1\left(\widetilde{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{X}}_{k,m}\right)}{\lambda_l\left(\widetilde{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{X}}_{k,m}\right)}, \tag{3.13}$$

   which represents the ratio of the largest eigenvalue and the l-th largest eigenvalue of the sample covariance matrix of $\widetilde{\mathbf{X}}_{k,m}$, with l chosen to be greater than r_k.
5. The "best" sample $m \in [M_0]$ for estimating $\mathbf{A}_k$ is the one that maximizes $\mathrm{ER}_{l,m}$. We denote by $S_{\text{-}k,\max} := \times_{j\in[K]\setminus\{k\}} S_{j,\max}$ the corresponding product set, and

$$s_{\text{-}k,\max} := \prod_{j\in[K]\setminus\{k\}} s_{j,\max} := \prod_{j\in[K]\setminus\{k\}} \left\|\mathbf{A}_j^{\mathrm{T}}\mathbf{1}_{d_j,S_{j,\max}}\right\|^2.$$

6. Repeat steps 3,4,5 until each $k \in [K]$ is covered.

The justification of step 4 is as follows. With Assumption (L1) and (R2) satisfied, we have by Lemma 3.2 that the eigenvalue-ratio $\mathrm{ER}_{l,m}$ in (3.13) has

$$\mathrm{ER}_{l,m} \asymp_P d_k^{\alpha_{k,1}}\left[\frac{d_{\text{-}k}}{s_{\text{-}k,m}}\left(1 + \frac{d_k}{T}\right)\right]^{-1}. \tag{3.14}$$

$\mathrm{ER}_{l,m}$ can be seen as a measure of the signal-to-noise ratio in a specific sample $\widetilde{\mathbf{X}}_{k,m}$, since for the sample covariance matrix of $\widetilde{\mathbf{X}}_{k,m}$, its largest eigenvalue reflects the largest

magnitude of signal observed, and its l-th largest eigenvalue with $l \geq r_k + 1$ reflects the approximate level of noise contained in the sample. Hence, for a specific sample, (3.14) implies that $\mathrm{ER}_{l,m}$ should have magnitude proportional to $s_{\text{-}k,m}$, and larger $\mathrm{ER}_{l,m}$ observed means the sample has accumulated a larger magnitude of signal. Thus, the sample that maximised $\mathrm{ER}_{l,m}$ in fact asymptotically maximizes the product of signals, from $s_{\text{-}k,m}$ to $s_{\text{-}k,\max}$.

To better understanding the above algorithm, consider the case when $K = 2$ and we want to estimate $\mathbf{A}_1$. Then $M_0 = M_{2,0}$, and for each random sample $m \in [M_0]$, we generate $S_{2,m} \subseteq [d_2]$ (with $|S_{2,m}| = \lfloor d_2/2 \rfloor$) to be the m-th index set. We next create $\widetilde{\mathbf{X}}_{1,m} := (\mathbf{X}_1 \mathbf{1}_{d_2,S_{2,m}}, \ldots, \mathbf{X}_T \mathbf{1}_{d_2,S_{2,m}})^{\mathrm{T}}$, and calculate the eigenvalue ratio $\mathrm{ER}_{l,m}$ as in (3.13). Finally, among all M_0 samples generated, we find the one which maximizes $\mathrm{ER}_{l,m}$ as the "best" sample for estimating $\mathbf{A}_1$, with the largest accumulated signal to be defined as $s_{\text{-}1,max} = s_{2,max}$.

In step 4 of the above algorithm, we need to choose l such that $l \geq r_k + 1$. One way to choose l is to use expert opinion. A typical value of l we use depends on the user's idea of the maximum value of r_k. Suppose for an economic data set, we expect $r_k \leq 8$. Then we can use $l = 9$ for constructing ER_l. For a more data-driven way, note from Lemma 3.2 that for a particular sample with product set $S_{\text{-}k,m} \subseteq [d_{\text{-}k}]$,

$$\lambda_i\Big(\widetilde{\mathbf{X}}_{k,m}^{\mathrm{T}}\Big(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\Big)\widetilde{\mathbf{X}}_{k,m}/T\Big) \asymp_P \begin{cases} d_k^{\alpha_{k,i}}, & i \in [r_k]; \\ \frac{d_{\text{-}k}}{s_{\text{-}k,m}}\Big(1 + \frac{d_k}{T}\Big), & r_k + 1 \leq i \leq \lfloor c\min(T, d_k)\rfloor - r_k, \end{cases}$$

where $s_{\text{-}k,m}$ is defined in (3.10), and $\widetilde{\mathbf{X}}_{k,m}$ in (3.12). Hence for $d_k \asymp T$, if we have a sample $S_{\text{-}k,m}$ such that $d_{\text{-}k}/s_{\text{-}k,m} = O(1)$, then plotting the ordered-eigenvalues from the largest to smallest, we would expect to see a large dip at the $(r_k + 1)$th position. If we do not see such a dip, then we can generate another sample $S_{\text{-}k,m}$ and try again. Obtaining a sample with $d_{\text{-}k}/s_{\text{-}k,m} = O(1)$ should not take long. See the section below.

3.2.4 How many samples do we need

In most applications with $d_k = O(T)$ for each $k \in [K]$, if the ratio $d_{\text{-}k}/s_{\text{-}k,\max} = O(1)$, then Assumption (L2) is automatically satisfied, and the rate of convergence in (3.18) in Theorem 3.1 becomes $d_k^{-\alpha_{k,1}}$ when we choose $z_k = 1$ there. One way to achieve this is to have $s_{k,max} \asymp d_k$ for each $k \in [K]$.

The number of samples required to achieve $s_{k,max} \asymp d_k$ depends on the signs of elements, or more specifically, mean of the elements in $\mathbf{A}_k$. If there exists a dense column in $\mathbf{A}_k$ with the majority of elements having the same sign (i.e. the column has a non-zero mean), then any sample with $|S_{k,m}| = \lfloor d_k/2 \rfloor$ can achieve $\mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}} \asymp d_k$, and thus $s_{k,m} = (\mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}})^2 \asymp d_k^2$, which is larger than d_k. Therefore, $s_{k,\max} = (\max_{m\in[M_{k,0}]} \mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}})^2 \asymp d_k^2$ can be easily achieved with just one sample (i.e. $M_{k,0} = 1$).

On the other hand, if all columns in $\mathbf{A}_k$ contain a significant number of elements with opposite signs (for example, half positive and half negative), we may require more samples to achieve $s_{k,max} \asymp d_k$. Consider the scenario where for each $k \in [K]$, $r_k = 1$ and $\mathbf{A}_k$ contains d_k i.i.d. standard normal random variables, with $\mathbf{A}_i$ independent of $\mathbf{A}_j$ for $i \neq j$. In this case, approximately half of the elements of $\mathbf{A}_k$ are positive and the other half are negative, and the column of $\mathbf{A}_k$ has mean zero. For each $S_{k,m} \subseteq [d_k]$ with $m \in [M_{k,0}]$, we want to choose the $S_{k,m}$ such that $\mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}}$ is the largest, and that $s_{k,\max} = (\max_{m\in[M_{k,0}]} \mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}})^2 \asymp_P d_k$. Now for each $m \in [M_{k,0}]$,

$$z_{k,m} := \frac{\mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}}}{\lfloor d_k/2 \rfloor^{1/2}} \sim N(0,1), \text{ and } \operatorname{corr}(z_{k,m_1}, z_{k,m_2}) = \frac{|S_{k,m_1} \cap S_{k,m_2}|}{\lfloor d_k/2 \rfloor}, \tag{3.15}$$

if we are choosing $|S_{k,m}| = \lfloor d_k/2 \rfloor$ for each $m \in [M_{k,0}]$. Then by Theorem 3.4 of Hartigan (2014), we have

$$P\Big(\max_{m\in[M_{k,0}]} z_{k,m} \geq \sigma(N + L_\alpha - \frac{1}{2}\log(N + L_\alpha)) \Big) \geq 1 - 2\alpha, \text{ where}$$

$$N := \log(M_{k,0}^2/2\pi),\ L_\alpha := -2\log(-\log(\alpha)),\ \sigma := \min_{i\in[M_{k,0}]} \operatorname{var}(z_{k,i} - E(z_{k,i}|z_{k,1}, \ldots, z_{k,i-1})),$$

as long as $N + L_\alpha \geq 6$. With $\alpha = 0.025$, then $N + L_\alpha \geq 6$ implies $M_{k,0} \geq 186$, and with this we have

$$P(\max_{m\in[M_{k,0}]} z_{k,m} \geq 5.1\sigma) \geq 0.95, \tag{3.16}$$

meaning that $s_{k,\max} = (\max_{m\in[M_{k,0}]} \mathbf{A}_k^{\mathrm{T}} \mathbf{1}_{d_k,S_{k,m}})^2$ has order d_k with over 95% probability. Hence if $K = 2$, when we are estimating $\mathbf{A}_1$ and to sample fibres from $\mathrm{mat}_1(\mathscr{X}_t)$ using $S_{\text{-}1,\max} = S_{2,\max}$, we have when $M_0 = M_{2,0} \geq 186$ that over 95% probability we can have $s_{\text{-}1,\max} = s_{2,\max} \asymp d_2 = d_{\text{-}1}$.

In a more extreme scenario, suppose $r_k = 1$ and each element in $\mathbf{A}_k$ follows the distribution such that $\mathbb{P}\big((\mathbf{A}_k)_j = c_1\big) = 0.5$ and $\mathbb{P}\big((\mathbf{A}_k)_j = c_2\big) = 0.5$, with $c_1 > 0$ and $c_2 < 0$. This represents the case where half of the elements in $\mathbf{A}_k$ are strictly positive and

the other half are strictly negative. If $c_1+c_2\neq 0$, then any sample with $|S_{k,m}|=\lfloor d_k/2\rfloor$ can easily have $\mathbf{A}_k^{\mathrm{T}}\mathbf{1}_{d_k,S_{k,m}}\asymp d_k$, and thus achieve $s_{k,max}=(\mathbf{A}_k^{\mathrm{T}}\mathbf{1}_{d_k,S_{k,m}})^2\asymp d_k^2$. If $c_1+c_2=0$, then by applying the Central Limit Theorem, as $d_k\to\infty$, we have $z_{k,m}:=\frac{\mathbf{A}_k^{\mathrm{T}}\mathbf{1}_{d_k,S_{k,m}}}{\lfloor d_k/2\rfloor^{1/2}}\xrightarrow{D}N(0,1)$, and the arguments from (3.15) onwards apply. Thus, with $M_{k,0}\geq 186$, we have $s_{k,\max}=(\max_{m\in[M_{k,0}]}\mathbf{A}_k^{\mathrm{T}}\mathbf{1}_{d_k,S_{k,m}})^2\asymp d_k$ with over 95% probability.

To conclude, when there exists a column in $\mathbf{A}_k$ with mean of elements to be non-zero, then $s_{k,max}\asymp d_k^2$ can be easily achieved with any sample. If all columns in $\mathbf{A}_k$ have zero means (i.e. half positive and half negative elements which sums to zero), then we have $s_{k,max}\asymp d_k$ with high probability when $M_{k,0}\geq 186$. The value of $M_{k,0}$ in practice to achieve $s_{k,\max}\asymp d_k$ should be smaller than 186 since the constant 5.1σ above in (3.16) can be made smaller. In fact, in practice, we find that around $M_{k,0}=15$ does a perfect job in all our simulation settings in securing $s_{k,\max}\asymp d_k$. It means that with $K=3$, say we are estimating $\mathbf{A}_2$, then $M_0=M_{1,0}M_{3,0}=225$ works fine for securing $s_{\text{-}2,\max}\asymp d_1d_3=d_{\text{-}2}$. Indeed in all simulation settings, we use $M_0=200$ for $K=2$ or 3 and get very good performance overall.

We do not suggest explicit tuning of M_0, as our pre-averaging estimator is an initial estimator for feeding our iterative projection procedure. Simulation experiments in Section 3.4.2 has clearly shown that the practical performance of our iterative projection estimator remains at a good constant level no matter the initial M_0 we use.

3.2.5 Theoretical results for the pre-averaging estimator

In Section 3.2.3, we choose $S_{\text{-}k,\max}$ for summing the columns of $\mathrm{mat}_k(\mathscr{X}_t)$. To create stabler estimators, we can construct M different sets $S^{(m)}_{\text{-}k,\max}\subseteq[d_{\text{-}k}]$, $m\in[M]$ (we set $M=5$ in all our simulations), by choosing the best M from the M_0 samples in the procedure laid out in Section 3.2.3, and form $\widetilde{\mathbf{X}}_{k,1},\ldots,\widetilde{\mathbf{X}}_{k,M}$, where each $\widetilde{\mathbf{X}}_{k,i}$ is defined in (3.12). Note that we choose M to be a small number compared to M_0, ensuring that we only utilize samples with the largest signal accumulation. This approach leads to the best estimation accuracy of the pre-averaging estimator, as defined below. Let

$$\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k,agg}:=\frac{1}{M}\sum_{m=1}^{M}\frac{\widetilde{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{X}}_{k,m}}{T}. \tag{3.17}$$

The pre-averaging estimator $\widehat{\mathbf{Q}}_{k,pre,(z_k)}$ is defined as the z_k eigenvectors corresponding to the z_k largest eigenvalues of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k,agg}$, with the constraint $\widehat{\mathbf{Q}}_{k,pre,(z_k)}^{\mathrm{T}}\widehat{\mathbf{Q}}_{k,pre,(z_k)}=\mathbf{I}_{z_k}$, for

any $z_k \leq r_k$. The theoretical properties of $\widehat{\mathbf{Q}}_{k,pre,(z_k)}$ can be summarized in the following theorem.

Theorem 3.1. *Let M represent the number of selected samples from all M_0 samples, following the procedure outlined in Section 3.2.3. Let Assumption (E1), (E2), (F1), (L1), (L2), (R1), (R2) be satisfied for all M chosen random samples for constructing $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k,agg}$, and $r_{e,k} = O(d_k)$. Then*

$$\left\|\widehat{\mathbf{Q}}_{k,pre,(z_k)} - \mathbf{Q}_k\mathbf{H}_k\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}} c_{k,\max}\right), \quad \textit{where} \tag{3.18}$$

$$c_{k,\max} := \min\left\{1 + \frac{d_k}{T}, \frac{r_k d_k}{T}\right\}\frac{d_{\text{-}k}}{s_{\text{-}k,\max}} + d_k^{\alpha_{k,1}}\left(1 + \frac{d_k^2}{T^2}\right)\frac{d_{\text{-}k}^2}{s_{\text{-}k,\max}^2},$$

$$\mathbf{H}_k := T^{-1}\mathbf{D}_k^{1/2}\frac{1}{M}\sum_{m=1}^{M}\left[\widetilde{\mathbf{F}}_{k,m}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}\right]\mathbf{A}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k,pre,(z_k)}\widetilde{\mathbf{V}}_k^{-1}, \quad \textit{with}$$

$$\widetilde{\mathbf{F}}_{k,m} := (\mathrm{mat}_k(\mathscr{F}_{1,\text{-}k})\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,\max}^{(m)}}, \ldots, \mathrm{mat}_k(\mathscr{F}_{T,\text{-}k})\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,\max}^{(m)}}),$$

where $\mathbf{H}_k$ is a rotation matrix with rank($\mathbf{H}_k$) $= z_k$, and $\widetilde{\mathbf{V}}_k$ is diagonal, containing the z_k eigenvalues (in decreasing order) of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k,agg}$. Moreover, further assuming (L1'), there exists $\widehat{\mathbf{U}}_{k,pre,(z_k)}$ with $\widehat{\mathbf{U}}_{k,pre,(z_k)}^{\mathrm{T}}\widehat{\mathbf{U}}_{k,pre,(z_k)} = \mathbf{I}_{z_k}$ such that $\widehat{\mathbf{Q}}_{k,pre,(z_k)} = \widehat{\mathbf{U}}_{k,pre,(z_k)}\mathbf{P}_{k,pre,(z_k)}$ with $\mathbf{P}_{k,pre,(z_k)}$ being an orthogonal matrix, so that

$$\left\|\widehat{\mathbf{U}}_{k,pre,(z_k)} - \mathbf{U}_{k,(z_k)}\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}}\left[d_k^{2\alpha_{k,1}}\frac{r_k}{T} + c_{k,\max}\right]\right). \tag{3.19}$$

The matrix $\mathbf{U}_{k,(z_k)}$ is defined to be the matrix consisting of the first z_k columns of $\mathbf{U}_k$.

In the above theorem, Assumptions (E1), (E2), (F1), (L1), (R1), (R2) are naturally satisfied for any sample with $|S_{j,m}| = \lfloor d_j/2 \rfloor$, $j \in [d_{\text{-}k}]$, and Assumption (L2) will be satisfied for the M chosen samples by following the algorithm outlined in Section 3.2.3 with a suitable choice for M_0, as discussed in Section 3.2.4. The meanings for (3.18) and (3.19) are different. When $z_k < r_k$, (3.18) suggests that the estimated directions $\widehat{\mathbf{Q}}_{k,pre,(z_k)}$ will lie in the subspace spanned by the columns of $\mathbf{Q}_k$ (or $\mathbf{U}_k$), but it may not be "close" to the directions corresponding to the strongest z_k factors. However, with (3.19), we can conclude that $\widehat{\mathbf{U}}_{k,pre,(z_k)}$ will be "close" to the directions which correspond to the strongest z_k factors. As a compromise, (3.19) involves an extra rate $d_k^{2(\alpha_{k,1}-\alpha_{k,z_k})} r_k T^{-1}$ as compared to (3.18). Such a difference is especially notable when we set $z_k = 1$ and perform the iterative projection in Section 3.3. In addition, (3.19) requires an additional assumption (L1') that all population eigenvalues are distinct.

Remark: Suppose in (L2), the ratio $d_{-k}/s_{-k,max}$ is of order d_{-k}^{-1}, which can be achieved if, for instance, there exists a dense column in $\mathbf{A}_j$ (i.e., pervasive factor) having majority of elements of the same sign for each $j \in [K]$. Suppose further that the r_k's and K are constants, with $d_k \asymp T$ for each $k \in [K]$. The results from Theorem 3.1 implies that the projection matrix $\widehat{\mathbf{P}}_{k,pre} := \widehat{\mathbf{Q}}_{k,pre,(r_k)}\widehat{\mathbf{Q}}_{k,pre,(r_k)}^{\mathrm{T}}$ has error rate

$$\begin{aligned}\left\|\widehat{\mathbf{P}}_{k,pre} - \mathbf{Q}_k(\mathbf{Q}_k^{\mathrm{T}}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^{\mathrm{T}}\right\| &= \left\|\widehat{\mathbf{P}}_{k,pre} - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\right\| \\ &= O_P(d_k^{-\alpha_{k,r_k}}(d_{-k}^{-1/2} + d_k^{\alpha_{k,1}/2}d_{-k}^{-1})).\end{aligned} \tag{3.20}$$

This can be compared to the rates in Chen et al. (2022), which need the errors to be sub-Gaussian (compared to our Assumption (R1) where only bounded fourth moments is needed). While their σ^2 can be considered constant, their λ is such that $\lambda \asymp \prod_{k=1}^K d_k^{\alpha_{k,1}}$. The TIPUP procedure has rate (in our notations, using equation (47) in Chen et al. (2022), which has a faster rate of convergence than TOPUP)

$$\left\|\widehat{\mathbf{P}}_k - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\right\| = O_P\left(\frac{d_k^{1/2}}{T^{1/2}\prod_{k=1}^K d_k^{\alpha_{k,1}/2}} + \frac{d^{1/2}}{T^{1/2}\prod_{k=1}^K d_k^{\alpha_{k,1}}}\right). \tag{3.21}$$

When all factors are strong, i.e., $\alpha_{k,j} = 1$, the rate in (3.20) is faster than that in (3.21). When $\alpha_{k,1} = 1$ and $\alpha_{k,r_k} = 0.5$, i.e., the strongest factor is pervasive but the weakest factor is quite weak, then the two rates will be the same.

The rate in (3.20) can also be compared to Theorem 1 of Chen and Fan (2021) when $K = 2$, which under the same conditions laid out at the start of the remark, implies

$$\left\|\widehat{\mathbf{P}}_k - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\right\| = O_P(d_k^{-1/2}). \tag{3.22}$$

Our rate in (3.20) is $d_k^{-3/2}$ when all factors are strong, and is d_k^{-1} when $\alpha_{k,1} = 1$ and $\alpha_{k,r_k} = 0.5$. Both rates are faster than $d_k^{-1/2}$ in (3.22).

Indeed, the better performance of the iterative projection estimator, which uses the pre-averaging estimator as an initial estimator, is reflected in the empirical results in Section 3.4.

3.2.6 A discussion on optimality

Our pre-averaging estimator achieves a minimax optimal rate under certain scenarios over a certain localized set. For simplicity, suppose we only take $M = 1$ in (3.17), and assume

the data has mean 0. It means from (3.1) that

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k,agg} &= \frac{1}{T}\widetilde{\mathbf{X}}_{k,m}^{\mathrm{T}}\widetilde{\mathbf{X}}_{k,m} = \mathbf{M}^* + \mathbf{H}, \text{ where} \\
\mathbf{H} &:= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{A}_k\mathbf{F}_t\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}} + \mathbf{E}_t\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{F}_t^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}) \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}(\mathbf{E}_t\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}} - E[\mathrm{diag}(\mathbf{E}_t\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}})]), \\
\mathbf{M}^* &:= \frac{1}{T}\sum_{t=1}^{T}\mathbf{A}_k\mathbf{F}_t\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{F}_t^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}} + \frac{1}{T}\sum_{t=1}^{T}E[\mathrm{diag}(\mathbf{E}_t\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}})],
\end{aligned}$$

with $\mathbf{F}_t := \mathrm{mat}_k(\mathscr{F}_t)$, $\mathbf{E}_t := \mathrm{mat}_k(\mathscr{E}_t)$ and $\mathbf{q} := \mathbf{1}_{d_{\text{-}k},S_{\text{-}k,\max}}/\left\|\mathbf{1}_{d_{\text{-}k},S_{\text{-}k,\max}}\right\|$ (normalizing it does not affect the eigenvectors). Assume also $\mathbf{E}_t$ has only i.i.d. entries with finite 4th order moments (i.e., $\mathscr{E}_t = \boldsymbol{\varepsilon}_t$ in (E1), each element having the same finite variance), so that $E(\mathbf{H}) = \mathbf{0}$, and $T^{-1}\sum_{t=1}^{T}E[\mathrm{diag}(\mathbf{E}_t\mathbf{q}\mathbf{q}^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}})] = \sigma_\varepsilon^2\mathbf{I}_{d_k}$, where $\sigma_\varepsilon^2 = \mathrm{var}((\mathbf{E}_t)_{ij})$.

Let λ_j^* be the j-th largest eigenvalue of $\mathbf{M}^*$. The set of eigenvectors for $\mathbf{M}^*$ now coincides with the columns in $\mathbf{U}_k$ defined in (3.5), and we write $\mathbf{u}_j^*$ to be the j-th column of $\mathbf{U}_k$. Following equation (20c) in Cheng et al. (2021), define

$$\begin{aligned}
\mathscr{M}(\mathbf{M}^*) := \Big\{ & A \in \mathbb{R}^{d_k\times d_k} \text{ symmetric} \mid \mathrm{rank}(A) = r_k, \\
& \lambda_i(A) = \lambda_i^* \ \ (1 \le i \le r_k),\ \left\|\mathbf{u}_j(A) - \mathbf{u}_j^*\right\| \le \frac{c\sigma_{\min}\sqrt{d_k}}{|\lambda_j^*|} \Big\},
\end{aligned}$$

where $u_j(A)$ is the eigenvector corresponding to the j-th largest eigenvalue of A, and $\sigma_{\min}^2$ is the smallest value amongst of the variance of the elements of $\mathbf{H}$.

We can easily show that, as $T \to \infty$,

$$\lambda_j^* \asymp_P d_k^{\alpha_{k,j}}\mathbf{q}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{q} \asymp \frac{d_k^{\alpha_{k,j}}s_{\text{-}k,\max}}{d_{\text{-}k}}, \ \ j \in [r_k]; \ \ \sigma_{\min} \asymp \sqrt{\frac{\mathbf{q}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{q}}{T}} \asymp \sqrt{\frac{s_{\text{-}k,\max}}{Td_{\text{-}k}}}.$$

Then the conditions in Theorem 3 of Cheng et al. (2021) are satisfied, except that the elements of $\mathbf{H}$ are at most asymptotically normal as $T \to \infty$, and are dependent in general. The conclusion of the theorem is that

$$\inf_{\widehat{\mathbf{u}}_j}\sup_{A\in\mathscr{M}(\mathbf{M}^*)} E\left\|\widehat{\mathbf{u}}_j - \mathbf{u}_j(A)\right\| \ge \frac{C\sigma_{\min}\sqrt{d_k}}{|\lambda_j^*|} \asymp \frac{1}{d_k^{\alpha_{k,j}}}\sqrt{\frac{d_k}{T\mathbf{q}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{q}}} \asymp \frac{1}{d_k^{\alpha_{k,j}}}\sqrt{\frac{d}{Ts_{\text{-}k,\max}}}.$$

Similar to the remark at the end of Section 3.2.5, suppose $s_{-k,\max} \succeq d_{-k} d_k^{\alpha_{k,j}}$, which can be achieved if there exists a column in $\mathbf{A}_\ell$ having "enough" elements of the same sign for each $\ell \in [K]$ (if all are of the same sign, then $s_{-k,\max} \asymp d_{-k}^2$, which can be much larger than $d_{-k} d_k^{\alpha_{k,j}}$). Suppose also r_k and K are constants, and $d_\ell \asymp T$ for each $\ell \in [K]$. Then the minimax rate above is $d_k^{-\alpha_{k,j}} (d_{-k}/s_{-k,\max})^{1/2}$ for $j \in [r_k]$, which coincides with the rate from Theorem 3.1 for the pre-averaging estimator when $z_k = j \le r_k$:

$$\begin{aligned} \big\|\widehat{\mathbf{Q}}_{k,pre,(j)} - \mathbf{Q}_k \mathbf{H}_k\big\| &= O_P(d_k^{-\alpha_{k,j}} (d_{-k}/s_{-k,\max})^{1/2}) \text{ for } j \in [r_k], \text{ implying} \\ \big\|\widehat{\mathbf{P}}_{k,pre} - \mathbf{U}_k \mathbf{U}_k^{\mathrm{T}}\big\| &= O_P(d_k^{-\alpha_{k,r_k}} (d_{-k}/s_{-k,\max})^{1/2}), \end{aligned}$$

where $\widehat{\mathbf{P}}_{k,pre} := \widehat{\mathbf{Q}}_{k,pre,(r_k)} \widehat{\mathbf{Q}}_{k,pre,(r_k)}^{\mathrm{T}}$. Define $\mathbf{U}(A) := (\mathbf{u}_1(A), \ldots, \mathbf{u}_{r_k}(A))$, then

$$\begin{aligned} \sup_{A \in \mathscr{M}(\mathbf{M}^*)} \big\|\widehat{\mathbf{P}}_{k,pre} - \mathbf{U}(A)\mathbf{U}(A)^{\mathrm{T}}\big\| &\le \big\|\widehat{\mathbf{P}}_{k,pre} - \mathbf{U}_k \mathbf{U}_k^{\mathrm{T}}\big\| + \sup_{A \in \mathscr{M}(\mathbf{M}^*)} \big\|\mathbf{U}(A) - \mathbf{U}_k\big\| \\ &= O_P(d_k^{-\alpha_{k,r_k}} (d_{-k}/s_{-k,\max})^{1/2}). \end{aligned}$$

3.3 Re-estimation by Projection

While Yu et al. (2022), He et al. (2023a) and Barigozzi et al. (2023b) all deal with projection estimation of a factor loading matrix in the case of $K = 2$ or a general K, they all assume that all factors are pervasive. And in practice, they need to know the number of factors r_k in $\mathbf{A}_k$ for each $k \in [K]$ first in order to estimate a projection matrix $\mathbf{B}_k$ of size $d_{-k} \times r_{-k}$, where $r_{-k} := r/r_k$ with $r = r_1 \cdots r_K$.

In contrast, our projection method to be presented here does not need the estimation of each r_k first. This is because in our method, we are projecting in one direction only: the direction of the *strongest* factors, iteratively. Setting $z_k = 1$, the pre-averaging vector $\widehat{\mathbf{Q}}_{k,pre,(1)}$ is indeed asymptotically pointing to the direction of the strongest factors (see (3.19) in Theorem 3.1).

Projecting to the direction of the strongest factors is needed in our setting since there are weak factors. Their estimators have worse rate of convergence and estimation performance than pervasive ones. Using these worse estimated directions for projections will deteriorate the performance of the projection estimators. In Section 3.4, we demonstrated that under the presence of weak factors, our method provides the best performance of factor loading matrix estimation compared to all other state-of-the-art methods, including the projection estimation suggested by these three papers.

In (3.6), we demean the data first and change the projection direction to $\mathbf{q}_{\text{-}k}$, where

$$\mathbf{q}_{\text{-}k} := \mathbf{q}_K \otimes \cdots \otimes \mathbf{q}_{k+1} \otimes \mathbf{q}_{k-1} \otimes \cdots \otimes \mathbf{q}_1, \text{ with } \mathbf{q}_k := \mathbf{A}_k \mathbf{c}_k,\ k \in [K],$$

for some non-zero constant vectors $\mathbf{c}_k$. Then defining $\mathbf{c}_{\text{-}k} := \mathbf{c}_K \otimes \cdots \mathbf{c}_{k+1} \otimes \mathbf{c}_{k-1} \otimes \cdots \otimes \mathbf{c}_1$, we have $\mathbf{q}_{\text{-}k} = \mathbf{A}_{\text{-}k}\mathbf{c}_{\text{-}k}$, and we can construct the new projected data as

$$\begin{aligned} \mathbf{y}_t^{(k)} :&= \text{mat}_k(\mathscr{X}_t - \bar{\mathscr{X}})\mathbf{q}_{\text{-}k} \\ &= \mathbf{A}_k \text{mat}_k(\mathscr{F}_t - \bar{\mathscr{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{c}_{\text{-}k} + \text{mat}_k(\mathscr{E}_t - \bar{\mathscr{E}})\mathbf{q}_{\text{-}k}. \end{aligned} \tag{3.23}$$

Depending on the direction $\mathbf{c}_{\text{-}k}$, we can see from above that the signals from the factors are strengthened due to the term $\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{c}_{\text{-}k}$, while the noise level is retained or strengthened, depending on the level of cross-correlations among the noise fibres.

The projected data can also be used to estimate a finer projection direction, essentially iterating the projection step. See Theorem 3.2 below and the explanations followed. See simulation results regarding this in Section 3.4 as well.

3.3.1 Refining the projection direction

From Theorem 3.1, setting $z_k = 1$ there, we obtain $\widehat{\mathbf{q}}_{k,pre} := \widehat{\mathbf{Q}}_{k,pre,(1)} = \widehat{\mathbf{U}}_{k,pre,(1)}\mathbf{P}_{k,pre,(1)} = \pm\widehat{\mathbf{U}}_{k,pre,(1)}$ (WLOG we take the plus sign in the presentations hereafter). For each $k \in [K]$, we create the projected data $\mathbf{y}_t^{(k)}$ as in (3.23), using

$$\mathbf{q}_{\text{-}k} = \widehat{\mathbf{q}}_{\text{-}k,pre} := \widehat{\mathbf{q}}_{K,pre} \otimes \cdots \otimes \widehat{\mathbf{q}}_{k+1,pre} \otimes \widehat{\mathbf{q}}_{k-1,pre} \otimes \cdots \otimes \widehat{\mathbf{q}}_{1,pre}. \tag{3.24}$$

Then we define $\breve{\mathbf{q}}_k^{(1)}$ to be the eigenvector corresponding to the largest eigenvalue of the matrix

$$\widetilde{\mathbf{\Sigma}}_y^{(k)} := T^{-1}\sum_{t=1}^{T} \mathbf{y}_t^{(k)}\mathbf{y}_t^{(k)\mathrm{T}}.$$

The superscript (1) in $\breve{\mathbf{q}}_k^{(1)}$ signals that this is the first iterated estimator for $\mathbf{U}_{k,(1)}$. We can iterate this process to obtain refinement of projection direction. More formally, we introduce the following algorithm.

Algorithm for Iterative Projection Direction Refinement

1. Initialize $\breve{\mathbf{q}}_k^{(0)} = \widehat{\mathbf{q}}_{k,pre}$ for each $k \in [K]$.

2. For $i \geq 1$, at the i-th step, create projected data $\mathbf{y}_{t,i}^{(k)} := \mathrm{mat}_k(\mathscr{X}_t - \bar{\mathscr{X}})\check{\mathbf{q}}_k^{(i-1)}$ for each $k \in [K]$.

3. For each $k \in [K]$, define $\check{\mathbf{q}}_k^{(i)}$ the eigenvector corresponding to the largest eigenvalue of
$$\widetilde{\mathbf{\Sigma}}_{y,i}^{(k)} := T^{-1}\sum_{t=1}^{T}\mathbf{y}_{t,i}^{(k)}\mathbf{y}_{t,i}^{(k)\mathrm{T}}. \tag{3.25}$$

4. Replace i by $i+1$. Go back to step 2. Stop until after the procedure has been repeated for a fixed number of times.

We present a further assumption needed before presenting Theorem 3.2.

(RE1) *For a positive integer N, let $\mathscr{A}_{f,T} \in \mathbb{R}^{(N+1)T\times T}$ be defined as $\mathscr{A}_{f,T} := (\mathbf{a}_{f,1},\ldots,\mathbf{a}_{f,T})$, where*
$$\mathbf{a}_{f,t} := (\mathbf{0}_{t-1}^{\mathrm{T}}, a_{f,NT}, a_{f,NT-1},\ldots,a_{f,0},\mathbf{0}_{T-t}^{\mathrm{T}})^{\mathrm{T}},\quad t\in[T],$$
with $\mathbf{0}_j$ being a column vector of j zeros and the $a_{f,q}$'s are from Assumption (F1). Define $\mathscr{A}_{e,T}$ and $\mathscr{A}_{\varepsilon,T}$ similarly using coefficients from $\{a_{e,q}\}$ and $\{a_{\varepsilon,q}\}$ respectively from Assumption (E2). Then we assume that (with $\mathscr{A}$ can be either $\mathscr{A}_{f,T}, \mathscr{A}_{e,T}$ or $\mathscr{A}_{\varepsilon,T}$) $\left\|\mathscr{A}\right\|$ is uniformly bounded above, and
$$\begin{aligned}&\frac{1}{T}\mathrm{tr}(\mathscr{A}^{\mathrm{T}}\mathscr{A}) = 1 - o(T^{-2}d^{-2}),\\ &\frac{1}{T}\mathrm{tr}(\mathscr{A}^{\mathrm{T}}\mathscr{A})^2 \to a_1,\quad \frac{1}{T^2}\mathbf{1}_T^{\mathrm{T}}(\mathscr{A}^{\mathrm{T}}\mathscr{A})^2\mathbf{1}_T \to a_2,\quad \frac{1}{T^{3/2}}\mathbf{1}_T^{\mathrm{T}}\mathscr{A}^{\mathrm{T}}\mathscr{A}\mathbf{1}_T \to a_3,\end{aligned}$$
where $\mathbf{1}_T$ is a column vector of T ones, and the constants a_1, a_2 and a_3 can be different for $\mathscr{A} = \mathscr{A}_{f,T}, \mathscr{A}_{e,T}$ and $\mathscr{A}_{\varepsilon,T}$ respectively.

Consider a truncated linear process $\{y_t\}_{t\in[T]}$, and the original process $\{\widetilde{y}_t\}_{t\in[T]}$,
$$\widetilde{y}_t = \sum_{q\geq 0} a_q z_{t-q},\quad y_t = \sum_{q=0}^{NT} a_q z_{t-q},\ \text{with}\ \ \mathrm{var}(\widetilde{y}_t) = 1,$$
where $\{z_t\}$ is a sequence of i.i.d. random variables. Construct the matrix $\mathscr{A}$ using $\{a_q\}$ similar to those in Assumption (RE1). Then $\mathscr{A}^{\mathrm{T}}\mathscr{A}$ contains the variance of $\{y_t\}$ on the diagonal, and lag-k autocovariance on the k-th off-diagonal. The rates in (RE1) are then controlling how fast the a_q's are going to 0, and how much serial dependence between the y_t's are allowed. In particular, general linear processes with absolutely summable

autocovariance sequence, short range dependent processes like ARMA models, satisfy the assumption.

Theorem 3.2. *Let all the assumptions in Theorem 3.1 be satisfied, together with (RE1). Let* $g_s := \prod_{j=1}^{K} d_j^{\alpha_{j,1}}$, $r_e := \prod_{k=1}^{K} r_{e,k}$. *Assume further that for each* $k \in [K]$, $r = O(dg_s^{-1})$, $r_e = o(T)$, $d_k = O(g_s) = (r_e + \sqrt{T/r})$. *Then*

$$\left\|\check{\mathbf{q}}_k^{(1)} - \mathbf{U}_{k,(1)}\right\| = O_P\left\{\sqrt{\frac{r}{T}} + g_s^{-1/2} b_k \sqrt{\frac{rd}{T}}\right\}, \quad \textit{where}$$
$$b_k = K\sqrt{\frac{r_{\max}}{T}} + \sum_{j=1; j\neq k}^{K} d_j^{-\alpha_{j,1}} c_{j,max}^{1/2} = o(1).$$

Furthermore, if $rdg_s^{-1} = o(T)$, *then for an integer* $m \geq 1$,

$$\left\|\check{\mathbf{q}}_k^{(m+1)} - \mathbf{U}_{k,(1)}\right\| = O_P\left\{\sqrt{\frac{r}{T}} + g_s^{-1/2}\left\|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\right\|\sqrt{\frac{rd}{T}}\right\} = o_P(1),$$

and the Algorithm for Iterative Projection Direction Refinement will produce, after a certain number of iterations (say m),

$$\left\|\check{\mathbf{q}}_k^{(m)} - \mathbf{U}_{k,(1)}\right\| = O_P\left(\sqrt{\frac{r}{T}}\right).$$

To put the above results into perspective, assume a very common scenario that $d_1 \asymp \cdots \asymp d_K \asymp T$ (this is especially true in economic applications where T is small), with K and each r_k being constants for $k \in [K]$. While Barigozzi et al. (2023b) demonstrate that their one-step iteration projection estimator is sufficient to achieve good performance, which is computationally fast, we similarly show that, in certain scenarios especially when all factors are strong, our iterative projection can achieve the optimal rate in a single step. We first note that if all factors in $\mathbf{A}_k$ are pervasive, i.e., $\alpha_{k,1} = 1$ for all $k \in [K]$, then $g_s = d$, and hence $\left\|\check{\mathbf{q}}_k^{(1)} - \mathbf{U}_{k,(1)}\right\| = O_P(T^{-1/2})$, and any refinements will retain the same rate. Even if $\alpha_{k,1} < 1$ (i.e., the strongest factor corresponding to $\mathbf{A}_k$ is not pervasive), $\left\|\check{\mathbf{q}}_k^{(1)} - \mathbf{U}_{k,(1)}\right\|$ can still be $O_P(T^{-1/2})$, as long as $b_k^2 d/g_s = O(1)$, equivalent to $\alpha_{k,1} \geq 1/2$. The case of $\alpha_{k,1} = 1/2$ presents a significantly weak strongest factor corresponding to $\mathbf{A}_k$, and without the help of projection and strong factors from other modes' factor loading spaces, the typical rate for estimating such a weak factor would be $d_k^{-1/4}$ which is much worse than $T^{-1/2}$.

If $b_k^2 d/g_s = O(1)$ is not satisfied (for example, when the strongest factors for some or all modes are weak), then we require more iterations to achieve the optimal rate. To have

an idea on the value of m, from the last part of the proof of Theorem 3.2, we need

$$b_k\left(\sqrt{\frac{rd}{Tg_s}}\right)^m = O\left(\sqrt{\frac{r}{T}}\right).$$

Suppose $d_k \asymp T$, r_k is a constant and $d_{-k}/s_{-k,\max} \asymp 1$ (see Section 3.2.4 on how to achieve this). Then $b_k \asymp d_k^{-\alpha_{k,1}/2}$, and hence

$$m \geq \frac{\text{constant} + \alpha_{k,1}\log(d_k) - \log(T)}{\log(\frac{rd}{Tg_s})}.$$

Further, if $\alpha_{k,1} = 0.5$ (a very weak factor), and $d/g_s \asymp T^{0.95}$ (recall that we assume $rdg_s^{-1} = o(T)$), then as $T, d_k \to \infty$, we have $m \geq 10$. This is already quite extreme since $d/g_s \asymp T^{0.95}$ means that the strongest factors of some or all other $\mathbf{A}_k$'s are also weak. The fact that we are using $m = 30$ in our simulations in Section 3.4 throughout made sure that the rate $\sqrt{r/T}$ is reached, and we do not recommend users increase m further for saving computational time.

The fixed rate $O_P(\sqrt{r/T})$ in Theorem 3.2 comes from the fact that we need to distinguish the direction of the strongest factors from all other directions of weaker factors in order to find the "best" projection direction. In the case of studying the whole $\mathbf{U}_k$, we in fact may get a better rate of convergence even in the presence of weak factors.

Theorem 3.3. *Let all the assumptions in Theorem 3.2 be satisfied. Suppose we know the value of r_k, and perform an eigenanalysis on $\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}$ in (4.2) which utilized the projection direction $\check{\mathbf{q}}_k^{(m)}$ in Theorem 3.2, obtaining r_k eigenvectors as an estimator of the factor loading space of $\mathbf{A}_k$.*

Then there exists $\check{\mathbf{U}}_k \in \mathbb{R}^{d_k \times r_k}$ with $\check{\mathbf{U}}_k^{\mathrm{T}}\check{\mathbf{U}}_k = \mathbf{I}_{r_k}$ such that the r_k eigenvectors obtained above is $\check{\mathbf{U}}_k$ multiplied with some orthogonal matrix, with

$$\left\|\check{\mathbf{U}}_k - \mathbf{U}_k\right\| = O_P\left\{d_k^{\alpha_{k,1}-\alpha_{k,r_k}}\left[g_s^{-1} + \sqrt{\frac{r}{Tg_s}}\left(r_e^{1/2} + d_k^{1/2} + \sqrt{\frac{rd}{T}}\right)\right]\right\}.$$

Our projection estimator will consistently estimate the space spanned by $\mathbf{A}_k$ if the rate above is $o_p(1)$. Consider $d_1 \asymp \cdots \asymp d_K \asymp T$, with K and r_k being constants for $k \in [K]$. If all factors for $\mathbf{A}_k$ are pervasive, i.e., $\alpha_{k,j} = 1$ for all $j \in [r_k]$, then we have $\left\|\check{\mathbf{U}}_k - \mathbf{U}_k\right\| = O_P(T^{-1})$. When $K = 2$, this has the same rate as the average Frobenius error norm of the estimators of $\mathbf{A}_1$ and $\mathbf{A}_2$ in Theorem 3.1 and Theorem 4.1 of He et al. (2023a). This is also consistent with the rate in Corollary 3.1 of Barigozzi et al. (2023b),

Theorem 3.1 and 3.2 of He et al. (2022), and Theorem 3.1 of Yu et al. (2022) under the same scenario. When $K \geq 3$, under the same scenario, Corollary 3.1 of Barigozzi et al. (2023b) takes advantage and produces a faster rate of $O_P(T^{-3/2})$. However, if there exist some d_j with smaller order such that $d_j \asymp T^{1/2}$ for some j, then both Barigozzi et al. (2023b) and our Theorem 3.3 gives the same rate of $O_P(T^{-1})$.

The above rate in Theorem 3.3 can be greatly improved if the term $\sqrt{rd/T}$ can be removed. It is there because the estimated projection direction is correlated with the data in general. If we have independent noise tensor $\{\mathcal{E}_t\}$ (e.g., the setting in Chen et al. (2022)) we can split the data into half, and using only one half of it for projection direction estimation while the other half is for re-estimation only. Then the estimated projection direction will be independent of the re-estimation data, and hence the final rate indeed will be rid of this term (see Abadir et al. (2014); Lam (2016) for more discussions on sample splitting). When all the factors are strong, this improved rate will be the same as the one for TIPUP in equation (47) of Chen et al. (2022). We do not pursue this since our paper focuses on time series data with serial correlation in the noise. Moreover, the empirical performance of our projection method is very good already.

Remark: We have not included the asymptotic normality result for our factor loading estimators. This is because to estimate the covariance matrix in the asymptotic normality result, we somehow need to estimate the strength of different factors in the process, which poses another layer of practical and conceptual challenges. The accurate estimation of factor strength, and the inference of such, are important questions worthy of investigating as an independent topic since our NYC Taxi example in Section 4.4 has demonstrated that weak factors exist, and is indicative to follow procedures that do not assume pervasive factors in the first place. In Chapter 5, we propose a possible method to estimate factor strengths under certain identification conditions, which could pave the way for further research into the inference problem.

3.4 Simulation Experiments

In this section, we conduct simulation experiments to compare the performances of our iterative projection estimators (PROJ) to other state-of-the-art competitors. The pre-averaging estimator (PRE) is also presented with different M_0 and compared to PROJ.

3.4.1 Simulation settings

For generating our data, we use model (3.1), with elements in μ being i.i.d. standard normal in each repetition of experiment. For $k \in [K]$, each factor loading matrix $\mathbf{A}_k$ is generated independently with $\mathbf{A}_k = \mathbf{B}_k\mathbf{R}_k$, where the elements in $\mathbf{B}_k \in \mathbb{R}^{d_k \times r_k}$ are i.i.d. $U(u_1,u_2)$, and $\mathbf{R}_k \in \mathbb{R}^{r_k \times r_k}$ is diagonal with the j-th diagonal element being $d_k^{-\zeta_{k,j}}$, $0 \leq \zeta_{k,j} \leq 0.5$. Pervasive (strong) factors have $\zeta_{k,j} = 0$, while weak factors have $0 < \zeta_{k,j} \leq 0.5$.

The elements in $\mathscr{F}_t$ are independent standardized AR(5) with AR coefficients 0.7,0.3,-0.4,0.2 and -0.1. Same for the elements in $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$ in (3.2), but their AR coefficients are (-0.7,-0.3,-0.4,0.2,0.1) and (0.8,0.4, -0.4,0.2,-0.1) respectively. The standard deviation of each element of $\boldsymbol{\varepsilon}_t$ is randomly generated with i.i.d. $|\mathscr{N}(0,1)|$. Each entry of the matrices $\mathbf{A}_{e,k} \in \mathbb{R}^{d_k \times r_{e,k}}, k \in [K]$ is generated with i.i.d. standard normal, but has an independent probability of 0.7 being set exactly to 0. Each experiment is repeated 500 times. We consider the simulation settings (I), (II), (III) and (IV), with sub-settings (a) and (b), detailed below:

(Ia) Two strong factors with $r_k = 2$, $\zeta_{k,j} = 0$ for all k, j, and $u_1 = -2$, $u_2 = 2$ (elements in $\mathbf{A}_k$ have mean 0)).

(IIa) One strong factor and one weak factor with $r_k = 2$, $\zeta_{k,1} = 0$ and $\zeta_{k,2} = 0.2$ for all k; $u_1 = -2$, $u_2 = 2$.

(IIIa) Two weak factors with $r_k = 2$, $\zeta_{k,1} = 0.1$ and $\zeta_{k,2} = 0.2$ for all k; $u_1 = -2$, $u_2 = 2$.

(IVa) Four strong factors with $r_k = 4$, $\zeta_{k,j} = 0$ for all k, j; $u_1 = -2$, $u_2 = 2$.

Setting (Ib) to (IVb) are the same as (Ia) to (IVa) respectively, except that $u_1 = 0$, $u_2 = 2$, so that the elements in $\mathbf{A}_k$ have non-zero mean, leading to larger $s_{k,max}$.

Setting (I)(II)(III) and (IV) are designed to test the performance of estimation methods under different profiles of factor strengths. In Setting (I), we have two strong factors with $\alpha_{k,1} = \alpha_{k,2} = 1$ for each mode k, which is consistent with the pervasive factor assumptions of Barigozzi et al. (2023b); Chen and Fan (2021); He et al. (2023a, 2022); Yu et al. (2022). In Setting (II), $\alpha_{k,1} = 1$ and $\alpha_{k,2} = 0.6$, so the factor strengths differ and we have one strong factor and one weak factor. In Setting (III), even the strongest factor becomes weak, as $\alpha_{k,1} = 0.8$ and $\alpha_{k,2} = 0.6$. In Setting (IV), we may encounter what we refer to as 'pseudo weak factors'. This is because even if we generate four factors to be equally strong, the four population eigenvalues are likely to be separated, leading to an effect that certain factors seems to behave 'weaker' than the others.

In each Setting (I)-(IV), the distinction between sub-settings (a) and (b) is intended to highlight the impact of signal accumulation through pre-averaging. In sub-setting (a), the mean of each element of $\mathbf{A}_k$ is 0, whereas in sub-setting (b), all entries of $\mathbf{A}_k$ share the same sign. This difference potentially results in greater signal accumulation (larger $s_{k,max}$) through pre-averaging in sub-setting (b). For a more detailed analysis, please refer to the simulation results in Section 3.4.2.

To test the performance of different estimation methods under heavy-tailed distributions, we consider two distributions for the innovation processes of $\mathscr{F}_t$, $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$: 1) i.i.d. standard normal; 2) i.i.d. t_3. Thus, there are totally sixteen profiles considered. For all profiles above, we set $r_{e,k} = 2$ for all k. Note that the innovations t_3 do not have bounded fourth moments as required by Assumption (R1). We test the robustness of different methods under violations of this assumption.

3.4.2 Effect of M_0 in pre-averaging and projection

In the pre-averaging procedure, as described in Section 3.2, we generate a total of M_0 random samples. Following the algorithm outlined in Section 3.2.3, we then select the 'best' M samples from these generated samples, which achieve the largest signal accumulation, to construct the pre-averaging estimator. We test empirically the effect of using different M_0 in the pre-averaging procedure. We fix $M = 5$, and consider $M_0 = 200$, 400 and 800, respectively. For each M_0, we first calculate the pre-averaging estimator $\widehat{\mathbf{Q}}_{k,pre}$ for each $k \in [K]$, and obtain $\check{\mathbf{q}}_k^{(0)} = \widehat{\mathbf{q}}_{k,pre} = \widehat{\mathbf{Q}}_{k,pre,(1)}$. Then we calculate $\check{\mathbf{q}}_k^{(m)}$ according to the Algorithm for Iterative Projection Direction Refinement in Section 3.3.1 for $m = 1, \ldots, 29$. Finally, we obtain the iterative projection estimator $\check{\mathbf{U}}_k$ by utilising $\check{\mathbf{q}}_k^{(29)}$ as the projection direction for each $k \in [K]$.

In factor models, we can only estimate $\mathbf{A}_k$ up to rotations. Therefore, to evaluate the accuracy of factor loading estimators, we aim to compare the column spaces spanned by the columns of $\widehat{\mathbf{Q}}_k$ and $\mathbf{A}_k$. Recall that $\mathbf{U}_k$ is defined as the orthogonal basis of the true factor loading spaces, while $\widehat{\mathbf{Q}}_k$ represents the orthogonal basis of the estimated factor loading spaces. A natural measure of the distance between the column spaces of $\widehat{\mathbf{Q}}_k$ and $\mathbf{U}_k$ is given by $\|\widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k^{\mathrm{T}} - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\|$, i.e. the spectral norm of the difference between the estimated and true underlying projection matrices, which also equals the sine of the largest principle angle between the column spaces of $\widehat{\mathbf{Q}}_k$ and $\mathbf{U}_k$. Thus, a smaller value of $\|\widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k^{\mathrm{T}} - \mathbf{U}_k\mathbf{U}_k^{\mathrm{T}}\|$ indicates a more accurate estimator of the factor loading spaces. In all subsequent sections, we employ this commonly adopted measure to assess the accuracy of estimation (Chen

and Fan, 2021; Chen et al., 2022; Han et al., 2020), although other metrics to compare factor spaces are also possible (Barigozzi et al., 2023b; He et al., 2022).

For ease of presentation, we only display results under t_3 distributed errors for each setting, and consider the following three settings of different dimensions:

i. $K = 2$, $T = 100$, $d_1 = d_2 = 40$;

ii. $K = 2$, $T = 200$, $d_1 = d_2 = 80$;

iii. $K = 3$. $T = 200$, $d_1 = d_2 = d_3 = 15$.

We omit the results for $\mathbf{A}_2$ (and $\mathbf{A}_3$) and only display the estimation accuracy for $\mathbf{A}_1$ in Figures 3.1 to 3.4, since the results for $\mathbf{A}_1$ and $\mathbf{A}_2$ (and $\mathbf{A}_3$) are very similar.

From Figure 3.1 to 3.3, both the pre-averaging and iterative projection estimators have better estimation accuracy as T and d_k increase. Regarding the effect of M_0, the first thing to observe is that in each sub-setting (a), the initial pre-averaging estimator performs slightly better with larger M_0, while in sub-setting (b), M_0 does not significantly affect the accuracy of the pre-averaging estimator. This is natural, as the mean of each element of $\mathbf{A}_k$ in sub-setting (a) is 0 while it is non-zero in sub-setting (b), where pre-averaging estimators take advantage. In fact, in sub-setting (b), all entries of $\mathbf{A}_k$ have the same sign, so $s_{k,max} \asymp d_k^2$ (as the strongest factor in $\mathbf{A}_k$ is pervasive) can be easily achieved regardless of M_0. For sub-setting (a), around half of the entries of $\mathbf{A}_k$ are positive and the other half are negative, so it is more difficult to achieve the theoretical maximum order $s_{k,max} \asymp d_k^2$ (this happens when the majority of entries in the chosen sample S_k are of the same sign). However, in this case, we can still easily achieve $s_{k,max} \asymp d_k$ with a small M_0 (as discussed in Section 3.2.4), which is good enough to serve as an initial estimator for the Algorithm for Iterative Projection Direction Refinement.

The second thing to observe is that in most settings and dimensions, the differences in performance of the initial pre-averaging estimators (caused by changing M_0) do not affect the estimation accuracy of the subsequent iterative projection estimators. This is because in our iterative projection, we only need to utilize the strongest factor direction, which is estimated most accurately by the pre-averaging procedure (see Theorem 3.1 and Section 3.3.1 for more details). Hence, while increasing M_0 can lead to better performances of $\widehat{\mathbf{Q}}_{k,pre}$ as a whole, a small M_0 is usually sufficient to provide an accurate initial direction $\check{\mathbf{q}}_k^{(0)} = \widehat{\mathbf{q}}_{k,pre} = \widehat{\mathbf{Q}}_{k,pre,(1)}$ for projection. This is especially true in Setting (Ia) and (IIa) where the strongest factor is pervasive. The only exception is Setting (IIIa) with $K = 2$,

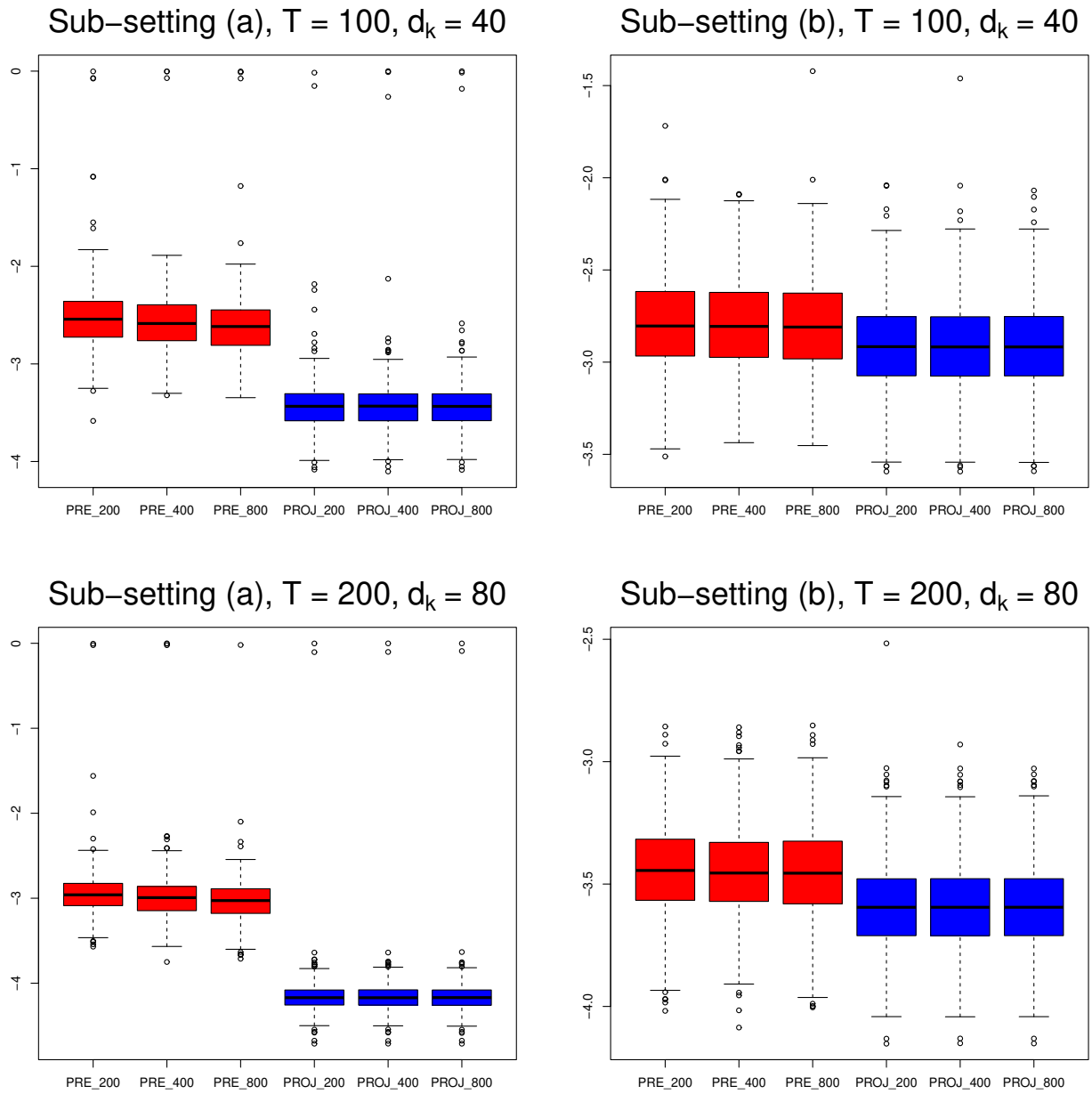


Fig. 3.1 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for Setting (I), $K = 2$. *Left:* Sub-setting (a). *Right:* Sub-setting (b).

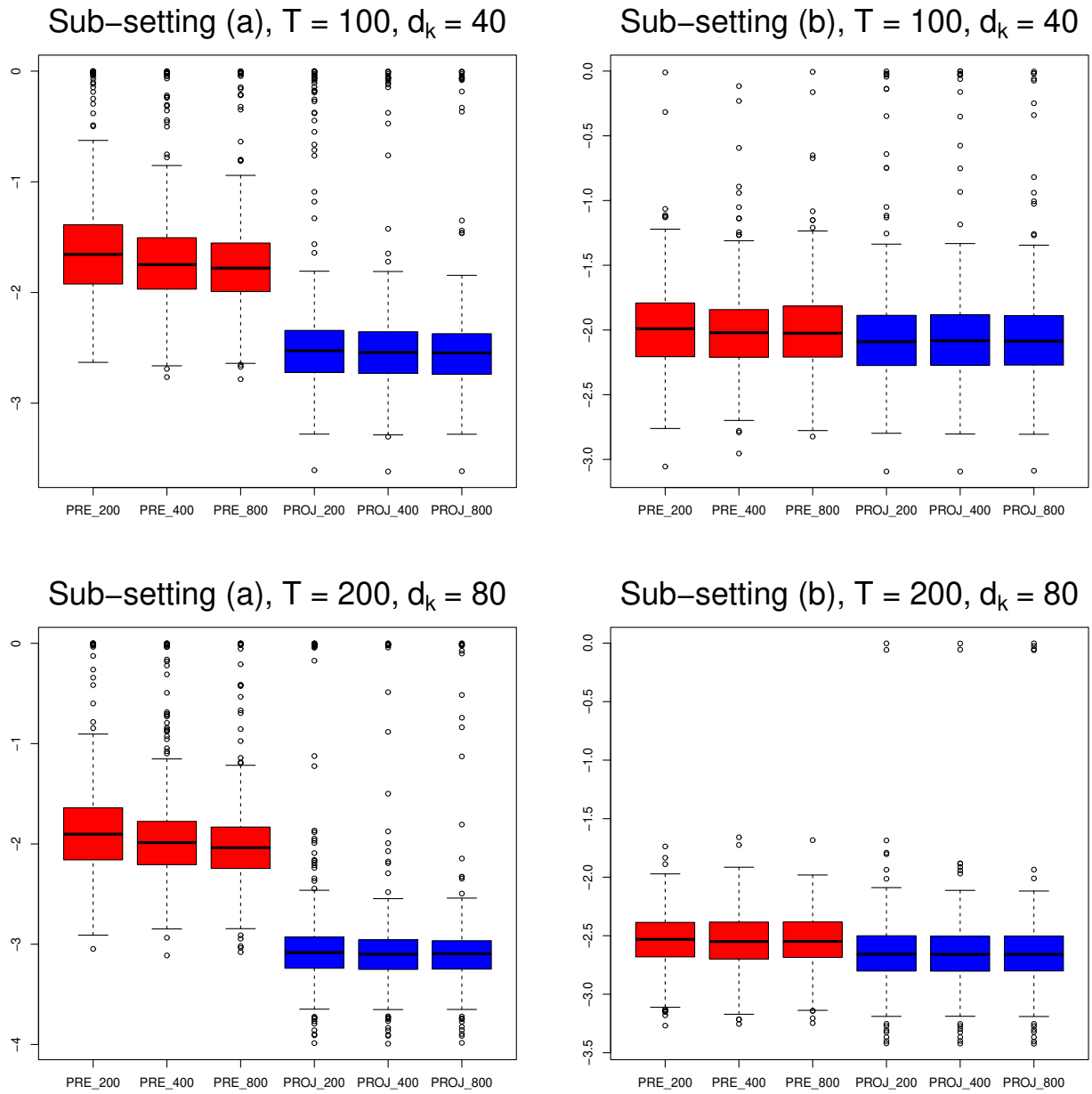


Fig. 3.2 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for Setting (II), $K = 2$. *Left:* Sub-setting (a). *Right:* Sub-setting (b).

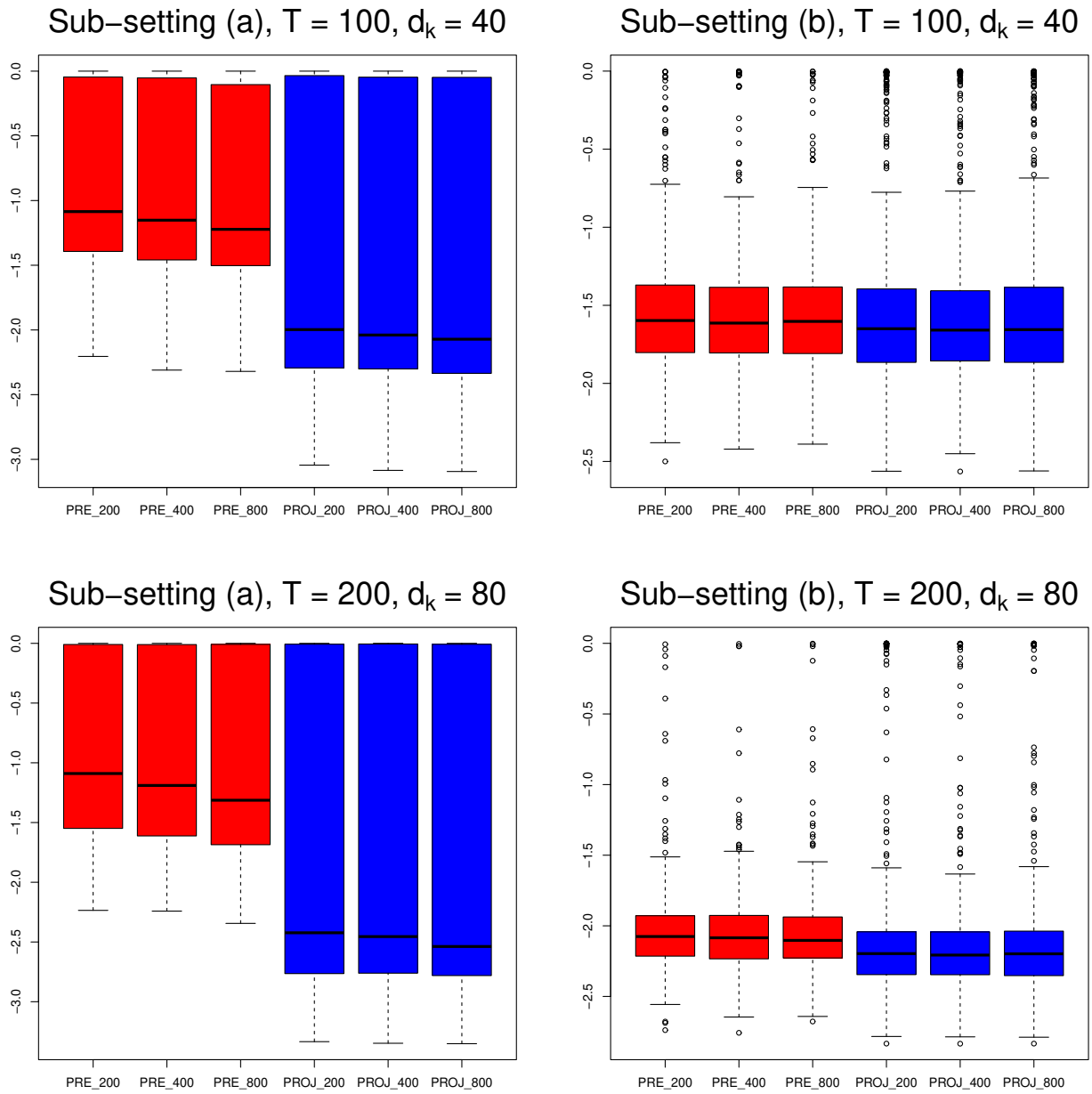


Fig. 3.3 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for Setting (III), $K = 2$. *Left:* Sub-setting (a). *Right:* Sub-setting (b).

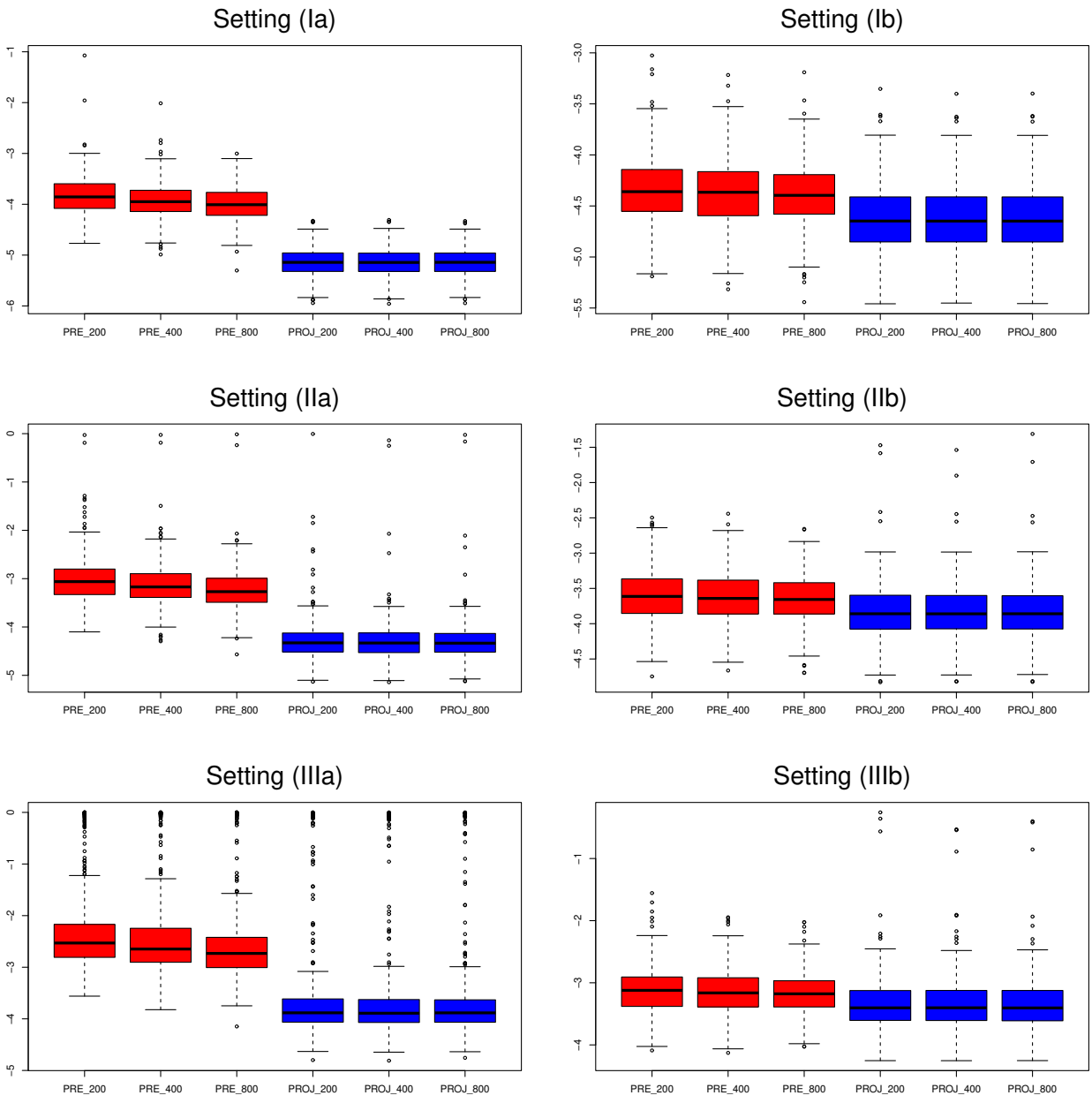


Fig. 3.4 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 3, d_1 = d_2 = d_3 = 15$.

where even the strongest factor is too weak to be estimated accurately, so increasing M_0 may lead to slightly better estimators after projection. However, such an issue is relieved when $K = 3$, since we have directions from more modes to be projected, which largely improves the estimation accuracy. In fact, a comparison between Figure 3.4 and Figure 3.1 to 3.3 reveals that both the pre-averaging and iterative projection estimator perform better as K increases.

Hence in general, increasing M_0 does not significantly affect the performance of our final estimator obtained by iterative projection in most scenarios. To save some computational time without sacrificing estimation accuracy, we use $M_0 = 200$ for the pre-averaging procedure in all of the subsequent simulations.

3.4.3 Comparison to state-of-the-art methods

We compare our iterative projection estimator (PROJ) in estimating the factor loading spaces with some state-of-the-art methods proposed by recent literature. All the twelve profiles in Section 3.4.1 and five settings of different dimensions are considered:

i. $K = 2, T = 100, d_1 = d_2 = 40$; ii. $K = 2, T = 200, d_1 = d_2 = 80$;

iii. $K = 3. T = 200, d_1 = d_2 = d_3 = 15$; iv. $K = 3. T = 200, d_1 = d_2 = d_3 = 25$;

v. $K = 4. T = 200, d_1 = d_2 = d_3 = d_4 = 15$.

When $K = 2$, the following methods designed for matrix-valued factor models are compared: The method of Wang et al. (2019) is TOPOP in Chen et al. (2022), but we omit its results since it performs much worse than iTIPUP in Han et al. (2020), which is the best one among the same type of estimators. The α-PCA estimator of Chen and Fan (2021) is implemented with $\alpha = 0$ (the performances for $\alpha \in \{-1, 0, 1\}$ are comparable according to Yu et al. (2022)). The projection method of Yu et al. (2022) and Barigozzi et al. (2023b) are referred to as PE (which is in the same spirit as HOOI). In addition, we also consider some robust procedures, including the robust tensor factor analysis (RTFA) proposed by He et al. (2022) and He et al. (2023a), and the Matrix Kendall's tau (MRTS) by He et al. (2022). For all the above methods which involve iterations, we set the number of iterations to be 30.

For settings with $K = 3$, we do not include α-PCA and MRTS, since they are only designed for $K = 2$. For the setting with $K = 4$, we further exclude RTFA in comparison, since it requires too much computational time, as can be observed in Table 3.1.

Figure 3.5 to 3.9 show the logarithm of estimation errors of $\mathbf{A}_1$ under the five different settings of dimensions for Setting (I)(II)(III), with normally and t_3 distributed errors, respectively. It can be seen that our iterative projection estimator (PROJ) generally outperforms all competitors in all settings and dimensions we consider, and is at least on par with other competitors.

More specifically, when $K = 2$, all methods perform reasonably well in sub-setting (a), but they all perform poorly in sub-setting (b) except our iterative projection estimator (PROJ). The only difference between sub-setting (a) and (b) is that the mean of each element of $\mathbf{A}_k$ in sub-setting (a) is 0 while it is non-zero in sub-setting (b). Whenever the mean of the elements are not 0, the pre-averaging estimator, and hence the iterative projection estimator, can take advantage since pre-averaging is based on summing rows of $\mathbf{A}_1$ (for estimating $\mathbf{A}_2$) or $\mathbf{A}_2$ (for estimating $\mathbf{A}_1$), which accumulates more signal when the sum is non-zero.

Regarding factor strengths, the advantage of PROJ is more notable in Setting (II) and (III), when other methods tend to give poorer estimates in the presence of weak factors in these settings. PE and RTFA perform on par with PROJ in Setting (Ia), (Ib) and (IIa), but they become less accurate in other settings. PROJ performs well when factor strengths differ, because it only projects onto the direction of the strongest factors, which are most accurately estimated in each iteration. In contrast, PE and RTFA project onto all factor directions, including those for weak factors that are poorly estimated. Consequently, using these less accurate directions for projection can degrade the performance of their estimators. Our method is also robust to heavy-tailed errors, and perform better than the robust procedure RTFA and MRTS in all scenarios. When $K = 3$ (and 4 as well), most methods take advantage of a larger K and perform better. Our iterative projection estimator still performs better than, or at least on par with all competitors.

Finally, Figure 3.10 compares the performance of different estimation methods under Setting (IV) with four strong factors. As discussed in Section 3.4.1, in such scenarios, the four population eigenvalues are likely to be separated, leading to an effect of 'pseudo weak factors'. Figure 3.10 shows that the performances of iTIPUP, RTFA, and α-PCA suffer from this effect of 'pseudo weak factors', compared to their performances in Setting (I) with only two strong factors. In contrast, PROJ and PE still perform well in this case.

To conclude, in general, pre-averaging estimator takes advantage of non-zero means in a factor loading matrix, and our iterative projection estimator performs better when there are weak factors, while on par with other estimators designed for pervasive factors when all factors are indeed pervasive. While integrating our pre-averaging or projection algorithm

with other iterative methods is feasible, we recommend using them together, particularly when dealing with tensor factor models containing weak factors. This is because our projection method is specifically tailored to handle the presence of weak factors, while the pre-averaging procedure is intended to generate the optimal initial direction for projection.

Table 3.1 records the average computational time of factor loading estimations under different methods and dimensions. For PROJ, we record the total computational time of the initial pre-averaging procedure and the iterative projection algorithm. It can be seen that the computational cost of our method is generally on par with iTIPUP and PE, while our method can be faster when d_{-k} is small, but become slower when d_{-k} grows large. Hence, in general, our method is more suitable for application when d_{-k} is not excessively large. The main reason for the increasing computational cost of our method with d_{-k} is that we need to sample over the d_{-k} fibers during the pre-averaging process. However, our iterative projection algorithm is fast on its own because it only involves the projection of a vector, whereas other iterative methods require the projection of a matrix. For other methods, MRTS is extremely computational expensive, and it gives poor estimation accuracy as well. α-PCA is the fastest because it does not require any iterations, but it also performs poorly as can be seen in Figure 3.5 and 3.6. The computational cost of RTFA also grows quickly when K and d_k increase, and it does not give good estimates when $K = 3$ (see Figure 3.7 and 3.8), which prompt us to not consider using it for $K = 4$.

		PROJ	iTIPUP	PE	α-PCA	RTFA	MRTS
$K = 2$	$T = 100, d_k = 40$	1.15	2.01	1.99	0.08	3.88	9.32
	$T = 200, d_k = 80$	4.63	4.63	4.55	0.32	15.65	49.73
$K = 3$	$T = 100, d_k = 15$	3.26	5.79	5.77	/	22.72	/
	$T = 200, d_k = 25$	10.66	10.05	10.04	/	151.65	/
$K = 4$	$T = 200, d_k = 15$	43.80	34.16	33.69	/	/	/

Table 3.1 Mean of the run time (in seconds) for factor loading estimations under different methods and different dimensions.

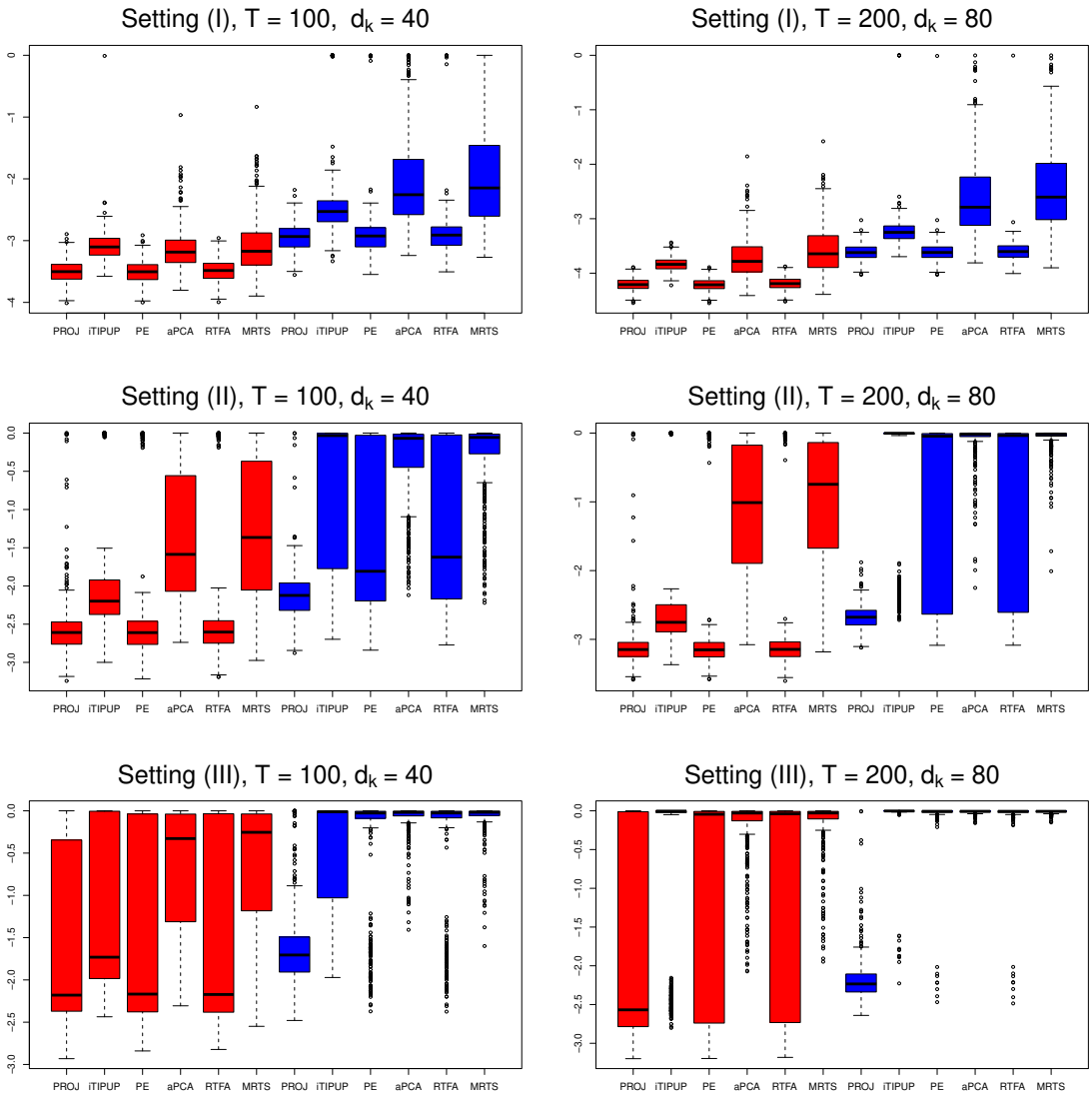


Fig. 3.5 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 2$, normally distributed errors. In each panel, the left six boxplots (in red) represent sub-setting (a), while the right six boxplots (in blue) represent sub-setting (b).

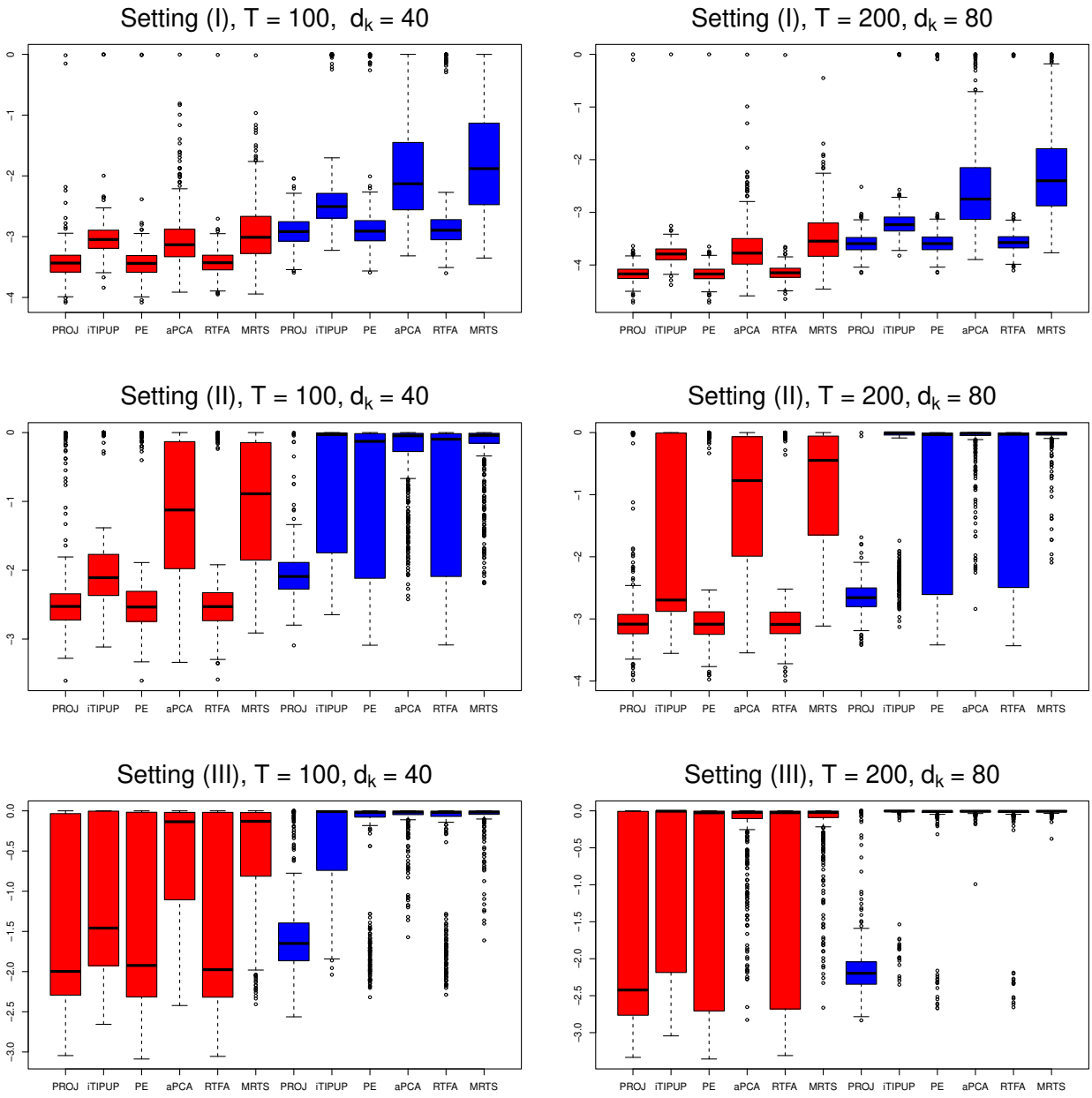


Fig. 3.6 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 2$, t_3-distributed errors. In each panel, the left six boxplots (in red) represent sub-setting (a), while the right six boxplots (in blue) represent sub-setting (b).

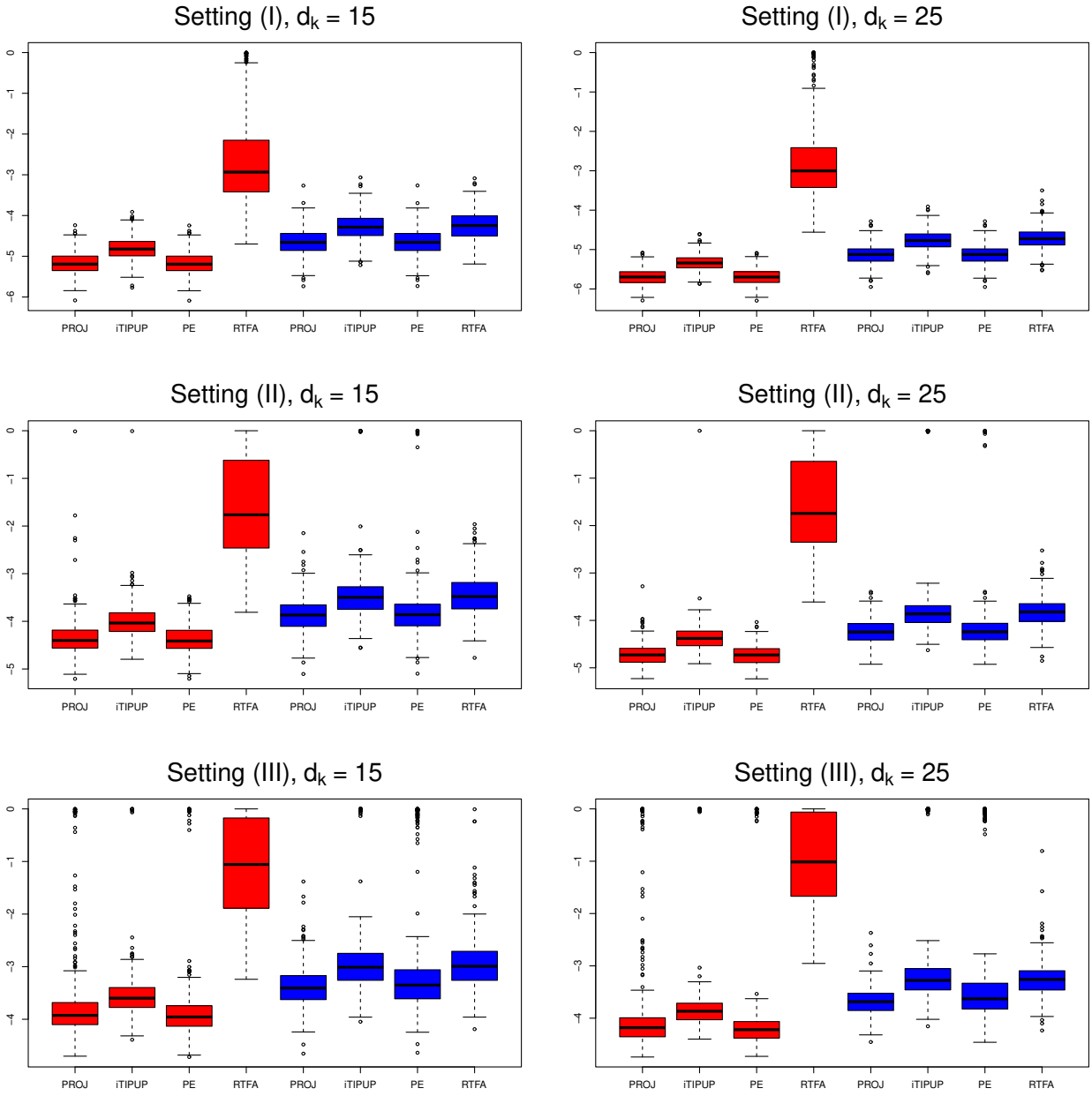


Fig. 3.7 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 3$, normally distributed errors. In each panel, the left four boxplots (in red) represent sub-setting (a), while the right four boxplots (in blue) represent sub-setting (b).

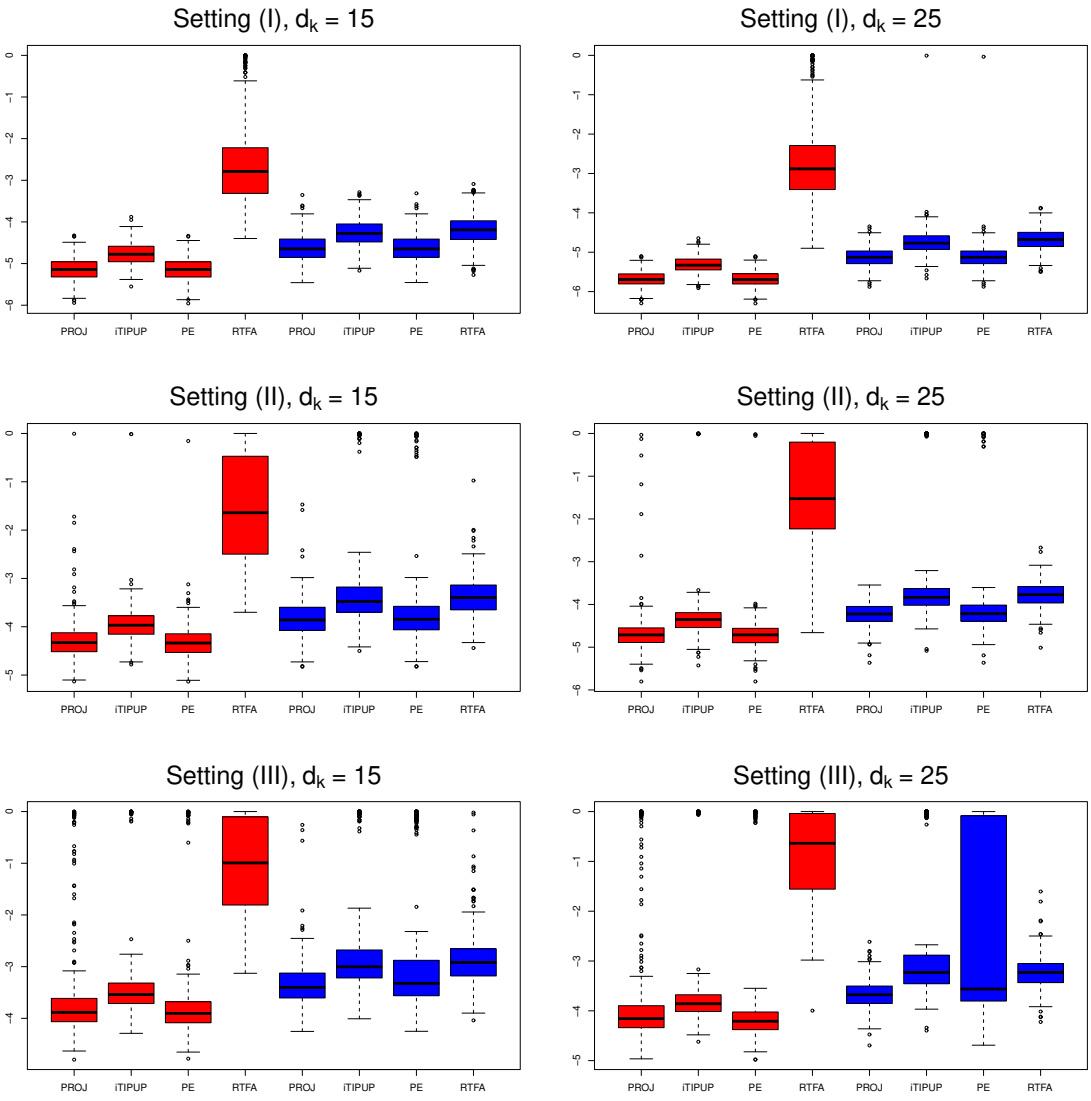


Fig. 3.8 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 3, T = 200$, t_3-distributed errors. In each panel, the left four boxplots (in red) represent sub-setting (a), while the right four boxplots (in blue) represent sub-setting (b).

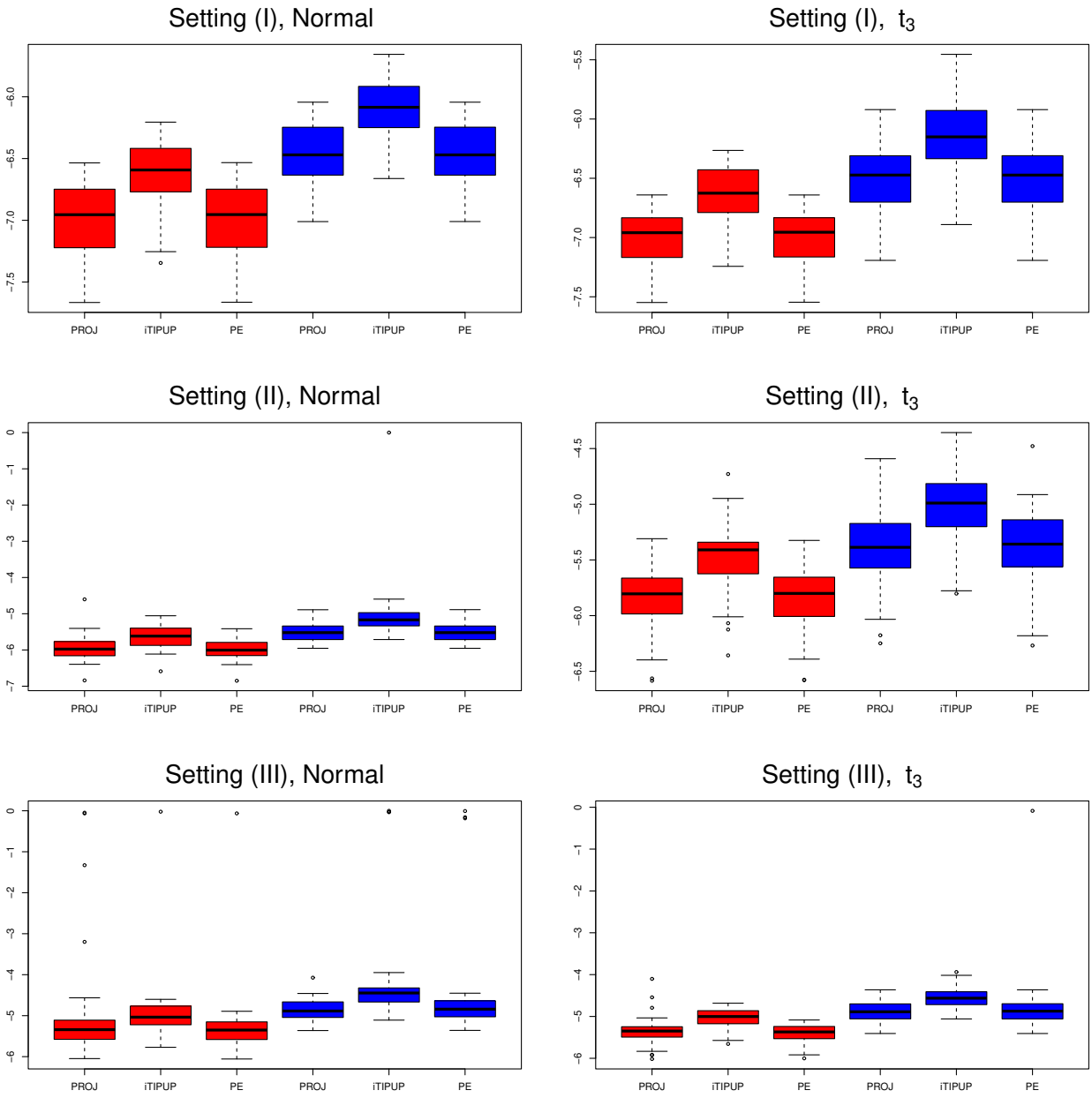


Fig. 3.9 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}}-\mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K=4, T=200, d_k=15$. In each panel, the left three boxplots (in red) represent sub-setting (a), while the right three boxplots (in blue) represent sub-setting (b).

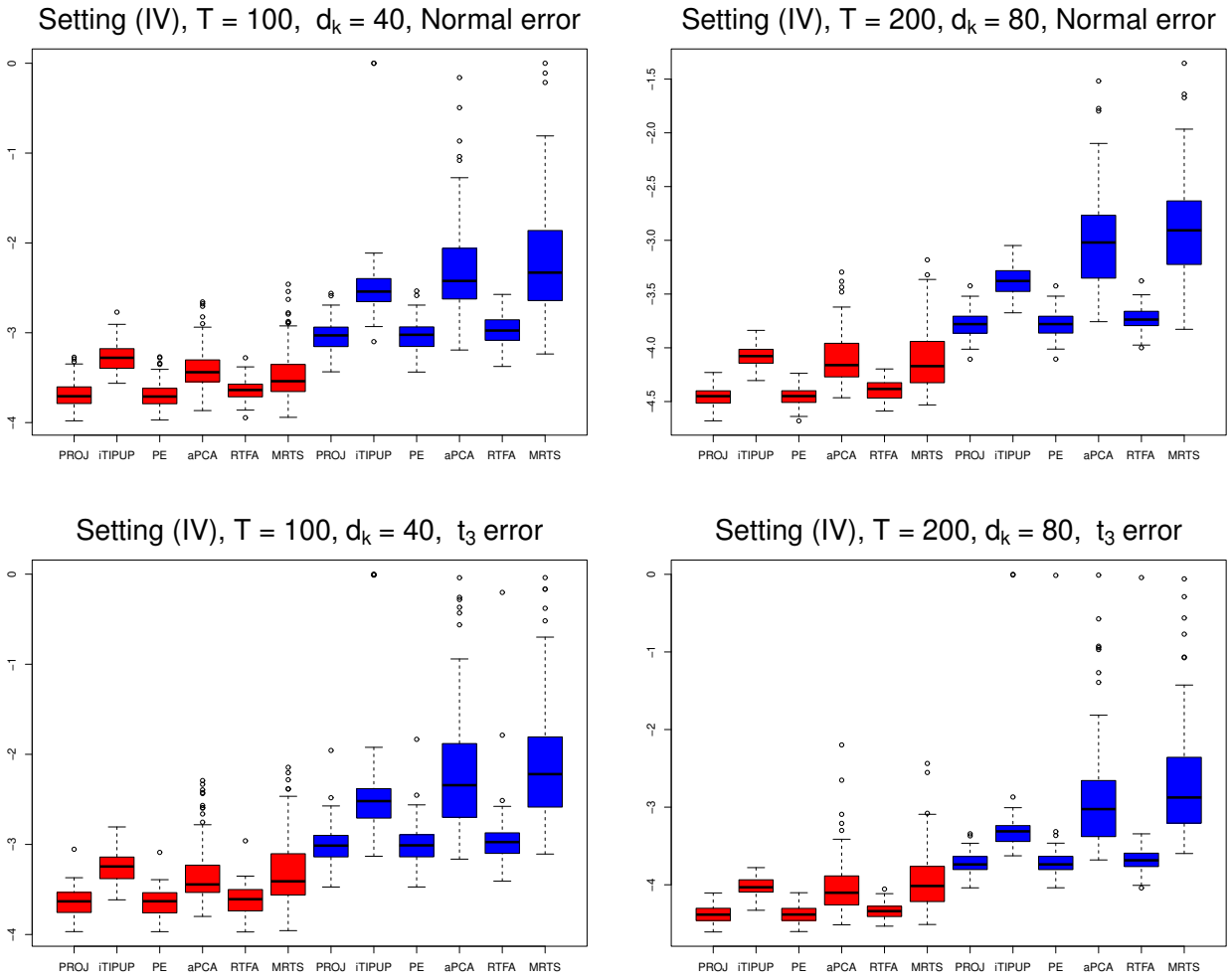


Fig. 3.10 Plot of estimation error $\left\|\widehat{\mathbf{Q}}_1\widehat{\mathbf{Q}}_1^{\mathrm{T}} - \mathbf{U}_1\mathbf{U}_1^{\mathrm{T}}\right\|$ (in log-scale) for $K = 2$, Setting (IV). In each panel, the left six boxplots (in red) represent sub-setting (a), while the right six boxplots (in blue) represent sub-setting (b).

3.5 Proof of Theorems

Before we present our proofs, we first introduce some notations which will be used in the proof of Theorem 3.1. Let $\mathbf{f}_{e,t,\ell}^{(k)} = (f_{e,t,\ell,j}^{(k)})$ to be the ℓ-th column of $\mathrm{mat}_k(\mathscr{F}_{e,t})$, and $\mathbf{f}_{t,\ell}^{(k)} = (f_{t,\ell,j}^{(k)})$ be the ℓ-th column vector in $\mathrm{mat}_k(\mathscr{F}_t)$. Then

$$f_{e,t,\ell,j}^{(k)} = \sum_{q\geq 0} a_{e,q} z_{e,t-q,\ell,j}^{(k)}, \quad j \in [r_{e,k}],\ \ell \in [r_{e,\text{-}k}]; \quad \boldsymbol{\varepsilon}_{t,\ell,j}^{(k)} = \sum_{q\geq 0} a_{\varepsilon,q} z_{\varepsilon,t-q,\ell,j}^{(k)}, \quad j \in [d_k],\ \ell \in [d_{\text{-}k}],$$
$$f_{t,\ell,j}^{(k)} = \sum_{q\geq 0} a_{f,q} z_{f,t-q,\ell,j}^{(k)}, \quad j \in [r_k]$$

where $z_{e,t,\ell,j}^{(k)}, z_{\boldsymbol{\varepsilon},t,\ell,j}^{(k)}$ and $z_{f,t,\ell,j}^{(k)}$ are the elements of $\mathscr{Z}_{e,t}$, $\mathscr{Z}_{\boldsymbol{\varepsilon},t}$, $\mathscr{Z}_{f,t}$ defined in Assumption (E2) and (F1), respectively. Furthermore, as introduced in Section 3.2, for a particular sample $S_{\text{-}k,m} \subseteq [d_{\text{-}k}]$ with $|S_{j,m}| \asymp d_j, j \neq k$, denote $\widetilde{\mathbf{x}}_{t,k,m} = \sum_{i\in S_{\text{-}k,m}} \mathbf{x}_{t,\text{-}k,i}$, $\widetilde{\mathbf{f}}_{t,k,m} = \sum_{i\in S_{\text{-}k,m}} \mathbf{f}_{t,\text{-}k,i}$, $\widetilde{\mathbf{e}}_{t,k,m} = \sum_{i\in S_{\text{-}k,m}} \mathbf{e}_{t,\text{-}k,i}$ and $\widetilde{\mu}_{k,m} = \sum_{i\in S_{\text{-}k,m}} \mu_{\text{-}k,i}$. Further, define $\ddot{\mathbf{f}}_{t,k,m} = \widetilde{\mathbf{f}}_{t,k,m}/s_{\text{-}k,m}^{\frac{1}{2}}$, write $\ddot{\mathbf{F}}_{k,m} = \left[\ddot{\mathbf{f}}_{1,k,m}, \ldots, \ddot{\mathbf{f}}_{T,k,m}\right] \in \mathbb{R}^{r_k \times T}$ and $\widetilde{\mathbf{F}}_{k,m} = \left[\widetilde{\mathbf{f}}_{1,k,m}, \ldots, \widetilde{\mathbf{f}}_{T,k,m}\right]$. Similarly, denote $\ddot{\mathbf{x}}_{t,k,m} = \widetilde{\mathbf{x}}_{t,k,m}/s_{\text{-}k,m}^{\frac{1}{2}}$ and $\ddot{\mathbf{X}}_{k,m} = (\ddot{\mathbf{x}}_{1,k,m}, ..., \ddot{\mathbf{x}}_{T,k,m})^{\mathrm{T}} \in \mathbb{R}^{T\times d_k}$, $\widetilde{\mathbf{X}}_{k,m} = (\widetilde{\mathbf{x}}_{1,k,m}, ..., \widetilde{\mathbf{x}}_{T,k,m})^{\mathrm{T}}$. Also let $\ddot{\mathbf{e}}_{t,k,m} = \widetilde{\mathbf{e}}_{t,k,m}/s_{\text{-}k,m}^{\frac{1}{2}}$, $\ddot{\mathbf{E}}_{k,m} = \left(\ddot{\mathbf{e}}_{1,k,m}, ..., \ddot{\mathbf{e}}_{T,k,m}\right)^{\mathrm{T}}$ and $\widetilde{\mathbf{E}}_{k,m} = \left(\widetilde{\mathbf{e}}_{1,k,m}, ..., \widetilde{\mathbf{e}}_{T,k,m}\right)^{\mathrm{T}}$. Finally, let $\ddot{\mu}_{k,m} = \widetilde{\mu}_{k,m}/s_{\text{-}k,m}^{\frac{1}{2}}$.

We present four important lemmas under our model assumptions. For convenience of presentation, we ignore the subscript m in Lemma 3.1, Lemma 3.2 and Lemma 3.4, since the results apply to all random samples $S_{\text{-}k,m} \subseteq [d_{\text{-}k}]$ with $|S_{j,m}| \asymp d_j, j \neq k$.

Lemma 3.1. *Under Assumption (E1), (E2), (L1), (L2), (R1), we have*

$$\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right) = O_p\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\left(1+\frac{d_k}{T}\right)\right), \tag{3.26}$$
$$\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right) = O_p\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\left(1+\frac{d_k}{T}\right)\right). \tag{3.27}$$

In addition, if Assumption (R2) is satisfied, then

$$P\left(\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right) \geq C\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\left(1+\frac{d_k}{T}\right)\right)\right) = 1, \tag{3.28}$$
$$P\left(\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right) \geq C\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\left(1+\frac{d_k}{T}\right)\right)\right) = 1, \tag{3.29}$$

for some $c \in (0,1]$ and $C > 0$.

In Lemma 3.1, (3.26) provides an upper bound for the largest eigenvalue of $\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k$ (and $\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k$), which facilitates the proof of Lemma 3.4. (3.28) suggests that at least $\lfloor c\min(T,d_k)\rfloor$ largest eigenvalues of $\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k$ (and $\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k$) are of the same order, which guarantees the validity of using eigenvalue ratio to detect the existence of factors, which will be further discussed in Remark 3.4.

Lemma 3.2. *Under Assumption (E1), (E2), (F1), (L1), (L2), (R1), we have*

$$\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\to\mathbf{\Sigma}_{F,k}$$

for some positive definite matrix $\mathbf{\Sigma}_{F,k}$, with all eigenvalues bounded away from 0 and infinity. For $j\in[r_k]$,

$$\lambda_j\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right)\asymp 1, \tag{3.30}$$

$$\lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right)\asymp d_k^{\alpha_{k,j}}, \tag{3.31}$$

$$\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right)\asymp d_k^{\alpha_{k,j}}, \tag{3.32}$$

$$\lambda_j\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right)\asymp 1, \tag{3.33}$$

$$\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right)\asymp d_k^{\alpha_{k,j}}. \tag{3.34}$$

and for $j\in[\min(z_k,r_k)]$,

$$\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)\asymp d_k^{\alpha_{k,j}}. \tag{3.35}$$

In addition, if Assumption (R2) is satisfied, then for $\min(z_k,r_k)+1\le j\le\lfloor c\min(T,d_k)\rfloor-r_k$,

$$\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)\asymp\frac{d_{-k}}{s_{-k}}\left(1+\frac{d_k}{T}\right). \tag{3.36}$$

The weak serial dependence of factors and errors are quantified in the following lemma.

Lemma 3.3. *Define $\mathbf{A}_{f,T}$ and $\mathbf{A}_{e,T}$ similar to $\mathbf{A}_{\varepsilon,T}$ in Assumption (R2). Define $\mathbf{A}_{f,T^2}$ to be the $T^2\times T^2$ fourth moment matrix of the MA process $f_{t,l,j}^{(k)}$ for any k,l,j, and define $\mathbf{A}_{e,T^2}$ and $\mathbf{A}_{\varepsilon,T^2}$ similarly. Then, under Assumption (E1), (E2) and (F1), we have the following*

results: For $\mathbf{A}_T$ *can either be* $\mathbf{A}_{f,T}$, $\mathbf{A}_{e,T}$ *or* $\mathbf{A}_{\varepsilon,T}$,

$$\mathbf{1}_T^{\mathrm{T}}\mathbf{A}_T\mathbf{1}_T = O(T),\ \ \|\mathbf{A}_T\|_F^2 = O(T). \tag{3.37}$$

For $\mathbf{A}_{T^2}$ *can either be* $\mathbf{A}_{f,T^2}$, $\mathbf{A}_{e,T^2}$ *or* $\mathbf{A}_{\varepsilon,T^2}$,

$$\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{T^2}\mathbf{1}_{T^2} = O(T^2),\ \ \|\mathbf{A}_{T^2}\|_F^2 = O(T^2). \tag{3.38}$$

Moreover,

$$vec(\mathbf{A}_{f,T})^{\mathrm{T}}vec(\mathbf{A}_{e,T}) = O(T),\ \ vec(\mathbf{A}_{f,T})^{\mathrm{T}}vec(\mathbf{A}_{\varepsilon,T}) = O(T), \tag{3.39}$$

$$\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T}\otimes\mathbf{A}_{e,T}\mathbf{1}_{T^2} = O(T^2),\ \ \mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T}\otimes\mathbf{A}_{\varepsilon,T}\mathbf{1}_{T^2} = O(T^2), \tag{3.40}$$

$$vec(\mathbf{A}_{f,T}\otimes\mathbf{A}_{f,T})^{\mathrm{T}}vec(\mathbf{A}_{e,T^2}) = O(T^2),\ \ vec(\mathbf{A}_{f,T}\otimes\mathbf{A}_{f,T})^{\mathrm{T}}vec(\mathbf{A}_{\varepsilon,T^2}) = O(T^2), \tag{3.41}$$

$$vec(\mathbf{A}_{f,T^2})^{\mathrm{T}}vec(\mathbf{A}_{e,T^2}) = O(T^2),\ \ vec(\mathbf{A}_{f,T^2})^{\mathrm{T}}vec(\mathbf{A}_{\varepsilon,T^2}) = O(T^2). \tag{3.42}$$

Given a particular sample, we can obtain $\widehat{\mathbf{Q}}_{k,(z_k)}$ as the z_k largest eigenvectors of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k} := \frac{\widetilde{\mathbf{X}}_k^{\mathrm{T}}(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}})\widetilde{\mathbf{X}}_k}{T}$, with normalisation $\widehat{\mathbf{Q}}_{k,(z_k)}^{\mathrm{T}}\widehat{\mathbf{Q}}_{k,(z_k)} = \mathbf{I}_{z_k}$. Let $\widetilde{\mathbf{V}}_k$ be the $z_k\times z_k$ diagonal matrix of the first z_k largest eigenvalues of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k}$ in decreasing order, and $\ddot{\mathbf{V}}_k := \widetilde{\mathbf{V}}_k/s_{\text{-}k}$. The theoretical properties of $\widehat{\mathbf{Q}}_{k,(z_k)}$ from a particular sample is presented in the following lemma.

Lemma 3.4. *Under Assumption (E1), (E2), (F1), (L1), (L2), (R1) and* $r_{e,k} = O(d_k)$*, for* $k\in[K]$*, let* $c_k := \frac{d_{\text{-}k}^2}{s_{\text{-}k}^2}\left(1+\frac{d_k^2}{T^2}\right)+\frac{d_{\text{-}k}}{s_{\text{-}k}}d_k^{\alpha_{k,1}}\min\left\{1+\frac{d_k}{T},\frac{r_k d_k}{T}\right\}$*, then*

$$\left\|\widehat{\mathbf{Q}}_{k,(z_k)}-\mathbf{Q}_k\ddot{\mathbf{H}}_k\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}}c_k\right), \tag{3.43}$$

where $\ddot{\mathbf{H}}_k = \frac{\mathbf{D}_k^{\frac{1}{2}}\ddot{\mathbf{F}}_k(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}})\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k,(z_k)}\ddot{\mathbf{V}}_k^{-1}}{T}$ *has* $rank(\ddot{\mathbf{H}}_k) = z_k$*. Moreover, further assuming (L1'), there exists* $\widehat{\mathbf{U}}_{k,(z_k)}$ *with* $\widehat{\mathbf{U}}_{k,(z_k)}^{\mathrm{T}}\widehat{\mathbf{U}}_{k,(z_k)} = \mathbf{I}_{z_k}$ *such that* $\widehat{\mathbf{Q}}_{k,(z_k)} = \widehat{\mathbf{U}}_{k,(z_k)}\mathbf{P}_{k,(z_k)}$ *with* $\mathbf{P}_{k,(z_k)}$ *an orthogonal matrix, and*

$$\left\|\widehat{\mathbf{U}}_{k,(z_k)}-\mathbf{U}_{k,(z_k)}\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}}\left[d_k^{2\alpha_{k,1}}\frac{r_k}{T}+c_k\right]\right), \tag{3.44}$$

where $\mathbf{U}_{k,(z_k)}$ *is the matrix consisting of the first* z_k *columns of* $\mathbf{U}_k$*.*

Remark 3.4. *Note that Lemma 3.2 implies the following result for eigenvalue ratio: For* $j \leq r_k - 1$,

$$\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} \asymp \frac{d_k^{\alpha_{k,j}}}{d_k^{\alpha_{k,j+1}}};$$

For $r_k + 1 \leq j \leq \lfloor c\min(T, d_k)\rfloor - r_k - 1$,

$$\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} \asymp 1;$$

For $j = r_k$,

$$\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} \asymp \frac{d_k^{\alpha_{k,r_k}}}{\frac{d_{\cdot k}}{s_{\cdot k}}\left(1+\frac{d_k}{T}\right)} \to \infty.$$

Therefore,

$$\frac{\lambda_1\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{r_k+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} \asymp \frac{d_k^{\alpha_{k,1}}}{\frac{d_{\cdot k}}{s_{\cdot k}}\left(1+\frac{d_k}{T}\right)} \to \infty. \tag{3.45}$$

(3.45) implies that at least one of ratio of subsequent eigenvalues $\left\{\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}, j \in [r_k]\right\}$ *goes to infinity. This is essential to detect the existence of factors. Note that in the special case when all factors have the same strength, we should have* $\frac{\lambda_{r_k}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{r_k+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} \to \infty$, *which recovers the result in Ahn and Horenstein (2013).*

In contrast, when there is no factor exist, we should have that

$$\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}=\frac{\lambda_j\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right)}\asymp 1$$

for $j=1,...,\lfloor c\min(T,d_k)\rfloor$ *by (3.28) in Lemma 3.1. Thus, every ratio (actually until* $\lfloor c\min(T,d_k)\rfloor$*) of subsequent eigenvalues remains bounded. In this way, we can use eigenvalue ratio to detect the existence of factors accordingly: when factors exist, at least one of the eigenvalue ratio will goes to infinity; when there is no factor, all eigenvalue ratios (until* $\lfloor c\min(T,d_k)\rfloor$*) remains bounded. Note that this is different from estimating the number of factors, since we can only know that there is factor exist, but we do not really know how many factors and what the factor strengths are. However, in the special case when all factors are of the same strength, we have* $\frac{\lambda_{r_k}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{r_k+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}\to\infty$ *and* $\frac{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}{\lambda_{j+1}\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)}\asymp 1$ *for* $j\neq r_k$*, which is how Ahn and Horenstein (2013) estimates the number of factors. In the same spirit, the eye-ball test of Lam and Yao (2012) also uses eigenvalue ratios in a similar manner to deduce the number of factors in a vector factor model.*

Proof of Lemma 3.1. By Assumption (E1), (E2), (R1), we have

$$\begin{aligned}\widetilde{\mathbf{e}}_{t,k}&=\mathbf{A}_{e,k}\widetilde{\mathbf{f}}_{e,t,k}+\sum_{\ell\in S_{\text{-}k}}(\boldsymbol{\Sigma}_{\varepsilon,\ell}^{(k)})^{1/2}\boldsymbol{\varepsilon}_{t,\ell}^{(k)}\\&:=\widetilde{\mathbf{e}}_{t,k,1}+\widetilde{\mathbf{e}}_{t,k,2}\end{aligned}$$

where $\widetilde{\mathbf{e}}_{t,k,1}:=\mathbf{A}_{e,k}\widetilde{\mathbf{f}}_{e,t,k}$ and $\widetilde{\mathbf{e}}_{t,k,2}=\sum_{\ell\in S_{\text{-}k}}(\boldsymbol{\Sigma}_{\varepsilon,\ell}^{(k)})^{1/2}\boldsymbol{\varepsilon}_{t,\ell}^{(k)}$, and $\widetilde{\mathbf{f}}_{e,t,k}=\sum_{\ell\in S_{\text{-}k}}\mathbf{f}_{t,l}^{(k)}$ is defined in a similar way as $\widetilde{\mathbf{f}}_{t,k}$. Using such decomposition, the error matrix can be written as

$$\widetilde{\mathbf{E}}_k=\widetilde{\mathbf{E}}_{k,1}+\widetilde{\mathbf{E}}_{k,2},$$

where $\widetilde{\mathbf{E}}_{k,1}=\left(\widetilde{\mathbf{e}}_{1,k,1},...,\widetilde{\mathbf{e}}_{T,k,1}\right)^{\mathrm{T}}$ and $\widetilde{\mathbf{E}}_{k,2}=\left(\widetilde{\mathbf{e}}_{1,k,2},...,\widetilde{\mathbf{e}}_{T,k,2}\right)^{\mathrm{T}}$. Then we can deal with $\widetilde{\mathbf{E}}_{k,1}$ and $\widetilde{\mathbf{E}}_{k,2}$ separately using random matrix theory. We first look at $\widetilde{\mathbf{E}}_{k,2}$. Similar to

assumptions in Ahn and Horenstein (2013), we further write the matrix $\widetilde{\mathbf{E}}_{k,2}$ as

$$\widetilde{\mathbf{E}}_{k,2} = L_{e,k,2}^{\frac{1}{2}} U_{e,k,2} R_{e,k,2}^{\frac{1}{2}},$$

where $U_{e,k,2} \in \mathbb{R}^{T \times d_k}$ are iid random variables with uniformly bounded fourth moment by Assumption (R1), and $L_{e,k,2} \in \mathbb{R}^{T \times T}$ and $R_{e,k,2} \in \mathbb{R}^{d_k \times d_k}$ are time-serial and cross-sectional covariance matrices. Define $\mathbf{\Sigma}_{\varepsilon}^{(k)} := \sum_{\ell \in S_{\text{-}k}} \mathbf{\Sigma}_{\varepsilon,\ell}^{(k)}$. By Assumption (E1), $R_{e,k,2} = \mathbf{\Sigma}_{\varepsilon}^{(k)}$, and $L_{e,k,2} = \mathbf{A}_{\varepsilon,T}$. Next, we need to bound the spectral norm of $L_{e,k,2}$ and $R_{e,k,2}$. Assumption (E1) implies $\left\|R_{e,k,2}\right\| = O(d_{\text{-}k})$. For $\mathbf{A}_{\varepsilon,T}$, since it is symmetric, $\|\mathbf{A}_{\varepsilon,T}\|_1 = \|\mathbf{A}_{\varepsilon,T}\|_\infty$, and

$$\begin{aligned} \|\mathbf{A}_{\varepsilon,T}\|_1 &= \max_t \sum_{s=1}^{T} |(\mathbf{A}_{\varepsilon,T})_{ts}| \\ &\leq 2 \sum_{v=0}^{T} \left| \sum_{q \geq 0} a_{\varepsilon,q} a_{\varepsilon,q+v} \right| \\ &\leq 2 \left(\sum_{q \geq 0} |a_{\varepsilon,q}| \right)^2 \leq C \end{aligned} \tag{3.46}$$

by Assumption (E2). Thus, $\left\|\mathbf{A}_{\varepsilon,T}\right\| \leq \sqrt{\|\mathbf{A}_{\varepsilon,T}\|_1 \|\mathbf{A}_{\varepsilon,T}\|_\infty} = \|\mathbf{A}_{\varepsilon,T}\|_1 \leq C$. Bai and Yin (1993) and Latala (2005) show that $\lambda_1\left(U_{e,k,2}U_{e,k,2}^{\mathrm{T}}/T\right) \to (1+\sqrt{d_k/T})^2$. Thus, similar to the result obtained by Ahn and Horenstein (2013) (see also Moon and Weidner (2015)), we have

$$\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right) \leq \left\|L_{e,k,2}\right\| \left\|R_{e,k,2}\right\| \left\|U_{e,k,2}U_{e,k,2}^{\mathrm{T}}/T\right\| = O_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right).$$

For $\widetilde{\mathbf{E}}_{k,1}$, note that the common error has a similar structure as the factor model. Similar to the definition of s_k and $s_{\text{-}k}$, define $s_{e,k} := \sum_{j \in S_{\text{-}k}} \left(\sum_{i=1}^{d_k} (\mathbf{A}_{e,k})_{ij}\right)^2$ and $s_{e,\text{-}k} := \prod_{l=1;l\neq k}^{K} s_{e,l}$. Define $\widetilde{\mathbf{F}}_{e,k}$ similar as $\widetilde{\mathbf{F}}_k$, then following the similar analysis as in Lemma 3.2, we have for $j \in [r_{e,k}]$,

$$\lambda_j\left(\frac{\widetilde{\mathbf{F}}_{e,k}\widetilde{\mathbf{F}}_{e,k}^{\mathrm{T}}}{T}\right) \asymp s_{e,\text{-}k}.$$

Thus,

$$\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right) = \lambda_1\left(\frac{\mathbf{A}_{e,k}\widetilde{\mathbf{F}}_{e,k}\widetilde{\mathbf{F}}_{e,k}^{\mathrm{T}}\mathbf{A}_{e,k}^{\mathrm{T}}}{T}\right) \leq \lambda_1\left(\mathbf{A}_{e,k}^{\mathrm{T}}\mathbf{A}_{e,k}\right)\lambda_1\left(\frac{\widetilde{\mathbf{F}}_{e,k}\widetilde{\mathbf{F}}_{e,k}^{\mathrm{T}}}{T}\right) = O_p(s_{e,\text{-}k}),$$

since $\lambda_1\left(\mathbf{A}_{e,k}^{\mathrm{T}}\mathbf{A}_{e,k}\right) = O(1)$ by Assumption (E1). Next, note that by definition,

$$s_{e,k} \leq r_{e,k}\|\mathbf{A}_{e,k}\|_1^2 = O(r_{e,k}),$$

so $s_{e,\text{-}k} = O(r_{e,\text{-}k})$. Since we assume $r_{e,k} = O(d_k)$, we have $\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right) = O_p(d_{\text{-}k})$.

Therefore,

$$\begin{aligned}
\lambda_1\left(\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}\widetilde{\mathbf{E}}_k}{T}\right) &\leq \left\|\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right\| + \left\|\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right\| + \left\|\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right\| + \left\|\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right\| \\
&= O_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right) + 2\left\|\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right\| \\
&\leq O_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right) + \sqrt{\left\|\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right\|\left\|\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right\|} \\
&= O_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right),
\end{aligned}$$

which implies (3.26), and

$$\begin{aligned}
\lambda_1\left(\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{E}}_k}{T}\right) &\leq \left\|\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}}{\sqrt{T}}\right\|^2\left\|\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right\| \\
&\leq 2\lambda_1\left(\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}\widetilde{\mathbf{E}}_k}{T}\right) \\
&= O_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right), \qquad (3.47)
\end{aligned}$$

which implies (3.27)

To show (3.28), we focus on $\widetilde{\mathbf{E}}_{k,2}$. Note that Assumption (R2) is parallel to Assumption (D) in Ahn and Horenstein (2013). Following Lemma A.7. in Ahn and Horenstein (2013), we have

$$P\left(\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{\max(T,d_k)}\right) \geq Cd_{\text{-}k}\right) = 1$$

for some $c \in (0,1]$ and $C > 0$. Thus,

$$P\left(\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\ddot{\mathbf{E}}_{k,2}^{\mathrm{T}}\ddot{\mathbf{E}}_{k,2}}{T}\right) \geq C\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right)\right) = 1$$

for some $c \in (0,1]$ and $C > 0$. Finally, by Weyl's inequalities,

$$\sqrt{\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}\widetilde{\mathbf{E}}_k}{T}\right)} \geq \sqrt{\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,2}}{T}\right)} - \sqrt{\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right)}$$
$$\asymp \sqrt{d_{\text{-}k}\left(1+\frac{d_k}{T}\right)},$$

which implies (3.28), since $\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\widetilde{\mathbf{E}}_{k,1}}{T}\right) = o_p\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right)$ under Assumption (R2). Similarly, following the same argument as Lemma A.7. in Ahn and Horenstein (2013), we can show that

$$P\left(\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\ddot{\mathbf{E}}_{k,2}^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,2}}{T}\right) \geq C\left(d_{\text{-}k}\left(1+\frac{d_k}{T}\right)\right)\right) = 1$$

for some $c \in (0,1]$ and $C > 0$, so

$$\sqrt{\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\widetilde{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{E}}_k}{T}\right)} \geq \sqrt{\lambda_{\lfloor c\min(T,d_k)\rfloor}\left(\frac{\widetilde{\mathbf{E}}_{k,2}^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{E}}_{k,2}}{T}\right)}$$
$$- \sqrt{\lambda_1\left(\frac{\widetilde{\mathbf{E}}_{k,1}^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{E}}_{k,1}}{T}\right)}$$
$$\asymp \sqrt{d_{\text{-}k}\left(1+\frac{d_k}{T}\right)},$$

which implies (3.29). This completes the proof of Lemma 3.1. □

Proof of Lemma 3.2. We consider the case $K = 3$ without loss of generality. By Assumption (F1), we know each element of $\mathscr{F}_t$ are independent with mean 0 and variance 1. Suppose we want to estimate $\mathbf{A}_1$. For $i, j \in [r_1]$, we have

$$E\left(\widetilde{f}_{t,1,i}\widetilde{f}_{t,1,j}\right) = \sum_{p=1}^{r_2}\sum_{q=1}^{r_3}\sum_{u=1}^{r_2}\sum_{v=1}^{r_3}\left(\sum_{\ell\in S_2}(\mathbf{A}_2)_{lp}\sum_{\ell\in S_3}(\mathbf{A}_3)_{lq}\sum_{\ell\in S_2}(\mathbf{A}_2)_{lu}\sum_{\ell\in S_3}(\mathbf{A}_3)_{lv}\right)E(f_{t,ipq}f_{t,juv})$$
$$= \sum_{p=1}^{r_2}\sum_{q=1}^{r_3}\left(\sum_{\ell\in S_2}(\mathbf{A}_2)_{lp}\right)^2\left(\sum_{\ell\in S_3}(\mathbf{A}_3)_{lq}\right)^2$$
$$\asymp s_2 s_3 = s_{\text{-}1}$$

when $i=j$, and 0 otherwise. Thus, $E\left(\ddot{f}_{t,1,i}\ddot{f}_{t,1,j}\right) \asymp 1$ for $i=j$, and 0 otherwise, which implies that $\mathbf{\Sigma}_{F,k}$ is a diagonal matrix with all diagonal elements bounded away from 0 and infinity.

Next, we show (3.30) – (3.32). To apply random matrix theory, similar to Ahn and Horenstein (2013), we can further write the matrix $\ddot{\mathbf{F}}_k$ as

$$\ddot{\mathbf{F}}_k = L_{f,k}^{\frac{1}{2}} U_{f,k} R_{f,k}^{\frac{1}{2}},$$

where $U_{f,k} \in \mathbb{R}^{r_k \times T}$ are iid random variables with uniformly bounded fourth moment by Assumption (R1), and $L_{f,k} \in \mathbb{R}^{r_k \times r_k}$ and $R_{f,k} \in \mathbb{R}^{T \times T}$ are cross-sectional and time-serial covariance matrices. By Assumption (F1), $L_{f,k} = I_{r_k}$, and $R_{f,k} = \mathbf{A}_{f,T}$. Following similar analysis as (3.46), we have $\|R_{f,k}\|_1 \leq C$. Thus, $\left\|R_{f,k}\right\| \leq \sqrt{\left\|R_{f,k}\right\|_1 \left\|R_{f,k}\right\|_\infty} = \left\|R_{f,k}\right\|_1 \leq C$. Since we assume $r_k = o(T)$, Bai and Yin (1993) and Latala (2005) show that $\lambda_1\left(U_{f,k}U_{f,k}^{\mathrm{T}}/T\right) - (1+\sqrt{r_k/T})^2 \to 0$ and $\lambda_{r_k}\left(U_{f,k}U_{f,k}^{\mathrm{T}}/T\right) - (1-\sqrt{r_k/T})^2 \to 0$. Similar to the result obtained by Ahn and Horenstein (2013) (see also Moon and Weidner (2015)), since the largest eigenvalues of $R_{f,k}$ and $L_{f,k}$ are bounded, we have

$$\lambda_1\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}/T\right) \leq \left\|L_{f,k}\right\|\left\|R_{f,k}\right\|\left\|U_{f,k}U_{f,k}^{\mathrm{T}}/T\right\| = O_p(1).$$

Similarly, since $\lambda_1(R_{f,k}) = \left\|R_{f,k}\right\| \geq 1$, we have

$$\lambda_{r_k}\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}/T\right) = \lambda_{r_k}\left(U_{f,k}R_{f,k}U_{f,k}^{T}/T\right) = \lambda_{r_k}\left(U_{f,k}^{T}U_{f,k}R_{f,k}/T\right) \geq \lambda_{r_k}\left(U_{f,k}^{T}U_{f,k}/T\right)\lambda_1(R_{f,k}) \geq 1$$

as $r_k, T \to \infty$. Thus, we have $\lambda_j\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) \asymp 1$ for $j \in [r_k]$.

Next, for $j \in [r_k]$, we know $\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right) = \lambda_j\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k}{T}\right)$ and

$$\lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right)\lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) \leq \lambda_j\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k}{T}\right) \leq \lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right)\lambda_1\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right).$$

Since $\lambda_j\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) \asymp 1$ for $j \in [r_k]$, it follows that $\lambda_j\left(\frac{\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\ddot{\mathbf{F}}_k}{T}\right) \asymp \lambda_j(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k)$ for $j \in [r_k]$. Similarly, $\lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right) = \lambda_j\left(\mathbf{D}_k^{\frac{1}{2}}\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\mathbf{D}_k^{\frac{1}{2}}\right) = \lambda_j\left(\mathbf{D}_k\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\right)$ and

$$\lambda_j(\mathbf{D}_k)\lambda_{r_k}\left(\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\right) \leq \lambda_j\left(\mathbf{D}_k\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\right) \leq \lambda_j(\mathbf{D}_k)\lambda_1\left(\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\right).$$

By Assumption (L1), $\lambda_j(\mathbf{D}_k) \asymp d_k^{\alpha_{k,j}}$ and $\lambda_j\left(\mathbf{D}_k^{-\frac{1}{2}}\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\mathbf{D}_k^{-\frac{1}{2}}\right) \asymp 1$ for $j \in [r_k]$ (this also holds when elements of $\mathbf{A}_k$ are random and independent). Thus, for $j \in [r_k]$, we have $\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right) \asymp \lambda_j(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k) \asymp d_k^{\alpha_{k,j}}$.

Next, we will show (3.33) – (3.34) accordingly. To start with, note that

$$\begin{aligned}\lambda_1\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) &\le \left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right\| + \left\|\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right\| \\ &\le \left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right\| + \left\|\frac{\ddot{\mathbf{F}}_k}{\sqrt{T}}\right\|^2\left\|\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right\| = 2\left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right\| \asymp 1.\end{aligned}$$

And

$$\begin{aligned}\lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) &= \lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}-\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) \\ &\ge \lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) - \lambda_1\left(\frac{\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T^2}\right). \qquad (3.48)\end{aligned}$$

We next obtain an upper bound for $\lambda_1\left(\frac{\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T^2}\right)$. Note that $\ddot{\mathbf{F}}_k\mathbf{1}_T \in \mathbb{R}^{r_k\times 1}$ is a random vector, with its i-th element to be $\sum_{t=1}^T \ddot{f}_{t,i}$. For its moment bounds, $E\left(\sum_{t=1}^T \ddot{f}_{t,i}\right)^2 = \mathbf{1}_T^{\mathrm{T}}\mathbf{A}_{f,T}\mathbf{1}_T = O(T)$ and $E\left(\sum_{t=1}^T \ddot{f}_{t,i}\right)^4 = \mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T^2}\mathbf{1}_{T^2} = O(T^2)$ by Lemma 3.3. Therefore, following similar argument in previous analysis, we can decompose $\ddot{\mathbf{F}}_k\mathbf{1}_T = \sqrt{T}\mathbf{I}_{r_k}U$, where $U \in \mathbb{R}^{r_k\times 1}$ has iid random entries with bounded second and fourth moments. Thus,

$$\lambda_1\left(\frac{\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T^2}\right) = \frac{1}{T^2}\lambda_1\left(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{F}}_k^{\mathrm{T}}\right) \le \frac{1}{T^2}O_p(r_kT) = O_p\left(\frac{r_k}{T}\right) = o_p(1),$$

which implies $\lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) \succeq 1$ by (3.48). Therefore, $\lambda_j\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right)$ for $j \in [r_k]$.

Next, for $j \in [r_k]$, it is easy to see

$$\begin{aligned}\lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right)\lambda_{r_k}\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right) &\le \lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right) \\ &\le \lambda_j\left(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k\right)\lambda_1\left(\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\right),\end{aligned}$$

which implies $\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right) \asymp d_k^{\alpha_{k,j}}$ for $j \in [r_k]$.

Observe that $\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)^2 = \mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}$, and $\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right) = \left(\mathbf{A}_k\ddot{\mathbf{F}}_k + \ddot{\mathbf{E}}_k^{\mathrm{T}}\right)\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)$. Hence to show (3.35), following the similar argument as Theorem 1 in Freyaldenhoven (2022), for $j \in [r_k]$,

$$\begin{aligned}\sqrt{\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right)} &\geq \sqrt{\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right)} - \sqrt{\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right)} \\ &\succeq \sqrt{d_k^{\alpha_{k,j}}} - \sqrt{O_p\left(\frac{d_{-k}}{s_{-k}}\left(1+\frac{d_k}{T}\right)\right)} \asymp \sqrt{d_k^{\alpha_{k,j}}}. \qquad (3.49)\end{aligned}$$

where the second line follows from Lemma 3.1 and the last line follows from Assumption (L2). Similarly, for $j \in [r_k]$,

$$\begin{aligned}\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right) &\leq \lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right) + \lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right) \\ &\quad + 2\sqrt{\lambda_j\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right)}\sqrt{\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right)} \\ &\asymp d_k^{\alpha_{k,j}} + O_p\left(\frac{d_{-k}}{s_{-k}}\left(1+\frac{d_k}{T}\right)\right) \asymp d_k^{\alpha_{k,j}}. \qquad (3.50)\end{aligned}$$

Therefore, $\lambda_j\left(\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T}\right) \asymp d_k^{\alpha_{k,j}}$ for $j \in [r_k]$.

Finally, to show (3.36), we apply similar technique as Lemma A.8 in Ahn and Horenstein (2013). For simplicity of expression, in the following proof, we write $\mathbf{M}_T = \mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}$ and $\mathbf{M} = (\ddot{\mathbf{F}}_k\mathbf{M}_T)^{\mathrm{T}}(\ddot{\mathbf{F}}_k\mathbf{M}_T\ddot{\mathbf{F}}_k^{\mathrm{T}})^{-1}\ddot{\mathbf{F}}_k\mathbf{M}_T$. Then $rank(\mathbf{M}) \leq r_k$, and

$$\ddot{\mathbf{X}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{X}}_k = \left(\mathbf{A}_k\ddot{\mathbf{F}}_k + \ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)\mathbf{M}_T\left(\mathbf{A}_k\ddot{\mathbf{F}}_k + \ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)^{\mathrm{T}} + \ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T - \mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k.$$

Hence, for $r_k + 1 \leq j \leq \lfloor c\min(T, d_k)\rfloor - r_k$,

$$\begin{aligned}\lambda_j\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T - \mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right) &\leq \lambda_j\left(\ddot{\mathbf{X}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{X}}_k\right) \\ &\leq \lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T - \mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right) \\ &\quad + \lambda_{r_k+1}\left[\left(\mathbf{A}_k\ddot{\mathbf{F}}_k + \ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)\mathbf{M}_T\left(\mathbf{A}_k\ddot{\mathbf{F}}_k + \ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)^{\mathrm{T}}\right] \\ &\leq \lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T - \mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right),\end{aligned}$$

since $\left(\mathbf{A}_k\ddot{\mathbf{F}}_k+\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)\mathbf{M}_T\left(\mathbf{A}_k\ddot{\mathbf{F}}_k+\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\right)^{\mathrm{T}}$ is positive semi-definite and has rank at most r_k. Similarly, we can show that

$$\begin{aligned}\lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T-\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right)&\leq\lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T-\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k+\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\ddot{\mathbf{E}}_k\right)\\&=\lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{E}}_k\right),\end{aligned}$$

$$\begin{aligned}\lambda_{j+r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{E}}_k\right)&\leq\lambda_j\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T-\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right)+\lambda_{r_k+1}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\ddot{\mathbf{E}}_k\right)\\&=\lambda_j\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{M}_T-\mathbf{M}\mathbf{M}_T\mathbf{M}^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k\right).\end{aligned}$$

Therefore, for $r_k+1\leq j\leq\lfloor c\min(T,d_k)\rfloor-r_k$,

$$\lambda_{j+r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{E}}_k\right)\leq\lambda_j\left(\ddot{\mathbf{X}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{X}}_k\right)\leq\lambda_{j-r_k}\left(\ddot{\mathbf{E}}_k^{\mathrm{T}}\mathbf{M}_T\ddot{\mathbf{E}}_k\right),$$

which implies (3.36) with the result of Lemma 3.1. This completes the proof of Lemma 3.2. □

Proof of Lemma 3.3. For $\mathbf{A}_T$ can either be $\mathbf{A}_{f,T}$, $\mathbf{A}_{e,T}$ or $\mathbf{A}_{\varepsilon,T}$, we have already shown that $\|\mathbf{A}_T\|_1=O(1)$ in the Proof of Lemma 3.1 and Lemma 3.2 (see (3.46) for example). Hence

$$\mathbf{1}_T^{\mathrm{T}}\mathbf{A}_T\mathbf{1}_T\leq\sum_{t,s=1}^{T}|(\mathbf{A}_T)_{ts}|\leq T\|\mathbf{A}_T\|_1=O(T).$$

In addition, it is not difficult to see every entry of $\mathbf{A}_T$ has absolute value bounded above by 1, hence,

$$\|\mathbf{A}_T\|_F^2=\sum_{t,s=1}^{T}|(\mathbf{A}_T)_{ts}|^2\leq\sum_{t,s=1}^{T}|(\mathbf{A}_T)_{ts}|\leq T\|\mathbf{A}_T\|_1=O(T).$$

Next, for $\mathbf{A}_{T^2}$ can either be $\mathbf{A}_{f,T^2}$, $\mathbf{A}_{e,T^2}$ or $\mathbf{A}_{\varepsilon,T^2}$, we have

$$\|\mathbf{A}_{T^2}\|_1=\max_{t_1,s_1}\sum_{t_2,s_2=1}^{T}|(\mathbf{A}_{T^2})_{t_1s_1,t_2s_2}|\leq 2\max_{q_1q_2}\sum_{q_3,q_4}\left|a_{q_1}a_{q_2}a_{q_3}a_{q_4}\right|\leq 2\left(\sum_{q\geq 0}|a_q|\right)^4\leq C.$$

Hence,

$$\mathbf{1}_T^{\mathrm{T}}\mathbf{A}_{T^2}\mathbf{1}_T\leq\sum_{t,s=1}^{T}|(\mathbf{A}_{T^2})_{ts}|\leq T^2\|\mathbf{A}_{T^2}\|_1=O(T^2).$$

Similarly, we can also observe that every entry of $\mathbf{A}_{T^2}$ has absolute value bounded above by 1. To see this, take $f^{(k)}_{t,l,j}$ for example, note that for the MA process $f_t := f^{(k)}_{t,l,j}$ for any (l,j,k), $|E(f_{t_1}f_{t_2}f_{t_3}f_{t_4})| \le 0.5\left[E(f^2_{t_1}f^2_{t_2})+E(f^2_{t_3}f^2_{t_4})\right] \le 0.25\left[E(f^4_{t_1})+E(f^4_{t_2})+E(f^4_{t_3})+E(f^4_{t_4})\right] \le 1$, because $E(f^4_{t_i}) = \sum_{q\ge 0} a_q^4 \le \left(\sum_{q\ge 0} a_q^2\right)^2 = 1$ for any t_i. Therefore,

$$\|\mathbf{A}_{T^2}\|_F^2 = \sum_{t,s=1}^{T} |(\mathbf{A}_{T^2})_{ts}|^2 \le \sum_{t,s=1}^{T} |(\mathbf{A}_{T^2})_{ts}| \le T^2\|\mathbf{A}_{T^2}\|_1 = O(T^2).$$

(3.39) – (3.42) can be easily implied by (3.37) and (3.38). To see this, first note that for any matrices A and B, $\|A\otimes B\|_F = \|A\|_F\|B\|_F$, and $|\mathrm{vec}(A)^{\mathrm{T}}\mathrm{vec}(B)| \le \|A\|_F\|B\|_F$ by Cauchy-Schwarz, which implies (3.39), (3.41) and (3.42). Finally, $\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T}\otimes\mathbf{A}_{e,T}\mathbf{1}_{T^2} \le T^2\|\mathbf{A}_{f,T}\otimes\mathbf{A}_{e,T}\|_1 = T^2\|\mathbf{A}_{f,T}\|_1\|\mathbf{A}_{e,T}\|_1 = O(T^2)$, which gives (3.40). This completes the proof of Lemma 3.3. □

Proof of Lemma 3.4. Note that we have

$$\begin{aligned}\frac{\ddot{\mathbf{X}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_k}{T} &= \frac{\left(\mathbf{A}_k\ddot{\mathbf{F}}_k+\ddot{\mathbf{E}}_k^{\mathrm{T}}\right)\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\left(\mathbf{A}_k\ddot{\mathbf{F}}_k+\ddot{\mathbf{E}}_k^{\mathrm{T}}\right)^{\mathrm{T}}}{T}\\ &= \mathbf{A}_k\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T}\mathbf{A}_k^{\mathrm{T}}+R_k, \end{aligned} \tag{3.51}$$

where

$$R_k = \frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}+\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}+\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}. \tag{3.52}$$

We estimate $\widehat{\mathbf{Q}}_{k,(z_k)}$ as the first z_k eigenvectors of (3.51). Let's define $\mathbf{U}_{k,(z_k)}$ to be the matrix consisting of the first z_k columns of $\mathbf{U}_k$, and $\mathbf{B}$ be its orthogonal complement. Then $\mathbf{U}_{k,(z_k)}$ is an invariant subspace for $\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}$ and

$$\begin{bmatrix}\mathbf{U}_{k,(z_k)}^{\mathrm{T}}\\ \mathbf{B}^{\mathrm{T}}\end{bmatrix}\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\begin{bmatrix}\mathbf{U}_{k,(z_k)} & \mathbf{B}\end{bmatrix} = \begin{bmatrix}\mathbf{G}_{k,(z_k)} & 0\\ 0 & \Lambda_{k,(z_k)}\end{bmatrix},$$

where $\mathbf{G}_{k,(z_k)}$ is a $z_k\times z_k$ diagonal matrix consisting the largest z_k eigenvalues of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k$, and $\Lambda_{k,(z_k)}$ is a $(d_k-z_k)\times(d_k-z_k)$ diagonal matrix where the first r_k-z_k entries are the z_k+1 to r_k eigenvalues of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k$, and the remaining entries are all 0's. In this way, we can apply Lemma 3 of Lam et al. (2011) and know that there exists $\widehat{\mathbf{U}}_{k,(z_k)}$ (with $\widehat{\mathbf{U}}_{k,(z_k)}^{\mathrm{T}}\widehat{\mathbf{U}}_{k,(z_k)} = \mathbf{I}_{z_k}$)

such that $\widehat{\mathbf{Q}}_{k,(z_k)} = \widehat{\mathbf{U}}_{k,(z_k)}\mathbf{P}_{k,(z_k)}$ where $\mathbf{P}_{k,(z_k)}$ is orthogonal matrix, so that

$$\begin{aligned}\left\|\widehat{\mathbf{U}}_{k,(z_k)} - \mathbf{U}_{k,(z_k)}\right\| &\leq \frac{8\left\|\mathbf{A}_k\left[\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right]\mathbf{A}_k^{\mathrm{T}} + R_k\right\|}{\operatorname{sep}(\mathbf{G}_{k,z_k}, \Lambda_{k,z_k})} \\ &\preceq \frac{\left\|\mathbf{A}_k\left[\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right]\mathbf{A}_k^{\mathrm{T}}\right\| + \left\|R_k\right\|}{d_k^{\alpha_{k,z_k}}}. \end{aligned} \tag{3.53}$$

Next, we bound the norms on the numerator of (3.53). For the first term,

$$\begin{aligned}\left\|\mathbf{A}_k\left[\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right]\mathbf{A}_k^{\mathrm{T}}\right\| &\leq \left\|\mathbf{A}_k\right\|^2\left\|\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right\| \\ &\asymp d_k^{\alpha_{k,1}}\left\|\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right\|.\end{aligned}$$

By Assumption (F1), $\{f_{t,l,j}^{(k)}\}$ is a linear process with absolutely summable autocovariance sequence. In Lemma 3.2, we have shown that elements of $\ddot{\mathbf{F}}_k$ retain the covariance structure of $\{f_{t,l,j}^{(k)}\}$, so each row of $\ddot{\mathbf{F}}_k$ is a linear process with absolutely summable autocovariance sequence, which satisfies Assumption (R2) in Lam (2021). Thus, applying the result from Lemma 3 (or more specifically, equation (8.26)) of Lam (2021) implies

$$\left\|\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right\| = O_p\left(\sqrt{\frac{r_k}{T}}\right), \tag{3.54}$$

which further gives

$$\left\|\mathbf{A}_k\left[\frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k}\right]\mathbf{A}_k^{\mathrm{T}}\right\| = O_p\left(d_k^{\alpha_{k,1}}\sqrt{\frac{r_k}{T}}\right).$$

Next, we bound the norm of R_k. First, note that $\left\|\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right\| \leq \left\|\mathbf{I}_T\right\| + \left\|\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right\| = 2$. Bounding the squared norm of each term on the right hand side of (3.52), we have

$$\begin{aligned}\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2 &\leq \left\|\mathbf{I}_T - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right\|^2\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k}{T^{\frac{1}{2}}}\right\|^2\left\|\frac{\ddot{\mathbf{E}}_k}{T^{\frac{1}{2}}}\right\|^2 \\ &\leq 4\lambda_1\left(\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right)\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right) \\ &= O_p\left(d_k^{\alpha_{k,1}}\right)O_p\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\left(1 + \frac{d_k}{T}\right)\right),\end{aligned} \tag{3.55}$$

where the last line follows from Lemma 3.1 and Lemma 3.2. Similarly,

$$\left\|\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\right\|^2=O_p\left(d_k^{\alpha_{k,1}}\right)O_p\left(\frac{d_{-k}}{s_{-k}}\left(1+\frac{d_k}{T}\right)\right).$$

As for the last term,

$$\left\|\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2\leq 4\lambda_1\left(\frac{\ddot{\mathbf{E}}_k^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right)^2=O_p\left(\frac{d_{-k}^2}{s_{-k}^2}\left(1+\frac{d_k^2}{T^2}\right)\right) \tag{3.56}$$

by Lemma 3.1 and Lemma 3.2.

For the first two terms on the right hand side of (3.52), there may exist potentially better bounds. To see this, note that we can equivalently write

$$\begin{aligned}\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2&\leq\|\mathbf{A}_k\|^2\left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k}{T}\right\|^2+\|\mathbf{A}_k\|^2\left\|\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right\|^2\\&=O_p\left(d_k^{\alpha_{k,1}}\right)\left(\left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k}{T}\right\|^2+\left\|\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right\|^2\right).\end{aligned}$$

Since $\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k}{T}$ and $\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}$ are $r_k\times d_k$ matrices with each element having a certain rate of convergence, we actually can have a better rate by simply counting the numbers of elements of them. More specifically, the (i,j) element of $\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k$ is

$$\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{ij}=\sum_{t=1}^{T}\ddot{f}_{t,k,i}\ddot{e}_{t,k,j},$$

where $\ddot{f}_{t,k,i}$ is the i-th entry of $\ddot{\mathbf{f}}_{t,k}$, and $\ddot{e}_{t,k,j}$ are defined similarly. We now focus on bounding the Frobenius norm of $\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k$. Note that $\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2=\sum_{(i,j)\in(r_k,d_k)}\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{ij}^2$ and

$$\begin{aligned}\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2&=E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2+O_p\left(\sqrt{Var(\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2)}\right)\\&=E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2+O_p\left(\sqrt{E\left(\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^4\right)-\left(E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2\right)^2}\right).\end{aligned} \tag{3.57}$$

Thus, we need to obtain bounds for $E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2$ and $E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^4$. We start with $E\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\|_F^2$. For each entry of $\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k$, we have

$$E\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{ij}^2=E\left(\sum_{t=1}^{T}\ddot{f}_{t,k,i}\ddot{e}_{t,k,j}\right)^2=\sum_{t=1}^{T}\sum_{s=1}^{T}E\left(\ddot{f}_{t,k,i}\ddot{f}_{s,k,i}\right)E\left(\ddot{e}_{t,k,i}\ddot{e}_{s,k,i}\right).$$

By Assumption (F1),

$$E\left(\ddot{f}_{t,k,i}\ddot{f}_{s,k,i}\right) = \sum_{q\geq 0} a_{f,q}a_{f,q-|t-s|} = \left(\mathbf{A}_{f,T}\right)_{ts}. \tag{3.58}$$

Similarly, by Assumption (E1) and (E2),

$$E\left(\ddot{e}_{t,k,j}\ddot{e}_{s,k,j}\right) = \frac{1}{s_{-k}}\left[s_{e,-k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}\left(\mathbf{A}_{e,T}\right)_{ts} + \left(\mathbf{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\left(\mathbf{A}_{\varepsilon,T}\right)_{ts}\right]. \tag{3.59}$$

Hence,

$$\begin{aligned} E\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{ij}^2 &= \frac{1}{s_{-k}}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\mathbf{A}_{f,T}\right)_{ts}\left[s_{e,-k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}\left(\mathbf{A}_{e,T}\right)_{ts} + \left(\mathbf{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\left(\mathbf{A}_{\varepsilon,T}\right)_{ts}\right] \\ &= \frac{1}{s_{-k}}\left[s_{e,-k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}\mathrm{vec}(\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{e,T}) + \left(\mathbf{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\mathrm{vec}(\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{\varepsilon,T})\right], \end{aligned}$$

and

$$\begin{aligned} E\left\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right\|_F^2 &= \sum_{i=1}^{r_k}\sum_{j=1}^{d_k} E\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{ij}^2 \\ &= \frac{r_k}{s_{-k}}\left[s_{e,-k}\mathrm{tr}(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}})\mathrm{vec}(\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{e,T}) + \mathrm{tr}(\mathbf{\Sigma}_{\varepsilon}^{(k)})\mathrm{vec}(\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{\varepsilon,T})\right] \\ &= O_p\left(r_k d_k T\frac{d_{-k}}{s_{-k}}\right), \end{aligned} \tag{3.60}$$

where the last line follows from Assumption (E1) and Lemma 3.3. Next, for $E\left\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right\|_F^4$, note that

$$E\left\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right\|_F^4 = \sum_{i_1=1}^{r_k}\sum_{j_1=1}^{d_k}\sum_{i_2=1}^{r_k}\sum_{j_2=1}^{d_k} E\left[\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{i_1j_1}^2\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{i_2j_2}^2\right],$$

and

$$\begin{aligned} E\left[\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{i_1j_1}^2\left(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right)_{i_2j_2}^2\right] &= E\left[\left(\sum_{t=1}^{T}\ddot{f}_{t,k,i_1}\ddot{e}_{t,k,j_1}\right)^2\left(\sum_{t=1}^{T}\ddot{f}_{t,k,i_2}\ddot{e}_{t,k,j_2}\right)^2\right] \\ &= \sum_{t_1=1}^{T}\sum_{s_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_2=1}^{T} E\left(\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\ddot{f}_{t_2,k,i_2}\ddot{f}_{s_2,k,i_2}\right) \\ &\quad \cdot E\left(\ddot{e}_{t_1,k,j_1}\ddot{e}_{s_1,k,j_1}\ddot{e}_{t_2,k,j_2}\ddot{e}_{s_2,k,j_2}\right). \end{aligned}$$

We deal with the terms separately. For the term related to $\ddot{f}$, we separate the cases for $i_1 = i_2$ and $i_1 \neq i_2$. When $i_1 = i_2$, by Assumption (F1), we have

$$E\left(\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\ddot{f}_{t_2,k,i_2}\ddot{f}_{s_2,k,i_2}\right) = E\left(\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\ddot{f}_{t_2,k,i_1}\ddot{f}_{s_2,k,i_1}\right) = \left(\mathbf{A}_{f,T^2}\right)_{t_1s_1,t_2s_2},$$

where $\left(\mathbf{A}_{f,T^2}\right)_{t_1s_1,t_2s_2}$ is the (t_1s_1,t_2s_2) entry of $\mathbf{A}_{f,T^2}$. When $i_1 \neq i_2$,

$$E\left(\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\ddot{f}_{t_2,k,i_2}\ddot{f}_{s_2,k,i_2}\right) = E\left(\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\right)E\left(\ddot{f}_{t_2,k,i_2}\ddot{f}_{s_2,k,i_2}\right) = \left(\mathbf{A}_{f,T}\right)_{t_1s_1}\left(\mathbf{A}_{f,T}\right)_{t_2s_2}.$$

Next, for the term related to $\ddot{e}$, by Assumption (E1) and (E2),

$$\begin{aligned}
&E\left(\ddot{e}_{t_1,k,j_1}\ddot{e}_{s_1,k,j_1}\ddot{e}_{t_2,k,j_2}\ddot{e}_{s_2,k,j_2}\right)\\
\leq&\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\left(\mathbf{A}_{e,T^2}\right)_{t_1s_1,t_2s_2} + \sum_{l=1}^{d_{\text{-}k}}(\mathbf{\Sigma}_{\varepsilon,l}^{(k)})_{j_1j_1}(\mathbf{\Sigma}_{\varepsilon,l}^{(k)})_{j_2j_2}\left(\mathbf{A}_{\varepsilon,T^2}\right)_{t_1s_1,t_2s_2}\right]\\
\leq&\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\left(\mathbf{A}_{e,T^2}\right)_{t_1s_1,t_2s_2} + (\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\left(\mathbf{A}_{\varepsilon,T^2}\right)_{t_1s_1,t_2s_2}\right].
\end{aligned} \tag{3.61}$$

Hence, when $i_1 = i_2$,

$$\begin{aligned}
&E\left[(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)_{i_1j_1}^2(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)_{i_2j_2}^2\right]\\
\leq&\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\mathrm{vec}(\mathbf{A}_{f,T^2})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{e,T^2})\right.\\
&\left.+(\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\mathrm{vec}(\mathbf{A}_{f,T^2})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{\varepsilon,T^2})\right],
\end{aligned}$$

and when $i_1 \neq i_2$,

$$\begin{aligned}
&E\left[(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)_{i_1j_1}^2(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)_{i_2j_2}^2\right]\\
\leq&\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\mathrm{vec}(\mathbf{A}_{f,T}\otimes\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{e,T^2})\right.\\
&\left.+(\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\mathbf{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\mathrm{vec}(\mathbf{A}_{f,T}\otimes\mathbf{A}_{f,T})^{\mathrm{T}}\mathrm{vec}(\mathbf{A}_{\varepsilon,T^2})\right].
\end{aligned}$$

Thus, by Lemma 3.3, we have for any (i_1, j_1, i_2, j_2),

$$\begin{aligned}
&E\left[(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)^2_{i_1j_1}(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)^2_{i_2j_2}\right] \\
\leq& O\left(T^2\frac{1}{s_{\text{-}k}^2}\right)\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}+(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
E\left\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right\|_F^4 &= \sum_{i_1=1}^{r_k}\sum_{j_1=1}^{d_k}\sum_{i_2=1}^{r_k}\sum_{j_2=1}^{d_k}E\left[(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)^2_{i_1j_1}(\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k)^2_{i_2j_2}\right] \\
&\leq r_k^2O\left(T^2\frac{s_{e,\text{-}k}^2}{s_{\text{-}k}^2}\right)\left\{\left[\mathrm{tr}(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}})\right]^2+\left[\mathrm{tr}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})\right]^2\right\}=O\left(r_k^2T^2d_k^2\frac{d_{\text{-}k}^2}{s_{\text{-}k}^2}\right),
\end{aligned} \tag{3.62}$$

by Assumption (E1). (3.62) together with (3.60) and (3.57) imply $\left\|\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k\right\|_F^2=O_p\left(r_kd_kT\frac{d_{\text{-}k}}{s_{\text{-}k}}\right)$.

Next, we will show $\left\|\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right\|_F^2=O_p\left(r_kd_k\frac{d_{\text{-}k}}{s_{\text{-}k}}\right)$ in a similar manner. First, note that

$$\left(\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right)_{ij}=\frac{1}{T}\left(\sum_{t=1}^{T}\ddot{f}_{t,k,i}\right)\left(\sum_{t=1}^{T}\ddot{e}_{t,k,j}\right),$$

and

$$\begin{aligned}
E\left(\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right)^2_{ij} &= \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{s_1=1}^{T}\sum_{t_2=1}^{T}\sum_{s_2=1}^{T}E\left(\ddot{f}_{t_1,k,i}\ddot{f}_{s_1,k,i}\right)E\left(\ddot{e}_{t_2,k,j}\ddot{e}_{s_2,k,j}\right) \\
&= \frac{1}{T^2}\sum_{t_1,t_2,s_1,s_2=1}^{T}(\mathbf{A}_{f,T})_{t_1s_1}\frac{1}{s_{\text{-}k}}\left[s_{e,\text{-}k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}(\mathbf{A}_{e,T})_{t_2s_2}+\left(\boldsymbol{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}(\mathbf{A}_{\varepsilon,T})_{t_2s_2}\right] \\
&= \frac{1}{T^2s_{\text{-}k}}\left[s_{e,\text{-}k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T}\otimes\mathbf{A}_{e,T}\mathbf{1}_{T^2}+\left(\boldsymbol{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T}\otimes\mathbf{A}_{\varepsilon,T}\mathbf{1}_{T^2}\right] \\
&= O\left(\frac{1}{s_{\text{-}k}}\right)\left[s_{e,\text{-}k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}+\left(\boldsymbol{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\right]
\end{aligned}$$

where the second line follows from (3.58) and (3.59), and the last line follows from Lemma 3.3. Hence,

$$\begin{aligned}
E\left\|\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right\|_F^2 &= \sum_{i=1}^{r_k}\sum_{j=1}^{d_k}O\left(\frac{1}{s_{\text{-}k}}\right)\left[s_{e,\text{-}k}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{jj}+\left(\boldsymbol{\Sigma}_{\varepsilon}^{(k)}\right)_{jj}\right] \\
&= O\left(r_k\frac{1}{s_{\text{-}k}}\right)\left[s_{e,\text{-}k}\mathrm{tr}(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}})+\mathrm{tr}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})\right]=O\left(r_kd_k\frac{d_{\text{-}k}}{s_{\text{-}k}}\right).
\end{aligned} \tag{3.63}$$

Next, we know

$$E\left\|\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right\|_F^4=\frac{1}{T^4}\sum_{i_1=1}^{r_k}\sum_{j_1=1}^{d_k}\sum_{i_2=1}^{r_k}\sum_{j_2=1}^{d_k}E\left[(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_1j_1}^2(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_2j_2}^2\right],$$

where

$$E\left[(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_1j_1}^2(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_2j_2}^2\right]=E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^2\left(\sum_{t=1}^T\ddot{f}_{t,k,i_2}\right)^2\right]E\left[\left(\sum_{t=1}^T\ddot{e}_{t,k,j_1}\right)^2\left(\sum_{t=1}^T\ddot{e}_{t,k,j_2}\right)^2\right].$$

For the terms related to $\ddot{f}$, when $i_1=i_2$,

$$\begin{aligned}E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^2\left(\sum_{t=1}^T\ddot{f}_{t,k,i_2}\right)^2\right]&=E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^4\right]=\sum_{t_1,t_2,s_1,s_2=1}^TE\left[\ddot{f}_{t_1,k,i_1}\ddot{f}_{s_1,k,i_1}\ddot{f}_{t_2,k,i_1}\ddot{f}_{s_2,k,i_1}\right]\\&=\sum_{t_1,t_2,s_1,s_2=1}^T\left(\mathbf{A}_{f,T^2}\right)_{t_1s_1,t_2s_2}=\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{f,T^2}\mathbf{1}_{T^2}=O(T^2)\end{aligned}$$

by Lemma 3.3. When $i_1\neq i_2$,

$$\begin{aligned}E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^2\left(\sum_{t=1}^T\ddot{f}_{t,k,i_2}\right)^2\right]&=E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^2\right]E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_2}\right)^2\right]\\&=\left[\sum_{t=1}^T\sum_{s=1}^TE\left(\ddot{f}_{t,k,i_1}\ddot{f}_{s,k,i_1}\right)\right]^2=\left(\mathbf{1}_T^{\mathrm{T}}\mathbf{A}_{f,T}\mathbf{1}_T\right)^2=O(T^2)\end{aligned}$$

by Lemma 3.3. Therefore, $E\left[\left(\sum_{t=1}^T\ddot{f}_{t,k,i_1}\right)^2\left(\sum_{t=1}^T\ddot{f}_{t,k,i_2}\right)^2\right]=O(T^2)$ for any (i_1,i_2). Finally, for terms with $\ddot{e}$,

$$\begin{aligned}&E\left[\left(\sum_{t=1}^T\ddot{e}_{t,k,j_1}\right)^2\left(\sum_{t=1}^T\ddot{e}_{t,k,j_2}\right)^2\right]=\sum_{t_1=1}^T\sum_{s_1=1}^T\sum_{t_2=1}^T\sum_{s_2=1}^TE\left[\ddot{e}_{t_1,k,j_1}\ddot{e}_{s_1,k,j_1}\ddot{e}_{t_2,k,j_2}\ddot{e}_{s_2,k,j_2}\right]\\
\leq&\sum_{t_1,t_2,s_1,s_2=1}^T\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\left(\mathbf{A}_{e,T^2}\right)_{t_1s_1,t_2s_2}+(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_1j_1}(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_2j_2}\left(\mathbf{A}_{\varepsilon,T^2}\right)_{t_1s_1,t_2s_2}\right]\\
=&\frac{1}{s_{\text{-}k}^2}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{e,T^2}\mathbf{1}_{T^2}+(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_1j_1}(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_2j_2}\mathbf{1}_{T^2}^{\mathrm{T}}\mathbf{A}_{\varepsilon,T^2}\mathbf{1}_{T^2}\right]\\
=&O\left(\frac{T^2}{s_{\text{-}k}^2}\right)\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}+(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_1j_1}(\mathbf{\Sigma}_\varepsilon^{(k)})_{j_2j_2}\right],\end{aligned}$$

where the third line follows from (3.61), and the last line from Lemma 3.3. Therefore,

$$E\left[(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_1j_1}^2(\ddot{\mathbf{F}}_k\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k)_{i_2j_2}^2\right]=O\left(\frac{T^4}{s_{\text{-}k}^2}\right)\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}+(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\right].$$

Hence,

$$\begin{aligned}E\left\|\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right\|_F^4&=O\left(\frac{1}{s_{\text{-}k}^2}\right)\sum_{i_1=1}^{r_k}\sum_{j_1=1}^{d_k}\sum_{i_2=1}^{r_k}\sum_{j_2=1}^{d_k}\left[s_{e,\text{-}k}^2\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_1j_1}\left(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}\right)_{j_2j_2}+(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_1j_1}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})_{j_2j_2}\right]\\&=r_k^2O\left(\frac{1}{s_{\text{-}k}^2}\right)\left\{s_{e,\text{-}k}^2\left[\mathrm{tr}(\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}})\right]^2+\left[\mathrm{tr}(\boldsymbol{\Sigma}_{\varepsilon}^{(k)})\right]^2\right\}=O\left(r_k^2d_k^2\frac{d_{\text{-}k}^2}{s_{\text{-}k}^2}\right).\end{aligned}\tag{3.64}$$

(3.64) together with (3.63) imply $\left\|\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k\right\|_F^2=O_p\left(r_kd_k\frac{d_{\text{-}k}}{s_{\text{-}k}}\right)$. Therefore, we can conclude that

$$\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2=O_p\left(d_k^{\alpha_{k,1}}\right)\left(\left\|\frac{\ddot{\mathbf{F}}_k\ddot{\mathbf{E}}_k}{T}\right\|^2+\left\|\frac{\ddot{\mathbf{F}}_k\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\ddot{\mathbf{E}}_k}{T}\right\|^2\right)=O_p\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}\frac{r_kd_k}{T}d_k^{\alpha_{k,1}}\right).\tag{3.65}$$

Compared with the original rate (3.55), we can obtain a potentially better rate (3.65) by directly bounding and counting the elements in a large matrix. In other words, the rate for $\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2$ will be the minimum between the two, which is

$$\left\|\frac{\mathbf{A}_k\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_k}{T}\right\|^2=O_p\left(\frac{d_{\text{-}k}}{s_{\text{-}k}}d_k^{\alpha_{k,1}}\min\left\{1+\frac{d_k}{T},\frac{r_kd_k}{T}\right\}\right).$$

Combining the rates (3.55), (3.65) and (3.56) yields that

$$\left\|R_k\right\|^2=\frac{d_{\text{-}k}^2}{s_{\text{-}k}^2}\left(1+\frac{d_k^2}{T^2}\right)+\frac{d_{\text{-}k}}{s_{\text{-}k}}d_k^{\alpha_{k,1}}\min\left\{1+\frac{d_k}{T},\frac{r_kd_k}{T}\right\}:=c_k.$$

Hence,

$$\left\|\widehat{\mathbf{U}}_{k,(z_k)}-\mathbf{U}_{k,(z_k)}\right\|^2=O_p\left(d_k^{-2\alpha_{k,z_k}}\left[d_k^{2\alpha_{k,1}}\frac{r_k}{T}+c_k\right]\right).$$

Finally, to show (3.43), let $\widetilde{\mathbf{V}}_k$ be the $z_k \times z_k$ diagonal matrix of the first z_k largest eigenvalues of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_k}$ in decreasing order, and $\ddot{\mathbf{V}}_k := \widetilde{\mathbf{V}}_k / s_{\text{-}k}$. Then it follows from (3.51) that

$$\begin{aligned} \widehat{\mathbf{Q}}_{k,(z_k)} - \mathbf{Q}_k \ddot{\mathbf{H}}_k &= R_k \widehat{\mathbf{Q}}_{k,(z_k)} \ddot{\mathbf{V}}_k^{-1}, \text{ implying} \\ \left\| \widehat{\mathbf{Q}}_{k,(z_k)} - \mathbf{Q}_k \ddot{\mathbf{H}}_k \right\|^2 &\le \left\| R_k \right\|^2 \left\| \ddot{\mathbf{V}}_k^{-1} \right\|^2 O_p\left(d_k^{-2\alpha_{k,z_k}} c_k \right). \end{aligned}$$

This completes the proof of Lemma 3.4.

Proof of Theorem 3.1. Define $s_{\text{-}k,pre} := \frac{1}{M} \sum_{m=1}^{M} s_{\text{-}k,m}$. First, by the definition of $s_{\text{-}k,m}$, we have $\lambda_j \left(\frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) \asymp s_{\text{-}k,m}$ for $j \in [r_k]$ and for each m. Then,

$$\begin{aligned} \lambda_1 \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) &\le \frac{1}{M} \sum_{m=1}^{M} \lambda_1 \left(\frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) \asymp s_{\text{-}k,pre}, \\ \lambda_{r_k} \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) &\ge \frac{1}{M} \sum_{m=1}^{M} \lambda_{r_k} \left(\frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) \asymp s_{\text{-}k,pre}. \end{aligned}$$

Therefore, for $j \in [r_k]$,

$$\lambda_j \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} \right) \asymp s_{\text{-}k,pre}.$$

Next, following the same analysis as in the proof of Lemma 3.2, it is easy to see that

$$\lambda_j \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\mathbf{A}_k \widetilde{\mathbf{F}}_{k,m} \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}} \mathbf{A}_k^{\mathrm{T}}}{T} \right) \asymp s_{\text{-}k,pre} \asymp \lambda_j \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\mathbf{A}_k \widetilde{\mathbf{F}}_{k,m} \left(\mathbf{I}_T - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^{\mathrm{T}} \right) \widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}} \mathbf{A}_k^{\mathrm{T}}}{T} \right),$$

for $j \in [r_k]$. On the other hand, by Lemma 3.1,

$$\lambda_1 \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\widetilde{\mathbf{E}}_{k,m}^{\mathrm{T}} \widetilde{\mathbf{E}}_{k,m}}{T} \right) \le \frac{1}{M} \sum_{m=1}^{M} \lambda_1 \left(\frac{\widetilde{\mathbf{E}}_{k,m}^{\mathrm{T}} \widetilde{\mathbf{E}}_{k,m}}{T} \right) = O_p \left(d_{\text{-}k,pre} \left(1 + \frac{d_k}{T} \right) \right),$$

where $d_{\text{-}k,pre} = \frac{1}{M} \sum_{m=1}^{M} d_{\text{-}k,m}$. Similarly,

$$\lambda_1 \left(\frac{1}{M} \sum_{m=1}^{M} \frac{\widetilde{\mathbf{E}}_{k,m}^{\mathrm{T}} \left(\mathbf{I}_T - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^{\mathrm{T}} \right) \widetilde{\mathbf{E}}_{k,m}}{T} \right) \le O_p \left(d_{\text{-}k,pre} \left(1 + \frac{d_k}{T} \right) \right).$$

If Assumptions (E1) – (R1) are satisfied for all M chosen samples, then $\frac{d_{\text{-}k,m}}{s_{\text{-}k,m}} \left(1 + \frac{d_k}{T} \right) = o\left(d_k^{\alpha_{k,z_k}} \right)$ for all m, which implies $\frac{d_{\text{-}k,pre}}{s_{\text{-}k,pre}} \left(1 + \frac{d_k}{T} \right) = o\left(d_k^{\alpha_{k,z_k}} \right)$. Therefore, if we define the

new normalized version $\ddot{\mathbf{X}}_{k,m} = \widetilde{\mathbf{X}}_{k,m}/s^{\frac{1}{2}}_{k,pre}$, and $\ddot{\mathbf{F}}_{k,m}$ and $\ddot{\mathbf{E}}_{k,m}$ similarly, then following (3.49) and (3.50), (we can write $\frac{1}{M}\sum_{m=1}^{M}\frac{\ddot{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_{k,m}}{T} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ by Cholesky decomposition, and similarly define $\mathbf{F}$ and $\mathbf{E}$, so that $\mathbf{X} = \mathbf{F}+\mathbf{E}$.) we have

$$\lambda_j\left(\frac{1}{M}\sum_{m=1}^{M}\frac{\ddot{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_{k,m}}{T}\right) \asymp 1$$

for $j \in [r_k]$. Then, we can make use of the decomposition

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}\frac{\ddot{\mathbf{X}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{X}}_{k,m}}{T} &= \frac{1}{M}\sum_{m=1}^{M}\frac{\left(\mathbf{A}_k\ddot{\mathbf{F}}_{k,m}+\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\right)\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\left(\mathbf{A}_k\ddot{\mathbf{F}}_{k,m}+\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\right)^{\mathrm{T}}}{T}\\ &= \mathbf{A}_k\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T}\right]\mathbf{A}_k^{\mathrm{T}}+R_{k,pre},\end{aligned} \tag{3.66}$$

where

$$\begin{aligned}R_{k,pre} = \frac{1}{M}\sum_{m=1}^{M}\Bigg[&\frac{\mathbf{A}_k\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T}+\frac{\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T}\\ &+\frac{\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T}\Bigg].\end{aligned} \tag{3.67}$$

Then, following similar argument as in the proof of Lemma 3.4, we can apply Lemma 3 of Lam et al. (2011) and know that there exists $\widehat{\mathbf{U}}_{k,pre,(z_k)}$ (with $\widehat{\mathbf{U}}_{k,pre,(z_k)}^{\mathrm{T}}\widehat{\mathbf{U}}_{k,pre,(z_k)} = \mathbf{I}_{z_k}$) such that $\widehat{\mathbf{Q}}_{k,pre,(z_k)} = \widehat{\mathbf{U}}_{k,pre,(z_k)}\mathbf{P}_{k,pre,(z_k)}$ where $\mathbf{P}_{k,pre,(z_k)}$ is an orthogonal matrix, so that

$$\left\|\widehat{\mathbf{U}}_{k,pre,(z_k)}-\mathbf{U}_{k,(z_k)}\right\| \preceq \frac{d_k^{\alpha_{k,1}}\left\|\frac{1}{M}\sum_{m=1}^{M}\frac{\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T}-\mathbf{I}_{r_k}\right\|+\left\|R_{k,pre}\right\|}{d_k^{\alpha_{k,z_k}}}.$$

To proceed, note that by (3.54), we have $\left\| \frac{\widetilde{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\widetilde{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T s_{-k,m}} - \mathbf{I}_{r_k} \right\| = O_p\left(\sqrt{r_k/T}\right)$ for each m, so $\left\| \frac{\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} - \frac{s_{-k,m}}{s_{-k,pre}}\mathbf{I}_{r_k} \right\| = O_p\left(\frac{s_{-k,m}}{s_{-k,pre}}\sqrt{\frac{r_k}{T}}\right)$, and thus,

$$\begin{aligned}
\left\| \frac{1}{M}\sum_{m=1}^{M} \frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - \mathbf{I}_{r_k} \right\| &= \frac{1}{M}\left\| \sum_{m=1}^{M} \frac{\ddot{\mathbf{F}}_k\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_k^{\mathrm{T}}}{T} - M\mathbf{I}_{r_k} \right\| \\
&= \frac{1}{M}\left\| \sum_{m=1}^{M} \left[\frac{\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} - \frac{s_{-k,m}}{s_{-k,pre}}\mathbf{I}_{r_k} \right] \right\| \\
&\le \frac{1}{M}\sum_{m=1}^{M} \left\| \frac{\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}}{T} - \frac{s_{-k,m}}{s_{-k,pre}}\mathbf{I}_{r_k} \right\| \\
&= \frac{1}{M}\sum_{m=1}^{M} O_p\left(\frac{s_{-k,m}}{s_{-k,pre}}\sqrt{\frac{r_k}{T}}\right) = O_p\left(\sqrt{\frac{r_k}{T}}\right).
\end{aligned}$$

To bound $\left\|R_{k,pre}\right\|^2$, we can bound each term on the right hand side of (3.67) by using the similar technique as in the proof of Lemma 3.4. We have that

$$\begin{aligned}
\left\| \frac{1}{M}\sum_{m=1}^{M} \frac{\mathbf{A}_k\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T} \right\|^2 &\le \frac{1}{M}\sum_{m=1}^{M} \left\| \frac{\mathbf{A}_k\ddot{\mathbf{F}}_{k,m}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T} \right\|^2 \\
&= O_p\left(d_k^{\alpha_{k,1}} \min\left\{1+\frac{d_k}{T}, \frac{r_k d_k}{T}\right\} \frac{\frac{1}{M}\sum_{m=1}^{M} d_{-k,m}s_{-k,m}}{s_{-k,pre}^2} \right),
\end{aligned}$$

and similarly,

$$\left\| \frac{1}{M}\sum_{m=1}^{M} \frac{\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{F}}_{k,m}^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}}{T} \right\|^2 = O_p\left(d_k^{\alpha_{k,1}} \min\left\{1+\frac{d_k}{T}, \frac{r_k d_k}{T}\right\} \frac{\frac{1}{M}\sum_{m=1}^{M} d_{-k,m}s_{-k,m}}{s_{-k,pre}^2} \right),$$

and

$$\begin{aligned}
\left\| \frac{1}{M}\sum_{m=1}^{M} \frac{\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T} \right\|^2 &\le \frac{1}{M}\sum_{m=1}^{M} \left\| \frac{\ddot{\mathbf{E}}_{k,m}^{\mathrm{T}}\left(\mathbf{I}_T-\frac{1}{T}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}\right)\ddot{\mathbf{E}}_{k,m}}{T} \right\|^2 \\
&= O_p\left(\left(1+\frac{d_k^2}{T^2}\right) \frac{\frac{1}{M}\sum_{m=1}^{M} d_{-k,m}^2}{s_{-k,pre}^2} \right).
\end{aligned}$$

Combining the above results give

$$c_{k,pre} := \left\|R_{k,pre}\right\|^2 = O_p\left(\min\left\{1+\frac{d_k}{T}, \frac{r_k d_k}{T}\right\} \frac{\frac{1}{M}\sum_{m=1}^{M} d_{-k,m}s_{-k,m}}{s_{-k,pre}^2} + d_k^{\alpha_{k,1}} \left(1+\frac{d_k^2}{T^2}\right) \frac{\frac{1}{M}\sum_{m=1}^{M} d_{-k,m}^2}{s_{-k,pre}^2} \right),$$

and thus,

$$\left\|\widehat{\mathbf{U}}_{k,pre,(z_k)} - \mathbf{U}_{k,(z_k)}\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}}\left[d_k^{2\alpha_{k,1}}\frac{r_k}{T} + c_{k,pre}\right]\right). \tag{3.68}$$

Further, let $\widetilde{\mathbf{V}}_{k,pre}$ be the $z_k \times z_k$ diagonal matrix of the first z_k largest eigenvalues of $\widehat{\mathbf{\Sigma}}_{\widetilde{\mathbf{x}}_{k,agg}}$ in decreasing order, and $\ddot{\mathbf{V}}_{k,pre} := \widetilde{\mathbf{V}}_{k,pre}/s_{\text{-}k,pre}$. Then it follows from (3.66) that

$$\begin{aligned}
&\widehat{\mathbf{Q}}_{k,pre,(z_k)} - \mathbf{Q}_k\ddot{\mathbf{H}}_{k,pre} = R_{k,pre}\widehat{\mathbf{Q}}_{k,pre,(z_k)}\ddot{\mathbf{V}}_{k,pre}^{-1}, \text{ implying} \\
&\left\|\widehat{\mathbf{Q}}_{k,pre,(z_k)} - \mathbf{Q}_k\ddot{\mathbf{H}}_{k,pre}\right\|^2 \leq \left\|R_{k,pre}\right\|^2\left\|\ddot{\mathbf{V}}_{k,pre}^{-1}\right\|^2 = O_p\left(d_k^{-2\alpha_{k,z_k}}c_{k,pre}\right).
\end{aligned} \tag{3.69}$$

Finally, (3.18) and (3.19) follow as we substitute $d_{\text{-}k,m} \asymp d_{\text{-}k}$ and $s_{\text{-}k,m} \asymp s_{\text{-}k,max}$ into (3.69) and (3.68), which is guaranteed by our choice of $S_{k,m}$ and M_0 as discussed in Section 3.2.3 and 3.2.4. This completes the proof of Theorem 3.1. □

Before the proof of Theorem 3.2, we decompose

$$\begin{aligned}
\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)} &= T^{-1}\sum_{t=1}^{T}\mathbf{y}_{t,m+1}^{(k)}\mathbf{y}_{t,m+1}^{(k)\mathrm{T}} = T^{-1}\sum_{t=1}^{T}\mathrm{mat}_k(\mathscr{X}_t - \bar{\mathscr{X}})\breve{\mathbf{q}}_{\text{-}k}^{(m)}\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathrm{mat}_k^{\mathrm{T}}(\mathscr{X}_t - \bar{\mathscr{X}}) \\
&= T^{-1}\sum_{t=1}^{T}\left(\mathbf{A}_k\mathrm{mat}_k(\mathscr{F}_t - \bar{\mathscr{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\breve{\mathbf{q}}_{\text{-}k}^{(m)} + \mathrm{mat}_k(\mathscr{E}_t - \bar{\mathscr{E}})\breve{\mathbf{q}}_{\text{-}k}^{(m)}\right)^{\otimes 2} \\
&= T^{-1}\sum_{t=1}^{T}\left\{\mathbf{A}_k\mathrm{mat}_k(\mathscr{F}_t - \bar{\mathscr{F}})\mathbf{A}_{\text{-}k}^{\mathrm{T}}\breve{\mathbf{q}}_{\text{-}k}^{(m)} + \mathbf{A}_{e,k}\mathrm{mat}_k(\mathscr{F}_{e,t} - \bar{\mathscr{F}}_{e,\cdot})\mathbf{A}_{e,\text{-}k}^{\mathrm{T}}\breve{\mathbf{q}}_{\text{-}k}^{(m)}\right. \\
&\qquad \left. + \left[\left(\mathbf{\Sigma}_{\varepsilon,1}^{(k)}\right)^{1/2}(\boldsymbol{\varepsilon}_{t,1}^{(k)} - \bar{\boldsymbol{\varepsilon}}_{\cdot,1}^{(k)}), \ldots, \left(\mathbf{\Sigma}_{\varepsilon,d_{\text{-}k}}^{(k)}\right)^{1/2}(\boldsymbol{\varepsilon}_{t,d_{\text{-}k}}^{(k)} - \bar{\boldsymbol{\varepsilon}}_{\cdot,d_{\text{-}k}}^{(k)})\right]\breve{\mathbf{q}}_{\text{-}k}^{(m)}\right\}^{\otimes 2} \\
&= \sum_{i=1}^{3}\widetilde{\mathbf{S}}_{ii,m} + \sum_{i,j=1;i<j}^{3}(\widetilde{\mathbf{S}}_{ij,m} + \widetilde{\mathbf{S}}_{ij,m}^{\mathrm{T}}), \text{ where}
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathbf{S}}_{11,m} &:= \mathbf{A}_k(\mathbf{I}_{r_k}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})T^{-1}\widetilde{\mathbf{F}}^{(k)}\mathbf{M}_T\widetilde{\mathbf{F}}^{(k)\mathrm{T}}(\mathbf{I}_{r_k}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_{12,m} &:= \mathbf{A}_k(\mathbf{I}_{r_k}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})T^{-1}\widetilde{\mathbf{F}}^{(k)}\mathbf{M}_T\widetilde{\mathbf{\Theta}}^{(k)\mathrm{T}}\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)}, \ldots, \mathbf{\Sigma}_{\varepsilon,d_{\text{-}k}}^{(k)})(\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\otimes\mathbf{I}_{d_k})^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_{13,m} &:= \mathbf{A}_k(\mathbf{I}_{r_k}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})T^{-1}\widetilde{\mathbf{F}}^{(k)}\mathbf{M}_T\widetilde{\mathbf{F}}_e^{(k)\mathrm{T}}(\mathbf{I}_{r_{e,k}}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{e,\text{-}k})^{\mathrm{T}}\mathbf{A}_{e,k}^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_{22,m} &:= (\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\otimes\mathbf{I}_{d_k})\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)}, \ldots, \mathbf{\Sigma}_{\varepsilon,d_{\text{-}k}}^{(k)})T^{-1}\widetilde{\mathbf{\Theta}}^{(k)}\mathbf{M}_T\widetilde{\mathbf{\Theta}}^{(k)}\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)}, \ldots, \mathbf{\Sigma}_{\varepsilon,d_{\text{-}k}}^{(k)})(\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\otimes\mathbf{I}_{d_k})^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_{23,m} &:= (\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\otimes\mathbf{I}_{d_k})\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)}, \ldots, \mathbf{\Sigma}_{\varepsilon,d_{\text{-}k}}^{(k)})T^{-1}\widetilde{\mathbf{\Theta}}^{(k)}\mathbf{M}_T\widetilde{\mathbf{F}}_e^{(k)\mathrm{T}}(\mathbf{I}_{r_{e,k}}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{e,\text{-}k})^{\mathrm{T}}\mathbf{A}_{e,k}^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_{33,m} &:= \mathbf{A}_{e,k}(\mathbf{I}_{r_{e,k}}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{e,\text{-}k})T^{-1}\widetilde{\mathbf{F}}_e^{(k)}\mathbf{M}_T\widetilde{\mathbf{F}}_e^{(k)\mathrm{T}}(\mathbf{I}_{r_{e,k}}\otimes\breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{e,\text{-}k})^{\mathrm{T}}\mathbf{A}_{e,k}^{\mathrm{T}},
\end{aligned}$$

with $\mathbf{M}_T = \mathbf{I}_T - T^{-1}\mathbf{1}_T\mathbf{1}_T^{\mathrm{T}}$, and

$$\begin{aligned}
\widetilde{\mathbf{F}}^{(k)} &:= [\mathrm{vec}(\mathrm{mat}_k^{\mathrm{T}}(\mathscr{F}_1)),\ldots,\mathrm{vec}(\mathrm{mat}_k^{\mathrm{T}}(\mathscr{F}_T))] =: [\mathbf{f}_1^{(k)},\ldots,\mathbf{f}_T^{(k)}],\\
\widetilde{\mathbf{F}}_e^{(k)} &:= [\mathrm{vec}(\mathrm{mat}_k^{\mathrm{T}}(\mathscr{F}_{e,1})),\ldots,\mathrm{vec}(\mathrm{mat}_k^{\mathrm{T}}(\mathscr{F}_{e,T}))] =: [\mathbf{f}_{e,1}^{(k)},\ldots,\mathbf{f}_{e,T}^{(k)}], \text{ where}\\
\mathbf{f}_t^{(k)} &:= (f_{t,1,1}^{(k)},\ldots,f_{t,r_{-k},1}^{(k)},\ldots,f_{t,1,r_k}^{(k)},\ldots,f_{t,r_{-k},r_k}^{(k)})^{\mathrm{T}},\\
\mathbf{f}_{e,t}^{(k)} &:= (f_{e,t,1,1}^{(k)},\ldots,f_{e,t,r_{e,-k},1}^{(k)},\ldots,f_{e,t,1,r_{e,k}}^{(k)},\ldots,f_{e,t,r_{e,-k},r_{e,k}}^{(k)})^{\mathrm{T}}, \text{ and}\\
\widetilde{\boldsymbol{\Theta}}^{(k)} &:= [\boldsymbol{\varepsilon}_1^{(k)},\ldots,\boldsymbol{\varepsilon}_T^{(k)}] \text{ with } \boldsymbol{\varepsilon}_t^{(k)} := (\boldsymbol{\varepsilon}_{t,1}^{(k)\mathrm{T}},\ldots,\boldsymbol{\varepsilon}_{t,d_{-k}}^{(k)\mathrm{T}})^{\mathrm{T}}.
\end{aligned}$$

With the notations in Assumption (E2) and (F1), for $t \in [T]$, define

$$\begin{aligned}
\mathbf{z}_{e,t}^{(k)} &:= (z_{e,t,1,1}^{(k)},\ldots,z_{e,t,r_{e,-k},1}^{(k)},\ldots,z_{e,t,1,r_{e,k}}^{(k)},\ldots,z_{e,t,r_{e,-k},r_{e,k}}^{(k)})^{\mathrm{T}},\\
\mathbf{z}_{\varepsilon,t}^{(k)} &:= (z_{\varepsilon,t,1,1}^{(k)},\ldots,z_{\varepsilon,t,1,d_k}^{(k)},\ldots,z_{\varepsilon,t,d_{-k},1}^{(k)},\ldots,z_{\varepsilon,t,d_{-k},d_k}^{(k)})^{\mathrm{T}},\\
\mathbf{z}_{f,t}^{(k)} &:= (z_{f,t,1,1}^{(k)},\ldots,z_{f,t,r_{-k},1}^{(k)},\ldots,z_{f,t,1,r_k}^{(k)},\ldots,z_{f,t,r_{-k},r_k}^{(k)})^{\mathrm{T}}.
\end{aligned}$$

We further split, for a fixed integer $N \geq 1$,

$$\begin{aligned}
&\widetilde{\mathbf{F}}^{(k)} := \mathbf{F}^{(k)} + \check{\mathbf{F}}^{(k)},\quad \widetilde{\mathbf{F}}_e^{(k)} := \mathbf{F}_e^{(k)} + \check{\mathbf{F}}_e^{(k)},\quad \widetilde{\boldsymbol{\Theta}}^{(k)} := \boldsymbol{\Theta}^{(k)} + \check{\boldsymbol{\Theta}}^{(k)}, \text{ where}\\
&\mathbf{F}^{(k)} := \left(\sum_{q=0}^{NT} a_{f,q}\mathbf{z}_{f,1-q}^{(k)},\ldots,\sum_{q=0}^{NT} a_{f,q}\mathbf{z}_{f,T-q}^{(k)}\right) = (\mathbf{z}_{f,1-NT}^{(k)},\ldots,\mathbf{z}_{f,T}^{(k)})\mathscr{A}_{f,T} =: \mathbf{Z}_f^{(k)}\mathscr{A}_{f,T},\\
&\mathbf{F}_e^{(k)} := \left(\sum_{q=0}^{NT} a_{e,q}\mathbf{z}_{e,1-q}^{(k)},\ldots,\sum_{q=0}^{NT} a_{e,q}\mathbf{z}_{e,T-q}^{(k)}\right) = (\mathbf{z}_{e,1-NT}^{(k)},\ldots,\mathbf{z}_{e,T}^{(k)})\mathscr{A}_{e,T} =: \mathbf{Z}_e^{(k)}\mathscr{A}_{e,T},\\
&\boldsymbol{\Theta}^{(k)} := \left(\sum_{q=0}^{NT} a_{\varepsilon,q}\mathbf{z}_{\varepsilon,1-q}^{(k)},\ldots,\sum_{q=0}^{NT} a_{\varepsilon,q}\mathbf{z}_{\varepsilon,T-q}^{(k)}\right) = (\mathbf{z}_{\varepsilon,1-NT}^{(k)},\ldots,\mathbf{z}_{\varepsilon,T}^{(k)})\mathscr{A}_{\varepsilon,T} =: \mathbf{Z}_\varepsilon^{(k)}\mathscr{A}_{\varepsilon,T},
\end{aligned}$$

with $\mathscr{A}_{f,T}$, $\mathscr{A}_{e,T}$ and $\mathscr{A}_{\varepsilon,T}$ defined in Assumption (RE1), and $\check{\mathbf{F}}^{(k)}$, $\check{\mathbf{F}}_e^{(k)}$ and $\check{\boldsymbol{\Theta}}^{(k)}$ the remainders of the truncations. Then we define

$$\begin{aligned}
&\widetilde{\mathbf{S}}_{ij,m} = \mathbf{S}_{ij,m} + \check{\mathbf{S}}_{ij,m}, \text{ where}\\
&\mathbf{S}_{11,m} := \mathbf{A}_k(\mathbf{I}_{r_k} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k})T^{-1}\mathbf{F}^{(k)}\mathbf{M}_T\mathbf{F}^{(k)\mathrm{T}}(\mathbf{I}_{r_k} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k})^{\mathrm{T}}\mathbf{A}_k^{\mathrm{T}},
\end{aligned} \tag{3.70}$$

and similarly for other $\mathbf{S}_{ij,m}$'s. We first prove a lemma on how large the quadratic form

$$\widehat{g}_{-k} := \check{\mathbf{q}}_{-k}^{(0)\mathrm{T}}\mathbf{A}_{-k}\mathbf{A}_{-k}^{\mathrm{T}}\check{\mathbf{q}}_{-k}^{(0)} \tag{3.71}$$

is, before showing how small each remainder term $\check{\mathbf{S}}_{ij,0}$ is in the next lemma.

Lemma 3.5. *Let all the assumptions in Theorem 3.2 be satisfied. Then*

$$\widehat{g}_{\text{-}k} \asymp_P \prod_{j=1;j\neq k}^{K} d_j^{\alpha_{j,1}}.$$

To put some perspectives into Lemma 3.5, recall that $\check{\mathbf{q}}_{\text{-}k}^{(0)} = \widehat{\mathbf{q}}_{\text{-}k,pre} = \widehat{\mathbf{U}}_{\text{-}k,pre,(1)}$ is estimating the direction $\mathbf{U}_{\text{-}k,(1)}$. Hence the quadratic form $\widehat{g}_{\text{-}k}$ is estimating

$$\mathbf{U}_{\text{-}k,(1)}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} = \mathbf{U}_{\text{-}k,(1)}^{\mathrm{T}}\mathbf{U}_{\text{-}k}\mathbf{G}_{\text{-}k}\mathbf{U}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} = (\mathbf{G}_{\text{-}k})_{11} \asymp \prod_{j=1;j\neq k}^{K} d_j^{\alpha_{j,1}},$$

where is the last rate is the result of (3.31) in Lemma 3.2. Hence $\widehat{g}_{\text{-}k}$ has the same rate as the quadratic form that it is estimating.

Proof of Lemma 3.5.

$$\begin{aligned}
\widehat{g}_{\text{-}k} &= \widehat{\mathbf{q}}_{\text{-}k,pre}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\widehat{\mathbf{q}}_{\text{-}k,pre} = \widehat{\mathbf{U}}_{\text{-}k,pre,(1)}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} \\
&= \mathbf{U}_{\text{-}k,(1)}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} + (\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)})^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}(\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)}) \\
&\quad + 2(\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)})^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} \\
&\geq \mathbf{U}_{\text{-}k,(1)}{}^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} + 2(\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)})^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} \\
&= \mathbf{U}_{\text{-}k,(1)}^{\mathrm{T}}\mathbf{U}_{\text{-}k}\mathbf{G}_{\text{-}k}\mathbf{U}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} + 2(\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)})^{\mathrm{T}}\mathbf{U}_{\text{-}k}\mathbf{G}_{\text{-}k}\mathbf{U}_{\text{-}k}^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)},
\end{aligned} \tag{3.72}$$

where the last line used the singular value decomposition of $\mathbf{A}_k$ in (3.5), and

$$\begin{aligned}
\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} &:= \mathbf{U}_{K,pre,(1)} \otimes \cdots \otimes \mathbf{U}_{k+1,pre,(1)} \otimes \mathbf{U}_{k-1,pre,(1)} \otimes \cdots \otimes \mathbf{U}_{1,pre,(1)}, \\
\mathbf{U}_{\text{-}k} &:= \mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_{k+1} \otimes \mathbf{U}_{k-1} \otimes \cdots \otimes \mathbf{U}_1, \quad \mathbf{G}_{\text{-}k} := \mathbf{G}_K \otimes \cdots \otimes \mathbf{G}_{k+1} \otimes \mathbf{G}_{k-1} \otimes \cdots \otimes \mathbf{G}_1, \\
\mathbf{U}_{\text{-}k,(1)} &:= \mathbf{U}_{K,(1)} \otimes \cdots \mathbf{U}_{k+1,(1)} \otimes \mathbf{U}_{k-1,(1)} \otimes \cdots \otimes \mathbf{U}_{1,(1)}.
\end{aligned}$$

But (3.31) in Lemma 3.2 proves that $\mathbf{G}_k$ has the j-th diagonal element with order $d_k^{\alpha_{k,j}}$ for $j \in [r_k]$ under Assumption (L1), which is the same as those in $\mathbf{D}_k$. At the same time, induction (proof omitted) easily gives

$$\left\|\widehat{\mathbf{U}}_{\text{-}k,pre,(1)} - \mathbf{U}_{\text{-}k,(1)}\right\| = O_P\left(\sum_{j=1;j\neq k}^{K}\left\|\widehat{\mathbf{U}}_{j,pre,(1)} - \mathbf{U}_{j,(1)}\right\|\right) = O_P\left(K\sqrt{\frac{r_{\max}}{T}} + \sum_{j=1;j\neq k}^{K} d_j^{-\alpha_{j,1}} c_{j,max}^{1/2}\right), \tag{3.73}$$

with $r_{\max} := \max_{k\in[K]} r_k$. Hence from the decomposition after (3.72), we then have, assuming the above rate is $o(1)$,

$$\widehat{g}_{-k} \asymp (1+o_P(1))(\mathbf{G}_{-k})_{11} \asymp_P \prod_{j=1;j\neq k}^{K} d_j^{\alpha_{j,1}}.$$

Lemma 3.6. *Let all the assumptions in Theorem 3.2 be satisfied. Then for each $k \in [K]$,*

$$\sum_{i,j=1;i\leq j}^{3} \left\|\check{\mathbf{S}}_{ij,0}\right\| = o_P(T^{-1}).$$

Proof of Lemma 3.6. Define

$$\begin{aligned}
\mathbf{Q}_{1,m}^{(k)} &:= \mathbf{A}_k(\mathbf{I}_{r_k} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{-k}), \\
\mathbf{Q}_{2,m}^{(k)} &:= (\check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \otimes \mathbf{I}_{d_k}) \mathrm{diag}^{1/2}(\boldsymbol{\Sigma}_{\varepsilon,1}^{(k)}, \ldots, \boldsymbol{\Sigma}_{\varepsilon,d_{-k}}^{(k)}), \\
\mathbf{Q}_{3,m}^{(k)} &:= \mathbf{A}_{e,k}(\mathbf{I}_{r_{e,k}} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{e,-k}).
\end{aligned} \tag{3.74}$$

Then

$$\begin{aligned}
&\left\|\mathbf{Q}_{1,0}^{(k)}\right\|^2 = (\check{\mathbf{q}}_{-k}^{(0)\mathrm{T}} \mathbf{A}_{-k} \mathbf{A}_{-k}^{\mathrm{T}} \check{\mathbf{q}}_{-k}^{(0)}) \left\|\mathbf{A}_k\right\|^2 = O_P\Big(\prod_{j=1;j\neq k}^{K} d_j^{\alpha_{j,1}} \cdot d_k^{\alpha_{k,1}} \Big) = O_P(g_s), \\
&\left\|\mathbf{Q}_{2,0}^{(k)}\right\|^2 = O_P\Big(\max_{j\in[d_{-k}]} \left\|\boldsymbol{\Sigma}_{\varepsilon,j}^{(k)}\right\| \Big) = O_P(1), \quad \left\|\mathbf{Q}_{3,0}^{(k)}\right\|^2 = (\check{\mathbf{q}}_{-k}^{(0)\mathrm{T}} \mathbf{A}_{e,-k} \mathbf{A}_{e,-k}^{\mathrm{T}} \check{\mathbf{q}}_{-k}^{(0)}) \left\|\mathbf{A}_{e,k}\right\|^2 = O_P(1),
\end{aligned}$$

where the first line used Lemma 3.5, and the results for $\mathbf{Q}_{2,0}^{(k)}$ and $\mathbf{Q}_{3,0}^{(k)}$ used finiteness of K and Assumption (E1). We also have, by Assumption (RE1),

$$\begin{aligned}
&E\left\|T^{-1/2}\mathbf{F}^{(k)}\right\|_F^2 = T^{-1} E\mathrm{tr}(\mathscr{A}_{f,T}^{\mathrm{T}} \mathbf{Z}_f^{(k)\mathrm{T}} \mathbf{Z}_f^{(k)} \mathscr{A}_{f,T}) = rT^{-1}\mathrm{tr}(\mathscr{A}_{f,T}\mathscr{A}_{f,T}^{\mathrm{T}}) = r(1-o(T^{-2}d^{-2})), \\
&E\left\|T^{-1/2}\check{\mathbf{F}}^{(k)}\right\|_F^2 = O(rT\cdot T^{-1}\cdot o(T^{-2}d^{-2})) = o(rT^{-2}d^{-2}), \\
&E\left\|T^{-1/2}\boldsymbol{\Theta}^{(k)}\right\|_F^2 = d(1-o(T^{-2}d^{-2})), \quad E\left\|T^{-1/2}\check{\boldsymbol{\Theta}}^{(k)}\right\|_F^2 = o(dT^{-2}d^{-2}) = o(T^{-2}d^{-1}), \\
&E\left\|T^{-1/2}\mathbf{F}_e^{(k)}\right\|_F^2 = r_e(1-o(T^{-2}d^{-2})), \quad E\left\|T^{-1/2}\check{\mathbf{F}}_e^{(k)}\right\|_F^2 = o(r_e T^{-2}d^{-2}).
\end{aligned}$$

Then writing $\check{\mathbf{S}}_{11,0} = I_1 + I_1^{\mathrm{T}} + I_2$, where

$$I_1 := T^{-1}\mathbf{Q}_{1,0}^{(k)}\check{\mathbf{F}}^{(k)}\mathbf{M}_T\mathbf{F}^{(k)}\mathbf{Q}_{1,0}^{(k)\mathrm{T}}, \quad I_2 := T^{-1}\mathbf{Q}_{1,0}^{(k)}\check{\mathbf{F}}^{(k)}\mathbf{M}_T\check{\mathbf{F}}^{(k)}\mathbf{Q}_{1,0}^{(k)\mathrm{T}},$$

we have for $a > 0$,

$$P(\|I_1\| \geq a) \leq \|\mathbf{Q}_{1,0}^{(k)}\|^2 E^{1/2} \|T^{-1/2}\mathbf{F}^{(k)}\|_F^2 \cdot E^{1/2} \|T^{-1/2}\check{\mathbf{F}}^{(k)}\|_F^2 / a$$
$$= o(g_s \cdot r^{1/2} \cdot r^{1/2} T^{-2/2} d^{-2/2}/a) = o(r g_s T^{-1} d^{-1}/a) = o(T^{-1}),$$

showing that $\|I_1\| = o_P(T^{-1})$. Similarly (details omitted), $\|I_2\| = o_P(T^{-2}d^{-1})$, so that $\|\check{\mathbf{S}}_{11,0}\| = o_P(T^{-1})$. Similar to the above arguments, we can show that (details omitted)

$$\|\check{\mathbf{S}}_{12,0}\| = o_P(r^{1/2} g_s^{1/2} T^{-1} d^{-1/2}) = o_P(T^{-1}), \quad \|\check{\mathbf{S}}_{13,0}\| = o_P(g_s^{1/2} r^{1/2} r_e^{1/2} T^{-1} d^{-1}) = o_P(T^{-1}),$$
$$\|\check{\mathbf{S}}_{22,0}\| = o_P(d^{1/2} \cdot T^{-1} d^{-1/2}) = o_P(T^{-1}), \quad \|\check{\mathbf{S}}_{23,0}\| = o_P(r_e^{1/2} T^{-1} d^{-1/2}) = o_P(T^{-1}),$$
$$\|\check{\mathbf{S}}_{33,0}\| = o_P(r_e^{1/2} \cdot r_e^{1/2} T^{-1} d^{-1}) = o_P(T^{-1}).$$

This completes the proof of the lemma. □

Before presenting the next lemma, define, for $j \in [d_{-k}]$ and $k \in [K]$, and m a non-negative integer,

$$\mathbf{S}'_{11,m} = (\check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{-k} \mathbf{A}_{-k}^{\mathrm{T}} \check{\mathbf{q}}_{-k}^{(m)}) \mathbf{A}_k \mathbf{A}_k^{\mathrm{T}},$$
$$\mathbf{S}''_{11,m} = \mathbf{A}_k (\mathbf{I}_{r_k} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{-k}) (T^{-1} \mathbf{F}^{(k)} \mathbf{M}_T \mathbf{F}^{(k)\mathrm{T}} - \mathbf{I}_r) (\mathbf{I}_{r_k} \otimes \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{-k})^{\mathrm{T}} \mathbf{A}_k^{\mathrm{T}},$$

so that $\mathbf{S}_{11,m} = \mathbf{S}'_{11,m} + \mathbf{S}''_{11,m}$, where $\mathbf{S}_{11,m}$ is defined in (3.70). Then

$$\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)} = T^{-1} \sum_{t=1}^{T} \mathbf{y}_{t,m+1}^{(k)} \mathbf{y}_{t,m+1}^{(k)\mathrm{T}} = \mathbf{S}'_{11,m} + \mathbf{S}''_{11,m} + \mathbf{S}_{22,m} + \mathbf{S}_{33,m} + \sum_{i,j=1;i<j}^{3} (\mathbf{S}_{ij,m} + \mathbf{S}_{ij,m}^{\mathrm{T}}) + \sum_{i,j=1}^{3} \check{\mathbf{S}}_{ij,m}$$
$$=: \mathbf{S}'_{11,m} + \mathbf{E}_m. \tag{3.75}$$

Lemma 3.7. *Let all the assumptions in Theorem 3.2 be satisfied. Then for $j \in [d_{-k}]$ and $k \in [K]$, defining $\check{g}_{-k}^{(m)} := \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{-k} \mathbf{A}_{-k}^{\mathrm{T}} \check{\mathbf{q}}_{-k}^{(m)}$, we have*

$$\|\mathbf{S}''_{11,m}\| = O_P\left(\check{g}_{-k}^{(m)} d_k^{\alpha_{k,1}} \sqrt{\frac{r}{T}}\right), \quad \|\mathbf{S}_{12,m}\| = O_P\left\{(\check{g}_{-k}^{(m)})^{1/2} d_k^{\alpha_{k,1}/2} \left(\sqrt{\frac{r d_k}{T}} + \|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{rd}{T}}\right)\right\},$$
$$\|\mathbf{S}_{13,m}\| = O_P\left((\check{g}_{-k}^{(m)})^{1/2} d_k^{\alpha_{k,1}/2} \sqrt{\frac{r r_e}{T}}\right),$$
$$\|\mathbf{S}_{22,m}\| = O_P\left(1 + \|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\|^2 \frac{d}{T} + \frac{d_k}{\sqrt{T}} + \|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{d_k d}{T}}\right),$$
$$\|\mathbf{S}_{23,m}\| = O_P\left\{\sqrt{\frac{r_e d_k}{T}} + \|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{r_e d}{T}}\right\}, \quad \|\mathbf{S}_{33,m}\| = O_P(1),$$

where $\mathbf{U}_{-k,(1)} := \mathbf{U}_{K,(1)} \otimes \cdots \otimes \mathbf{U}_{k+1,(1)} \otimes \mathbf{U}_{k-1,(1)} \otimes \cdots \otimes \mathbf{U}_{1,(1)}$.

Proof of Lemma 3.7. Using (3.5) and the fact that (3.31) proves that $\lambda_j(\mathbf{G}_k) \asymp d_k^{\alpha_{k,j}}$ for $j \in [r_k]$, we have

$$\begin{aligned}\left\|\mathbf{S}''_{11,m}\right\| &\leq (\check{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k}\mathbf{A}_{-k}^{\mathrm{T}}\check{\mathbf{q}}_{-k}^{(m)})\left\|\mathbf{A}_k\right\|^2 \cdot \left\|T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}} - \mathbf{I}_r\right\| \\ &= \check{g}_{-k}^{(m)}\left\|\mathbf{G}_k\right\| \cdot \left\|T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}} - \mathbf{I}_r\right\| = O_P\left(\check{g}_{-k}^{(m)}d_k^{\alpha_{k,1}}\sqrt{\frac{r}{T}}\right),\end{aligned}$$

where the last equality used Theorem 2.8 of Wang and Paul (2014), which can be applied since we have $r = o(T^{1/3})$ and with fourth order moments exist for the elements in $\mathbf{Z}_f^{(k)}$ from Assumption (L1) and (R1) respectively, and that by Assumption (RE1),

$$\begin{aligned}\left\|\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\right\| &\leq \left\|\mathbf{A}_{f,T}\right\|^2 < \infty, \\ \frac{1}{(N+1)T}\mathrm{tr}(\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}) &= \frac{1}{(N+1)T}(\mathrm{tr}(\mathscr{A}_{f,T}^{\mathrm{T}}\mathscr{A}_{f,T}) - T^{-1}\mathbf{1}_T^{\mathrm{T}}\mathscr{A}_{f,T}^{\mathrm{T}}\mathscr{A}_{f,T}\mathbf{1}_T) \to \frac{1}{N+1}, \\ \frac{1}{(N+1)T}\mathrm{tr}(\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}})^2 &= \frac{1}{(N+1)T}\big\{\mathrm{tr}(\mathscr{A}_{f,T}^{\mathrm{T}}\mathscr{A}_{f,T})^2 - 2T^{-1}\mathbf{1}_T^{\mathrm{T}}(\mathscr{A}_{f,T}^{\mathrm{T}}\mathscr{A}_{f,T})^2\mathbf{1}_T \\ &\qquad + T^{-2}(\mathbf{1}_T^{\mathrm{T}}\mathscr{A}_{f,T}^{\mathrm{T}}\mathscr{A}_{f,T}\mathbf{1}_T)^2\big\} \to \frac{a_1 - 2a_2 + a_3^2}{N+1}.\end{aligned}$$

For $\left\|\mathbf{S}_{12,m}\right\|$, define the notation $\mathbf{U}_S$ to be a sub-matrix of $\mathbf{U}$ restricted to the rows indexed by S. Let $\mathbf{Z}_j := (\mathbf{Z}_\varepsilon^{(k)})_{B_j} \in \mathbb{R}^{d_k \times (M+1)T}$, where $B_j := \{(j-1)d_k + 1, \ldots, jd_k\}$, $j = 1, \ldots, d_{-k}$. Let also $\mathbf{a}_{\ell,j} := (\mathbf{\Sigma}_{\varepsilon,j}^{(k)})_\ell^{\mathrm{T}}$, $\ell \in [d_k]$. Using the notation in (3.74), for $h \in [r]$ and $\ell \in [d_k]$, the (h,ℓ)-th element of $T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\mathbf{Z}_\varepsilon^{(k)\mathrm{T}}\mathbf{Q}_{2,m}^{(k)\mathrm{T}}$ is $J_1 + J_2$, where

$$\begin{aligned}J_1 &:= T^{-1}\sum_{j=1}^{d_{-k}}(\mathbf{U}_{-k,(1)})_j\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}(\mathbf{Z}_f^{(k)})_h, \\ J_2 &:= T^{-1}\sum_{j=1}^{d_{-k}}(\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)})_j\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}(\mathbf{Z}_f^{(k)})_h.\end{aligned}$$

We have $EJ_1 = 0$, and by Assumption (E1),

$$E|J_1|^2 = T^{-2}\sum_{j=1}^{d_{-k}}(\mathbf{U}_{-k,(1)})_j^2\left\|\mathbf{a}_{\ell,j}\right\|^2\left\|\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\right\|_F^2 \leq O(T^{-2}) \cdot \left\|\mathscr{A}_{f,T}\right\|\left\|\mathbf{M}_T\right\| \cdot \left\|\mathscr{A}_{\varepsilon,T}\right\|_F^2 = O(T^{-1}),$$

where the last equality is from Assumption (RE1). It means that $J_1 = O_P(T^{-1/2})$. For J_2, using the Cauchy-Schwarz inequality and that $\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}(\mathbf{Z}_f^{(k)})_h = O_P(T^{1/2})$

from the analysis of J_1 above,

$$J_2 \le \left\|\breve{\mathbf{q}}_{k}^{(m)} - \mathbf{U}_{k,(1)}\right\| \cdot T^{-1}\left(\sum_{j=1}^{d_k}(\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}(\mathbf{Z}_f^{(k)})_h)^2\right)^{1/2}$$
$$= \left\|\breve{\mathbf{q}}_{k}^{(m)} - \mathbf{U}_{k,(1)}\right\| \cdot T^{-1}O_P(d_kT)^{1/2} = O_P\left(\left\|\breve{\mathbf{q}}_{k}^{(m)} - \mathbf{U}_{k,(1)}\right\|\sqrt{\frac{d_k}{T}}\right).$$

With the above, we then have

$$\left\|\mathbf{S}_{12,m}\right\| = \left\|\mathbf{Q}_{1,m}^{(k)}T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\mathbf{Z}_\varepsilon^{(k)\mathrm{T}}\mathbf{Q}_{2,m}^{(k)\mathrm{T}}\right\| \le \left\|\mathbf{Q}_{1,m}^{(k)}\right\| \cdot \sqrt{rd_k} \cdot O_P(J_1+J_2)$$
$$= O_P\left\{(\breve{g}_{k}^{(m)})^{1/2}d_k^{\alpha_{k,1}/2}\left(\sqrt{\frac{rd_k}{T}} + \left\|\breve{\mathbf{q}}_{k}^{(m)} - \mathbf{U}_{k,(1)}\right\|\sqrt{\frac{rd}{T}}\right)\right\}.$$

For $\mathbf{S}_{13,m}$, using the definitions in (3.74) and the bounds for $\left\|\mathbf{Q}_{3,m}^{(k)}\right\|$ in Lemma 3.6,

$$\left\|\mathbf{S}_{13,m}\right\| \le \left\|\mathbf{Q}_{1,m}^{(k)}\right\| \cdot \left\|\mathbf{Q}_{3,m}^{(k)}\right\| \cdot \left\|T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{e,T}^{\mathrm{T}}\mathbf{Z}_e^{(k)\mathrm{T}}\right\|$$
$$\le (\breve{g}_{k}^{(m)})^{1/2}d_k^{\alpha_{k,1}/2} \cdot 1 \cdot O_P(T^{-1}\sqrt{rr_e}\left\|\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{e,T}^{\mathrm{T}}\right\|_F) = O_P\left((\breve{g}_{k}^{(m)})^{1/2}d_k^{\alpha_{k,1}/2}\sqrt{\frac{rr_e}{T}}\right),$$

where the last line used the fact that an element in $\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{e,T}^{\mathrm{T}}\mathbf{Z}_e^{(k)\mathrm{T}}$ is $O_P(\left\|\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{e,T}^{\mathrm{T}}\right\|_F) = O_P(T^{1/2})$ by a similar calculation for treating J_1 above.

For $\mathbf{S}_{22,m}$, decompose $\mathbf{Q}_{2,m}^{(k)} = \mathbf{Q}_{2,m,0}^{(k)} + \mathbf{Q}_{2,m,e}^{(k)}$, where

$$\mathbf{Q}_{2,m,0}^{(k)} := (\mathbf{U}_{k,(1)}^{\mathrm{T}} \otimes I_{d_k})\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)},\ldots,\mathbf{\Sigma}_{\varepsilon,d_k}^{(k)}), \quad \mathbf{Q}_{2,m,e}^{(k)} := ((\breve{\mathbf{q}}_{k}^{(m)} - \mathbf{U}_{k,(1)})^{\mathrm{T}} \otimes I_{d_k})\mathrm{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)},\ldots,\mathbf{\Sigma}_{\varepsilon,d_k}^{(k)}).$$

Then $\mathbf{S}_{22,m} = I_0 + I_1 + I_2 + I_2^{\mathrm{T}} + I_3$, where

$$I_0 := \mathbf{Q}_{2,m}^{(k)}\mathbf{Q}_{2,m}^{(k)\mathrm{T}},$$
$$I_1 := \mathbf{Q}_{2,m,0}^{(k)}(T^{-1}\mathbf{Z}_\varepsilon^{(k)}\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\mathbf{Z}_\varepsilon^{(k)\mathrm{T}} - \mathbf{I}_d)\mathbf{Q}_{2,m,0}^{(k)\mathrm{T}},$$
$$I_2 := \mathbf{Q}_{2,m,e}^{(k)}(T^{-1}\mathbf{Z}_\varepsilon^{(k)}\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\mathbf{Z}_\varepsilon^{(k)\mathrm{T}} - \mathbf{I}_d)\mathbf{Q}_{2,m,0}^{(k)\mathrm{T}},$$
$$I_3 := \mathbf{Q}_{2,m,e}^{(k)}(T^{-1}\mathbf{Z}_\varepsilon^{(k)}\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\mathbf{Z}_\varepsilon^{(k)\mathrm{T}} - \mathbf{I}_d)\mathbf{Q}_{2,m,e}^{(k)\mathrm{T}}.$$

Firstly,

$$\left\|I_0\right\| = \left\|\mathbf{Q}_{2,m}^{(k)}\right\|^2 \le \max_{j\in[d_k]}\left\|\mathbf{\Sigma}_{\varepsilon,j}^{(k)}\right\| = O(1).$$

Using the same notations as in the treatment of $\mathbf{S}_{12,m}$ before within this proof, for $\ell, h \in [d_k]$, the (ℓ, h) entry of I_1 is given by

$$(I_1)_{\ell,h} = T^{-1}\Bigg(\sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j \mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\Bigg)\mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}\Bigg(\sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j \mathbf{a}_{h,j}^{\mathrm{T}}\mathbf{Z}_j\Bigg)^{\mathrm{T}} - \sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j^2 \mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}.$$

Hence by Assumption (RE1) and (E1), writing $\mathbf{G} := \mathscr{A}_{\varepsilon,T}\mathbf{M}_T\mathscr{A}_{\varepsilon,T}^{\mathrm{T}}$,

$$E(I_1)_{\ell,h} = \sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j^2 \mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\{T^{-1}\mathrm{tr}(\mathbf{G}) - 1\} = O(T^{-1/2}).$$

Also,

$$\begin{aligned}
E(I_1)_{\ell,h}^2 &= T^{-2}\sum_{i,j=1}^{d_k}(\mathbf{U}_{k,(1)})_i^2(\mathbf{U}_{k,(1)})_j^2 E\{(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^{\mathrm{T}}\mathbf{a}_{h,i})(\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathrm{T}}\mathbf{a}_{h,j})\} \\
&\quad + T^{-2}\sum_{i\neq j}(\mathbf{U}_{k,(1)})_i^2(\mathbf{U}_{k,(1)})_j^2\{E(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{Z}_i\mathbf{G}\mathbf{Z}_j^{\mathrm{T}}\mathbf{a}_{h,j})^2 + E\{\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{Z}_i\mathbf{G}\mathbf{Z}_j^{\mathrm{T}}\mathbf{a}_{h,j}\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{Z}_j\mathbf{G}\mathbf{Z}_i^{\mathrm{T}}\mathbf{a}_{h,i}\}\} \\
&\quad - 2T^{-1}\mathrm{tr}(\mathbf{G})\Bigg(\sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg)^2 + \Bigg(\sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg)^2 \\
&= T^{-2}\sum_{i=1}^{d_k}(\mathbf{U}_{k,(1)})_i^4 E(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^{\mathrm{T}}\mathbf{a}_{h,i})^2 + T^{-2}\sum_{i\neq j}(\mathbf{U}_{k,(1)})_i^2(\mathbf{U}_{k,(1)})_j^2(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})(\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j})\mathrm{tr}^2(\mathbf{G}) \\
&\quad + T^{-2}\sum_{i\neq j}(\mathbf{U}_{k,(1)})_i^2(\mathbf{U}_{k,(1)})_j^2(\|\mathbf{a}_{\ell,i}\|^2\|\mathbf{a}_{h,j}\|^2 + (\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})(\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}))\mathrm{tr}(\mathbf{G}^2) \\
&\quad - (2T^{-1}\mathrm{tr}(\mathbf{G}) - 1)\Bigg(\sum_{j=1}^{d_k}(\mathbf{U}_{k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg)^2. \qquad (3.76)
\end{aligned}$$

Now define $\mathbf{w}_\ell := \mathbf{Z}_i^{\mathrm{T}}\mathbf{a}_{\ell,i}$. Then $E(\mathbf{w}_\ell) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{w}_\ell, \mathbf{w}_h) = \mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i}\mathbf{I}_{(M+1)T}$, with elements in $\mathbf{w}_\ell$ independent of each other for each $\ell \in [d_k]$ and $i \in [d_{\text{-}k}]$. Hence

$$
\begin{aligned}
E(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^{\mathrm{T}}\mathbf{a}_{h,i})^2 = E(\mathbf{w}_\ell^{\mathrm{T}}\mathbf{G}\mathbf{w}_h) &= E\Bigg(\sum_{j=1}^{(N+1)T}(\mathbf{G})_{jj}(\mathbf{w}_\ell)_j(\mathbf{w}_h)_j + \sum_{j_1\neq j_2}(\mathbf{G})_{j_1j_2}(\mathbf{w}_\ell)_{j_1}(\mathbf{w}_h)_{j_2}\Bigg)^2 \\
&= \sum_{j=1}^{(N+1)T}(\mathbf{G})_{jj}^2 E[(\mathbf{w}_\ell)_j^2(\mathbf{w}_h)_j^2] + \sum_{j_1\neq j_2}(\mathbf{G})_{j_1j_1}(\mathbf{G})_{j_2j_2}(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2 \\
&+ \sum_{j_1\neq j_2}(\mathbf{G})_{j_1j_2}^2(\|\mathbf{a}_{\ell,i}\|^2\|\mathbf{a}_{h,i}\|^2 + (\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2) \\
&= \sum_{j=1}^{(N+1)T}(\mathbf{G})_{jj}^2\Big\{\|\mathbf{a}_{h,i}\|^2\|\mathbf{a}_{\ell,i}\|^2 + (\nu_4-3)\mathbf{a}_{h,i}^{\mathrm{T}}\mathrm{diag}(\mathbf{a}_{\ell,i}\mathbf{a}_{\ell,i}^{\mathrm{T}})\mathbf{a}_{h,i} + 2(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2\Big\} \\
&+ \sum_{j_1\neq j_2}(\mathbf{G})_{j_1j_1}(\mathbf{G})_{j_2j_2}(\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2 + \sum_{j_1\neq j_2}(\mathbf{G})_{j_1j_2}^2(\|\mathbf{a}_{\ell,i}\|^2\|\mathbf{a}_{h,i}\|^2 + (\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2) \\
&= \|\mathbf{a}_{h,i}\|^2\|\mathbf{a}_{\ell,i}\|^2\mathrm{tr}(\mathbf{G}^2) + (\mathbf{a}_{\ell,i}^{\mathrm{T}}\mathbf{a}_{h,i})^2(\mathrm{tr}^2(\mathbf{G}) + \mathrm{tr}(\mathbf{G}^2)) \\
&+ (\nu_4-3)\mathrm{tr}(\mathrm{diag}^2(\mathbf{G}))\mathbf{a}_{h,i}^{\mathrm{T}}\mathrm{diag}(\mathbf{a}_{\ell,i}\mathbf{a}_{\ell,i}^{\mathrm{T}})\mathbf{a}_{h,i}, \qquad (3.77)
\end{aligned}
$$

where $\nu_4 := E(\mathbf{Z}_i)_{11}^4 < \infty$ by Assumption (R1), and we used Lemma (A.2) of Li et al. (2019). Substitute this back to (3.76), we have $E(I_1)_{\ell,h}^2 = H_1 + H_2 + H_3 + H_4$, where

$$
\begin{aligned}
H_1 &:= (T^{-2}\mathrm{tr}^2(\mathbf{G}) - 2T^{-1}\mathrm{tr}(\mathbf{G}) + 1)\Bigg(\sum_{j=1}^{d_{\text{-}k}}(\mathbf{U}_{\text{-}k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg)^2, \\
H_2 &:= T^{-2}\mathrm{tr}(\mathbf{G}^2)\Bigg(\sum_{j=1}^{d_{\text{-}k}}(\mathbf{U}_{\text{-}k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg)^2, \\
H_3 &:= T^{-2}\mathrm{tr}(\mathbf{G}^2)\Bigg(\sum_{j=1}^{d_{\text{-}k}}(\mathbf{U}_{\text{-}k,(1)})_j^2\mathbf{a}_{\ell,j}^{\mathrm{T}}\mathbf{a}_{\ell,j}\Bigg)\Bigg(\sum_{j=1}^{d_{\text{-}k}}(\mathbf{U}_{\text{-}k,(1)})_j^2\mathbf{a}_{h,j}^{\mathrm{T}}\mathbf{a}_{h,j}\Bigg), \\
H_4 &:= T^{-2}\mathrm{tr}(\mathrm{diag}^2(\mathbf{G}))\sum_{j=1}^{d_{\text{-}k}}(\mathbf{U}_{\text{-}k,(1)})_j^4(\nu_4-3)\mathbf{a}_{h,j}^{\mathrm{T}}\mathrm{diag}(\mathbf{a}_{\ell,j}\mathbf{a}_{\ell,j}^{\mathrm{T}})\mathbf{a}_{h,j}.
\end{aligned}
$$

By Assumption (RE1), we have

$$
\begin{aligned}
|H_1|, |H_2|, |H_3| &= O(T^{-1})\cdot O(\max_{j\in[d_{\text{-}k}]}\|\mathbf{\Sigma}_{\varepsilon,j}^{(k)}\|_{\max}^2) = O(T^{-1}), \\
|H_4| &\le T^{-2}\mathrm{tr}(\mathbf{G}^2)(\nu_4-3)\max_{j\in[d_{\text{-}k}]}\|\mathbf{\Sigma}_{\varepsilon,j}^{(k)}\|_{\max}^2 = O(T^{-1}).
\end{aligned}
$$

Hence we can conclude that $\|I_1\| = O_P(d_kT^{-1/2})$.

For I_2, define $\widehat{e}_j := (\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)})_j$. Then for $\ell, h \in [d_k]$, the (ℓ,h) entry of I_2 is given by

$$(I_2)_{\ell,h} = \sum_{j_1=1}^{d_{-k}} \widehat{e}_{j_1} \left(T^{-1} \mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{Z}_{j_1} \mathbf{G} \sum_{j_2=1}^{d_{-k}} (\mathbf{U}_{-k,(1)})_{j_2} \mathbf{Z}_{j_2}^{\mathrm{T}} \mathbf{a}_{h,j_2} - (\mathbf{U}_{-k,(1)})_{j_1} \mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1} \right) =: \sum_{j_1=1}^{d_{-k}} \widehat{e}_{j_1} g_{j_1,\ell,h}.$$

By Assumption (RE1) and (E1),

$$E(g_{j_1,h,\ell}) = (T^{-1}\mathrm{tr}(\mathbf{G}) - 1)(\mathbf{U}_{-k,(1)})_{j_1} \mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1} = O(T^{-1/2}).$$

Also, similar to the treatment of I_1 and using (3.77),

$$\begin{aligned}
E(g_{j_1,h,\ell}^2) &= T^{-2} \sum_{j=1}^{d_{-k}} (\mathbf{U}_{-k,(1)})_j^2 E(\mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{Z}_{j_1} \mathbf{G} \mathbf{Z}_j^{\mathrm{T}} \mathbf{a}_{h,j})^2 \\
&\quad - 2T^{-1}\mathrm{tr}(\mathbf{G})(\mathbf{U}_{-k,(1)})_{j_1}^2 (\mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1})^2 + (\mathbf{U}_{-k,(1)})_{j_1}^2 (\mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1})^2 \\
&= \sum_{j=1}^{d_{-k}} (\mathbf{U}_{-k,(1)})_j^2 T^{-2} \mathrm{tr}(\mathbf{G}^2) \|\mathbf{a}_{h,j}\|^2 \|\mathbf{a}_{\ell,j_1}\|^2 \\
&\quad + (T^{-1}\mathrm{tr}(\mathbf{G}) - 1)^2 (\mathbf{U}_{-k,(1)})_{j_1}^2 (\mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1})^2 + T^{-2}\mathrm{tr}(\mathbf{G}^2)(\mathbf{U}_{-k,(1)})_{j_1}^2 (\mathbf{a}_{\ell,j_1}^{\mathrm{T}} \mathbf{a}_{h,j_1})^2 \\
&\quad + T^{-2} (\mathbf{U}_{-k,(1)})_{j_1}^2 (\nu_4 - 3) \mathrm{tr}(\mathrm{diag}^2(\mathbf{G})) \mathbf{a}_{h,j_1}^{\mathrm{T}} \mathrm{diag}(\mathbf{a}_{\ell,j_1} \mathbf{a}_{\ell,j_1}^{\mathrm{T}}) \mathbf{a}_{h,j_1} = O(T^{-1}),
\end{aligned}$$

so that we can conclude that $\|I_2\| = O_P\left(d_k \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{d_{-k}}{T}}\right) = O_P\left(d_k^{1/2} \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{d}{T}}\right)$.

For I_3, using Theorem 2 of Latała (2004),

$$\begin{aligned}
\|I_3\| &\le \|\mathbf{Q}_{2,m,e}^{(k)}\|^2 \|T^{-1} \mathbf{Z}_{\varepsilon}^{(k)} \mathbf{G} \mathbf{Z}_{\varepsilon}^{(k)\mathrm{T}} - \mathbf{I}_d\| \le \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\|^2 \max_{j \in [d_{-k}]} \|\mathbf{\Sigma}_{\varepsilon,j}^{(k)}\| \cdot O_P(1 + d/T) \\
&= O_P(\|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\|^2 (1 + d/T)).
\end{aligned}$$

Hence

$$\begin{aligned}
\|\mathbf{S}_{22,m}\| &= \|I_0\| + \|I_1\| + 2\|I_2\| + \|I_3\| \\
&= O_P\left(1 + \frac{d_k}{\sqrt{T}} + \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{d_k d}{T}} + \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\|^2 \left(\frac{d}{T} + 1\right)\right) \\
&= O_P\left(1 + \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\|^2 \frac{d}{T} + \frac{d_k}{\sqrt{T}} + \|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\| \sqrt{\frac{d_k d}{T}}\right).
\end{aligned}$$

For $\mathbf{S}_{23,m}$, the treatment is exactly parallel to the treatment of $\mathbf{S}_{12,m}$, so that

$$\begin{aligned}\left\|\mathbf{S}_{23,m}\right\| &= \left\|\mathbf{Q}_{3,m}^{(k)}\right\| \cdot \sqrt{r_e d_k} \cdot O_P(J_1 + J_2) \\ &= O_P\left\{\sqrt{\frac{r_e d_k}{T}} + \left\|\breve{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\right\| \sqrt{\frac{r_e d}{T}}\right\}.\end{aligned}$$

Finally, by Assumption (RE1) and Theorem 2 of Latała (2004) and that $r_e = o(T)$,

$$\left\|\mathbf{S}_{33,m}\right\| \leq \left\|\mathbf{Q}_{3,m}^{(k)}\right\|^2 O_P\left(1 + \sqrt{\frac{r_e}{T}}\right) = O_P(1).$$

This completes the proof of the lemma. □

Proof of Theorem 3.2. Firstly, (3.31) in Lemma 3.2 proves that $\mathbf{G}_k$ has the j-th diagonal element with order $d_k^{\alpha_{k,j}}$ for $j \in [r_k]$ under Assumption (L1), which is the same as those in $\mathbf{D}_k$. Define $\mathbf{B}_k \in \mathbb{R}^{d_k \times d_k - r_k}$ to be an orthogonal compliment of $\mathbf{U}_k = (\mathbf{U}_{k,(1)}, \mathbf{U}_{k,(2:r_k)})$, in the sense that $\mathbf{V}_k := (\mathbf{U}_k, \mathbf{B}_k)$ has $\mathbf{V}_k \mathbf{V}_k^{\mathrm{T}} = \mathbf{V}_k^{\mathrm{T}} \mathbf{V}_k = \mathbf{I}_{d_k}$. Then using the SVD of $\mathbf{A}_k$ in (3.5),

$$\mathbf{V}_k^{\mathrm{T}} \mathbf{S}'_{11,m} \mathbf{V}_k = \begin{pmatrix} \breve{g}_{-k}^{(m)} \mathbf{G}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

with the (j,j) element of $\breve{g}_{-k}^{(m)} \mathbf{G}_k$ of order $\breve{g}_{-k}^{(m)} d_k^{\alpha_{k,j}}$, $j \in [r_k]$. By Assumption (L1'), we then have

$$\mathrm{sep}(\mathbf{U}_{k,(1)}^{\mathrm{T}} \mathbf{S}'_{11,m} \mathbf{U}_{k,(1)}, (\mathbf{U}_{k,(2:r_k)}, \mathbf{B}_k)^{\mathrm{T}} \mathbf{S}'_{11,m} (\mathbf{U}_{k,(2:r_k)}, \mathbf{B}_k)) \asymp \breve{g}_{-k}^{(m)} d_k^{\alpha_{k,1}},$$

where

$$\mathrm{sep}(\mathbf{D}_1, \mathbf{D}_2) := \min_{\lambda \in \lambda(\mathbf{D}_1), \mu \in \lambda(\mathbf{D}_2)} |\lambda - \mu|.$$

Then by Lemma 3 of Lam et al. (2011), which is Theorem 8.1.10 in Golub and Van Loan (1996), since $\mathbf{U}_{k,(1)}$ is an eigenvector of $\mathbf{S}'_{11,m}$ and hence the span of $\mathbf{U}_{k,(1)}$ is an invariant subspace for $\mathbf{S}'_{11,m}$, for the matrix $\mathbf{E}_m$ in (3.75), if

$$\left\|\mathbf{E}_m\right\| = o_P(\breve{g}_{-k}^{(m)} d_k^{\alpha_{k,1}}),$$

then there exists $\breve{\mathbf{q}}_k^{(m+1)}$ which is an eigenvector of $\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}$ such that

$$\left\|\breve{\mathbf{q}}_k^{(m+1)} - \mathbf{U}_{k,(1)}\right\| = O(\left\|\mathbf{E}_m\right\| / (\breve{g}_{-k}^{(m)} d_k^{\alpha_{k,1}})). \tag{3.78}$$

We prove by induction on the integer m for the following statements:

I(m): $\check{g}_{-k}^{(m)} d_k^{\alpha_{k,1}} \asymp_P g_s$ for each $k \in [K]$.

II(m): For each $k \in [K]$,

$$\big\|\check{\mathbf{q}}_k^{(m+1)} - \mathbf{U}_{k,(1)}\big\| = O_P\bigg\{\sqrt{\frac{r}{T}} + g_s^{-1/2}\big\|\check{\mathbf{q}}_{-k}^{(m)} - \mathbf{U}_{-k,(1)}\big\|\sqrt{\frac{rd}{T}}\bigg\} = o_P(1).$$

The statement I(0) is proved exactly by Lemma 3.5. For statement II(0), using Lemma 3.6 and Lemma 3.7 at $m = 0$ and combine with the result that $\check{g}_{-k}^{(0)} = \widehat{g}_{-k} \asymp_P g_s/d_k^{\alpha_{k,1}}$ from I(0), we have from (3.75) that

$$\begin{aligned}
\big\|\mathbf{E}_0\big\| &= O_P\bigg(\big\|\mathbf{S}''_{11,0}\big\| + \big\|\mathbf{S}_{22,0}\big\| + \big\|\mathbf{S}_{33,0}\big\| + \sum_{i<j=1}^{3}(\big\|\mathbf{S}_{ij,0}\big\| + \big\|\mathbf{S}_{ij,0}^{\mathrm{T}}\big\|) + \sum_{i,j=1}^{3}\big\|\check{\mathbf{S}}_{ij,0}\big\|\bigg) \\
&= O_P\bigg(g_s\sqrt{\frac{r}{T}} + g_s^{1/2}\big\|\check{\mathbf{q}}_{-k}^{(0)} - \mathbf{U}_{-k,(1)}\big\|\sqrt{\frac{rd}{T}} + \big\|\check{\mathbf{q}}_{-k}^{(0)} - \mathbf{U}_{-k,(1)}\big\|^2\frac{d}{T}\bigg) \\
&= O_P\bigg(g_s\sqrt{\frac{r}{T}} + g_s^{1/2}b_k\sqrt{\frac{rd}{T}} + b_k^2\frac{d}{T}\bigg),
\end{aligned}$$

where the dominating rates in the second line are the results of the explicit rate assumptions presented at the beginning of Theorem 3.2, and the last line used (3.73) and the notation b_k used in Theorem 3.2. Hence from (3.78),

$$\begin{aligned}
\big\|\check{\mathbf{q}}_k^{(1)} - \mathbf{U}_{k,(1)}\big\| &= O_P(\big\|\mathbf{E}_0\big\|/(\check{g}_{-k}^{(0)}d_k^{\alpha_{k,1}})) = O_P(\big\|\mathbf{E}_0\big\|/g_s) \\
&= O_P\bigg(\sqrt{\frac{r}{T}} + g_s^{-1/2}\big\|\check{\mathbf{q}}_{-k}^{(0)} - \mathbf{U}_{-k,(1)}\big\|\sqrt{\frac{rd}{T}} + g_s^{-1}\big\|\check{\mathbf{q}}_{-k}^{(0)} - \mathbf{U}_{-k,(1)}\big\|^2\frac{d}{T}\bigg) \\
&= O_P\bigg(\sqrt{\frac{r}{T}} + g_s^{-1/2}b_k\sqrt{\frac{rd}{T}} + g_s^{-1}b_k^2\frac{d}{T}\bigg) = O_P\bigg(\sqrt{\frac{r}{T}} + g_s^{-1/2}b_k\sqrt{\frac{rd}{T}}\bigg) = o_P(1),
\end{aligned}$$

which is statement II(0), and the first statement of Theorem 3.2 is proved. Hence both I(m) and II(m) are true at $m = 0$.

Assume that both statements are true at a non-negative integer m. Then with II(m), like (3.72),

$$\begin{aligned}
\check{g}_{-k}^{(m+1)} &= \check{\mathbf{q}}_{-k}^{(m+1)\mathrm{T}}\mathbf{A}_{-k}\mathbf{A}_{-k}^{\mathrm{T}}\check{\mathbf{q}}_{-k}^{(m+1)} \\
&\geq \mathbf{U}_{-k,(1)}^{\mathrm{T}}\mathbf{U}_{-k}\mathbf{G}_{-k}\mathbf{U}_{-k}^{\mathrm{T}}\mathbf{U}_{-k,(1)} + 2(\check{\mathbf{q}}_{-k}^{(m+1)} - \mathbf{U}_{-k,(1)})^{\mathrm{T}}\mathbf{U}_{-k}\mathbf{G}_{-k}\mathbf{U}_{-k}^{\mathrm{T}}\mathbf{U}_{-k,(1)} \\
&= (1 + o_P(1))(\mathbf{G}_{-k})_{11} \asymp_P \prod_{j=1;j\neq k}^{K} d_j^{\alpha_{j,1}} = g_s d_k^{-\alpha_{k,1}},
\end{aligned}$$

which proves statement I($m+1$). With this, then using the definitions in (3.74), we have

$$\left\|\mathbf{Q}_{1,m+1}^{(k)}\right\| = (\check{\mathbf{q}}_{\text{-}k}^{(m+1)\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\check{\mathbf{q}}_{\text{-}k}^{(m+1)})\left\|\mathbf{A}_k\right\|^2 = \check{g}_{\text{-}k}^{(m+1)} \cdot d_k^{\alpha_{k,1}} = O_P(g_s), \quad \left\|\mathbf{Q}_{2,m+1}^{(k)}\right\| = O(1) = \left\|\mathbf{Q}_{3,m+1}^{(k)}\right\|.$$

Then the rest of the proofs of Lemma 3.6 follow through and we can conclude that

$$\sum_{i<j=1}^{3} \left\|\check{\mathbf{S}}_{ij,m+1}\right\| = o_P(T^{-1}).$$

With the above, and Lemma 3.7 at $m+1$, we have from (3.75) that

$$\begin{aligned}\left\|\mathbf{E}_{m+1}\right\| &= O_P\left(\left\|\mathbf{S}''_{11,m+1}\right\| + \left\|\mathbf{S}_{22,m+1}\right\| + \left\|\mathbf{S}_{33,m+1}\right\| + \sum_{i<j=1}^{3}(\left\|\mathbf{S}_{ij,m+1}\right\| + \left\|\mathbf{S}_{ij,m+1}^{\mathrm{T}}\right\|) + \sum_{i,j=1}^{3}\left\|\check{\mathbf{S}}_{ij,m+1}\right\|\right) \\ &= O_P\left(g_s\sqrt{\frac{r}{T}} + g_s^{1/2}\left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\|\sqrt{\frac{rd}{T}} + \left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\|^2\frac{d}{T}\right),\end{aligned}$$

so that by (3.78) and the proved I($m+1$),

$$\begin{aligned}\left\|\check{\mathbf{q}}_k^{(m+2)} - \mathbf{U}_{k,(1)}\right\| &= O_P(\left\|\mathbf{E}_{m+1}\right\|/(\check{g}_{\text{-}k}^{(m+1)}d_k^{\alpha_{k,1}})) = O_P(\left\|\mathbf{E}_{m+1}\right\|/g_s) \\ &= O_P\left(\sqrt{\frac{r}{T}} + g_s^{-1/2}\left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\|\sqrt{\frac{rd}{T}} + g_s^{-1}\left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\|^2\frac{d}{T}\right) \\ &= O_P\left(\sqrt{\frac{r}{T}} + g_s^{-1/2}\left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\|\sqrt{\frac{rd}{T}}\right),\end{aligned}$$

which is $o_P(1)$ since $\left\|\check{\mathbf{q}}_{\text{-}k}^{(m+1)} - \mathbf{U}_{\text{-}k,(1)}\right\| = o_P(1)$ by assumption II(m), and that $rdg_s^{-1} = o(T)$. This proves statement II($m+1$), and hence we have proved the statements I(m) and II(m) for any non-negative integers m by induction.

To prove the second part of Theorem 3.2, define $e_{m,k} := \left\|\check{\mathbf{q}}_k^{(m)} - \mathbf{U}_{k,(1)}\right\|$ and $e_{m,\text{-}k} := \left\|\check{\mathbf{q}}_{\text{-}k}^{(m)} - \mathbf{U}_{\text{-}k,(1)}\right\|$. We have from an argument similar to the one in (3.73) that

$$e_{m,\text{-}k} = O_P\left(\sum_{j=1;j\neq k}^{K} e_{m,j}\right). \tag{3.79}$$

We see from statement II(m) that the coefficient rate of $\left\|\check{\mathbf{q}}_{\text{-}k}^{(m)} - \mathbf{U}_{\text{-}k,(1)}\right\|$ is, for each $k \in [K]$, $u_k := g_s^{-1/2}\sqrt{rd/T}$, which is $o(1)$ by the the assumption $rdg_s^{-1} = o(T)$ in Theorem 3.2.

The statement II$(m+n-1)$ then implies that

$$e_{m+n,k} = O_P\left(\sqrt{\frac{r}{T}} + u_k e_{m+n-1,\text{-}k}\right) = O_P\left(\sqrt{\frac{r}{T}} + u_k \sum_{j=1;j\neq k}^{K} e_{m+n-1,j}\right).$$

Defining $\mathbf{e}_m := (e_{m,1},\ldots,e_{m,K})^{\mathrm{T}}$, the above becomes

$$\mathbf{e}_{m+n} = O_P\left(\sqrt{\frac{r}{T}}\mathbf{1}_K + \mathbf{W}\mathbf{e}_{m+n-1}\right), \quad \text{where } \mathbf{W} := \begin{pmatrix} 0 & u_1 & u_1 & \cdots & u_1 \\ u_2 & 0 & u_2 & \cdots & u_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_K & u_K & u_K & \cdots & 0 \end{pmatrix},$$

with $\|\mathbf{W}\|_\infty \le K \max_{k\in[K]} u_k = o(1)$ since K is assumed finite. Hence iterating the above $n-1$ more times, we have

$$\begin{aligned} \mathbf{e}_{m+n} &= O_P\left(\sqrt{\frac{r}{T}}(\mathbf{I}_K + \mathbf{W} + \cdots + \mathbf{W}^{n-1})\mathbf{1}_K + \mathbf{W}^n\mathbf{e}_m\right), \text{ implying} \\ \|\mathbf{e}_{m+n}\|_{\max} &= O_P\left(\sqrt{\frac{r}{T}}(1 + \|\mathbf{W}\|_\infty + \cdots + \|\mathbf{W}\|_\infty^{n-1}) + \|\mathbf{W}\|_\infty^n \|\mathbf{e}_m\|_{\max}\right) \\ &= O_P\left(\sqrt{\frac{r}{T}} + \|\mathbf{W}\|_\infty^n \|\mathbf{e}_m\|_{\max}\right) = O_P\left(\sqrt{\frac{r}{T}}\right) \end{aligned}$$

for n large enough. This completes the proof of the theorem. □

Proof of Theorem 3.3. Using the notations in (3.70) and in Lemma 3.7, since $\mathbf{S}_{11,m}$ is sandwiched by $\mathbf{A}_k$ and $\mathbf{A}_k^{\mathrm{T}}$, implying that it is sandwiched by $\mathbf{U}_k$ and $\mathbf{U}_k^{\mathrm{T}}$, the span of the columns of $\mathbf{U}_k$ forms an invariant subspace for $\mathbf{S}_{11,m}$. Now Let $\mathbf{B}_k$ be the orthogonal complement of $\mathbf{U}_k$, in the sense that $\mathbf{U} := (\mathbf{U}_k, \mathbf{B}_k)$ is an orthogonal matrix. Then we have

$$\mathbf{U}^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U} = \begin{pmatrix} \mathbf{U}_k^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \text{ where } \operatorname{sep}(\mathbf{U}_k^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U}_k, \mathbf{0}) = \lambda_{r_k}(\mathbf{U}_k^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U}_k), \text{ and}$$

$$\begin{aligned} \lambda_{r_k}(\mathbf{U}_k^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U}_k) &= \lambda_{r_k}\big(\mathbf{V}_k\mathbf{G}_k\mathbf{V}_k^{\mathrm{T}}(\mathbf{I}_{r_k} \otimes \breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}}(\mathbf{I}_{r_k} \otimes \breve{\mathbf{q}}_{\text{-}k}^{(m)\mathrm{T}}\mathbf{A}_{\text{-}k})^{\mathrm{T}}\big) \\ &\ge \lambda_{r_k}(\mathbf{G}_k)\lambda_{r_k}(\breve{g}_{\text{-}k}^{(m)}\mathbf{I}_{r_k})\lambda_r(T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}}) \asymp_P (\breve{g}_{\text{-}k}^{(m)} d_k^{\alpha_{k,r_k}}), \end{aligned}$$

where we used Theorem 2.8 of Wang and Paul (2014) (applicable since $r_k = o(T^{1/3})$ by Assumption (L1) and all variables are of bounded fourth moments by Assumption (R1)) to

conclude that

$$\lambda_r(T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}}) \asymp_P (N+1)\left(\frac{\mathrm{tr}(\mathscr{A}_{f,T}\mathbf{M}_t\mathscr{A}_{f,T}^{\mathrm{T}})}{(N+1)T}+\sqrt{\frac{r}{(N+1)T}}\right)=1+O(T^{-1/2}), \tag{3.80}$$

and the last equality used Assumption (RE1). Hence Lemma 3 of Lam et al. (2011) implies that there exists $\check{\mathbf{U}}_k$ with $\check{\mathbf{U}}_k^{\mathrm{T}}\check{\mathbf{U}}_k=\mathbf{I}_{r_k}$ and

$$\left\|\check{\mathbf{U}}_k-\mathbf{U}_k\right\|=O_P\big\{[\left\|\mathbf{S}_{12,m}\right\|+\left\|\mathbf{S}_{13,m}\right\|+\left\|\mathbf{S}_{22,m}\right\|+\left\|\mathbf{S}_{23,m}\right\|+\left\|\mathbf{S}_{33,m}\right\|+\sum_{i,j=1}^{3}\left\|\check{\mathbf{S}}_{ij,m}\right\|]/\mathrm{sep}(\mathbf{U}_k^{\mathrm{T}}\mathbf{S}_{11,m}\mathbf{U}_k,\mathbf{0})\big\},$$

such that $\check{\mathbf{U}}_k$ equals the r_k eigenvectors corresponding to the first r_k largest eigenvalues of $\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}$ multiplied with an orthogonal matrix, if we can show that the above rate is $o_P(1)$. To this end, statement I(m) in the proof of Theorem 3.2 shows that $\check{g}_{\text{-}k}^{(m)} \asymp_P g_s/d_k^{\alpha_{k,1}}$. And from the rate assumptions in Theorem 3.2 and the results of Lemma 3.7, we have

$$\begin{aligned}\left\|\check{\mathbf{U}}_k-\mathbf{U}_k\right\| &= O_P\left\{\left[1+\sqrt{\frac{rg_s}{T}}(r_e^{1/2}+d_k^{1/2}+\left\|\check{\mathbf{q}}_{\text{-}k}^{(m)}-\mathbf{U}_{\text{-}k,(1)}\right\|d^{1/2})\right]/(g_s d_k^{\alpha_{k,r_k}-\alpha_{k,1}})\right\}\\ &= O_P\left\{d_k^{\alpha_{k,1}-\alpha_{k,r_k}}\left[g_s^{-1}+\sqrt{\frac{r}{Tg_s}}(r_e^{1/2}+d_k^{1/2}+\left\|\check{\mathbf{q}}_{\text{-}k}^{(m)}-\mathbf{U}_{\text{-}k,(1)}\right\|d^{1/2})\right]\right\}\end{aligned} \tag{3.81}$$

If m is large enough, then following Theorem 3.2, we have by (3.79) that

$$\left\|\check{\mathbf{q}}_{\text{-}k}^{(m)}-\mathbf{U}_{\text{-}k,(1)}\right\|=O_P\left(\sum_{j=1;j\neq k}^{K}\left\|\check{\mathbf{q}}_j^{(m)}-\mathbf{U}_{j,(1)}\right\|\right)=O_P\left(K\sqrt{\frac{r}{T}}\right).$$

Substituting the above rate into (3.81) completes the proof. □

Chapter 4

Rank Estimation in Time Series Tensor Factor Models by Bootstrapped Correlation Thresholding

4.1 Introduction

High dimensional time series data, often observed in tensor format, is becoming increasingly prevalent across various fields. An effective and widely adopted approach approach for dimension reduction of such high dimensional tensor time series is to employ a factor model structure (Chen et al., 2022; Han et al., 2020; Wang et al., 2019), similar to the Tucker decomposition for tensors (refer to Chapter 2 for an introduction to tensor factor modelling). In factor modelling, an important step is to accurately determine the number of relevant factors in the model. Underestimating the number of relevant factors can result in a loss of signal, while overestimating it may introduce more noise, both of which can negatively impact estimation accuracy and forecasting performance. By specifying the correct number of relevant factors, we can extract the significant signals from the factors to the greatest extent possible, while minimizing the inclusion of excessive noise. Thus, the accuracy of estimation and forecasting procedures depends on the specification of the number of factors (Ahn and Horenstein, 2013; Bai, 2003; Lam and Yao, 2012), making it an important consideration in the modelling process.

Over the past few decades, numerous methods have emerged for determining the number of common factors required for modelling high dimensional vector time series (i.e., when the tensor order is $K = 1$). The predominant approach involves leveraging

the behaviour of the eigenvalues of the covariance matrix under the approximate factor model assumptions with weak serial dependence of the idiosyncratic components (Bai and Ng, 2002), or the singular values of the autocovariance matrix under the assumption of 'white noise' (Lam and Yao, 2012). According to the assumption of factor models, the eigenvalues (or the singular values) corresponding to the r common components increase with the cross-sectional units d and diverge to infinity, while the remaining eigenvalues, representing idiosyncratic components, remain bounded. Under the approximate factor model assumptions with weak serial and cross-sectional dependence of noise series, Bai and Ng (2002) propose two information criteria to determine the number of factors by separating diverging eigenvalues from the rest using threshold functions. Additionally, Ahn and Horenstein (2013) estimate the number of factors by maximising the ratio of consecutive eigenvalues of the sample covariance matrix. In the realm of assumptions with independent noise series (Lam et al., 2011), Lam and Yao (2012) also employ an eigenvalues ratio-based estimator but using the sample autocovariance matrix instead. Other efforts to estimate the number of factors for vector factor modelling include contributions from Amengual and Watson (2007); Bai and Ng (2007); Hallin and Liška (2007); Kapetanios (2010); Kong (2017); Li et al. (2017); Luo and Li (2016); Onatski (2010, 2012); Ye and Weiss (2003).

All of the above studies focus on determining the number of factors for vector time series (i.e., order-1 tensor time series). The extension of factor modelling to matrix and tensor time series has attracted significant interest in recent years. For tensor factor model (2.7) with $K > 1$, the number of factors in each mode r_k, $k \in [K]$ actually defines the rank of the core tensor $\mathscr{C}_t$. Under tensor factor model assumptions with independent $\{\mathscr{E}_t\}$, Han et al. (2022) extend information criteria (IC) and eigenvalue ratio (ER) based methods to consistently estimate the rank of the core tensor. However, penalty functions are needed to be specified for both IC and ER methods, and potential tuning parameters are needed for fine tuning of performance, which can be computationally expensive.

Under the assumption of weak serial dependence of $\{\mathscr{E}_t\}$, recent studies have extended eigenvalue ratio (ER) based methods (Ahn and Horenstein, 2013; Lam and Yao, 2012) to estimate the number of factors. These methods involve maximising the ratio of consecutive eigenvalues on different variants of "sample covariance matrices," which are intricately linked with the corresponding factor loading estimation procedures. For matrix factor models (i.e., an order-2 tensor), Chen and Fan (2021) introduce an α-PCA method that performs eigenanalysis on aggregated statistics in both first and second moments. He et al. (2022) conduct eigenanalysis on the sample matrix Kendall's tau. Yu et al. (2022)

define the "sample covariance matrix" by projecting the observation matrix onto the row or column factor space, while He et al. (2023a) proposes a similar robust method using a weighted version of the projection matrix. Barigozzi et al. (2023b) and Barigozzi et al. (2023a) further generalize the projection method to estimate the core tensor rank for tensor factor models for a general K, and prove consistency of the corresponding estimators. However, the literature mentioned above all assume the existence of solely pervasive factors in the model, a limitation that can be restrictive.

All the previously discussed studies rely on the examination of the eigenvalues of the sample covariance matrix or its variations. However, a challenge arises in comparing these eigenvalues due to the heterogeneous scales of observed variables, making it challenging to establish a precise relationship between these eigenvalues and the number of common factors. In practice, it is common to re-scale all variables to have a variance of one before conducting any principal component analysis. As a result, many of the above-mentioned methods are often applied through the eigenvalue ratio (ER) test on the sample correlation matrices. However, in the presence of weak factors, the eigenvalues of the sample covariance matrix or correlation matrix often diverge at varying rates, resulting in the eigenvalue ratio (ER)-based methods being potentially less effective. In the context of the vector factor model, Lam and Yao (2012) propose a two-step estimation procedure to sequentially estimate the number of strong factors and weak factors using the eigenvalue ratio-based estimator. However, this procedure assumes that the factors have only two layers of strengths, i.e. $r_{(1)}$ strong factors with strength $\alpha_{(1)} = 1$, and $r_{(2)}$ weak factors with the same strength $\alpha_{(2)} < 1$. Consequently, it becomes challenging to generalize this approach when the actual model may consist of a spectrum of different factor strengths, as specifying the number of layers of strengths can be difficult in practice.

In analyzing the correlation matrix, in contrast to the eigenvalue ratio (ER)-based methods, Fan et al. (2022) propose an alternative approach by introducing an adjusted correlation thresholding method to determine the number of factors for a vector factor model. They demonstrate that, under certain mild conditions, the number of eigenvalues greater than 1 of the population correlation matrix is the same as the number of common factors, and establish an optimal threshold to account for sampling variabilities and biases. Building on this idea, Lam (2021) extend this method to estimate the core tensor rank for a general order-K tensor factor model by thresholding the eigenvalues of the total model-k correlation matrix.

In this chapter, we further extend the correlation thresholding method proposed by Fan et al. (2022) and Lam (2021) to our tensor factor model, as defined in Chapter 3.

This model not only assumes the presence of both serial and cross-correlations in the idiosyncratic components but also allows for a mixed strength of factors. While Lam (2021) consider using the total model-k correlation matrix directly, we introduce our core tensor rank estimators through correlation analysis on the projected data defined in Chapter 3.3, and provide theoretical guarantees for our estimators. Additionally, we present a bootstrap method for tuning parameter selection for practical implementation, and thus, our method is abbreviated as BCorTh (Bootstrapped Correlation Thresholding). Simulation studies and real data analyses are conducted to compare BCorTh with other state-of-the-art methods. Empirical studies demonstrate that BCorTh can effectively identify weak factors when they are present.

The rest of this chapter is organized as follows. Section 4.2 provides theoretical justifications for using correlation analysis in finding the rank of the core tensor. It also introduces a fibre bootstrapping technique in determining the tuning parameter of the procedure. Section 4.3 presents our simulation studies on various settings, comparing BCorTh to other state-of-the-art estimators. Section 4.4 analyses a set of matrix-valued portfolio return data and a tensor-valued NYC taxi data set, demonstrating the performance of all the rank and factor loading estimators proposed in Chapter 3 and Chapter 4. All the proofs are presented in Section 4.6.

Finally, all our methods in Chapter 3 and Chapter 4 are written into an R package `TensorPreAve` published on CRAN and GitHub. Please see Section 4.5 for a very brief explanation on how to use it.

4.2 Core Tensor Rank Estimation Using Projected Data

In Chapter 3, the tensor factor model for each $\mathcal{X}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t \in [T]$, is defined as

$$\mathcal{X}_t = \mu + \mathcal{C}_t + \mathcal{E}_t = \mu + \mathcal{F}_t \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K + \mathcal{E}_t. \tag{4.1}$$

Here, $\mathcal{F}_t \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ represents the core tensor. In this chapter, our objective is to provide an estimator for the rank of the core tensor, denoted as r_k for $k \in [K]$.

Our model assumes the presence of both weak serial and cross-correlations in the idiosyncratic components $\mathcal{E}_t$. Additionally, we allow for a spectrum of different factor strengths in each $\mathbf{A}_k$. To avoid redundancy, please refer to Chapter 3.2.1 for a detailed explanation of the assumptions underlying the model (4.1). We assume that all these assumptions hold throughout the remainder of this chapter.

In Chapter 3.3, we present an *Algorithm for Iterative Projection Direction Refinement* and conduct eigenanalysis on the associated covariance matrix corresponding to the final projection step to estimate the factor loading spaces. Interestingly, the same projected data can also be utilised to estimate the rank of the core tensor using correlation analysis. We briefly recap the algorithm as follows: The superscript (i) in $\check{\mathbf{q}}_k^{(i)}$ signals that this is the i-th iterated estimator for $\mathbf{U}_{k,(1)}$ (i.e., the direction corresponding to the strongest factor), and $\widehat{\mathbf{q}}_{k,pre}$ can be obtained using the pre-averaging procedure, as described in Chapter 3.2.

Algorithm for Iterative Projection Direction Refinement

1. Initialize $\check{\mathbf{q}}_k^{(0)} = \widehat{\mathbf{q}}_{k,pre}$ for each $k \in [K]$.
2. For $i \geq 1$, at the i-th step, create projected data $\mathbf{y}_{t,i}^{(k)} := \text{mat}_k(\mathscr{X}_t - \bar{\mathscr{X}})\check{\mathbf{q}}_k^{(i-1)}$ for each $k \in [K]$.
3. For each $k \in [K]$, define $\check{\mathbf{q}}_k^{(i)}$ the eigenvector corresponding to the largest eigenvalue of
$$\widetilde{\mathbf{\Sigma}}_{y,i}^{(k)} := T^{-1} \sum_{t=1}^{T} \mathbf{y}_{t,i}^{(k)} \mathbf{y}_{t,i}^{(k)\mathrm{T}}. \tag{4.2}$$
4. Replace i by $i+1$. Go back to step 2. Stop until after the procedure has been repeated for a fixed number of times.

With the projected data and the associated covariance matrix $\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}$ defined in (4.2), define the associated correlation matrix $\widetilde{\mathbf{R}}_{y,m+1}^{(k)}$ as

$$\widetilde{\mathbf{R}}_{y,m+1}^{(k)} := \text{diag}^{-1/2}(\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)})\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}\text{diag}^{-1/2}(\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}),\ k \in [K]. \tag{4.3}$$

Our estimator for r_k for each $k \in [K]$ is then defined to be

$$\widehat{r}_k := \max\{j : \lambda_j(\widetilde{\mathbf{R}}_{y,m+1}^{(k)}) > 1 + \eta_T,\ j \in [d_k]\}, \tag{4.4}$$

where $\eta_T \to 0$ as $T \to \infty$, and its practical choice will be discussed in Section 4.2.2. This estimator is inspired by the one in Fan et al. (2022) for independent observations from a vector factor model.

4.2.1 Main results

The following assumption is needed for all the theorems in this section.

(RE2) (Model Parameters) *For each* $k \in [K]$, *we assume that for each* $j \in [d_k]$, $c_1 \leq \lambda_j(\mathrm{diag}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})) \leq c_2$ *for some* $0 < c_1, c_2 < \infty$ *as* $T, d_k \to \infty$. *Moreover,* $r_k = o(d_k^{1-\alpha_{k,1}+\alpha_{k,r_k}})$.

Assumption (RE2) ensures that each row of $\mathbf{A}_k$ has at least one non-zero value, meaning that at least one factor drives the dynamics of the corresponding element in $\mathbf{y}_{t,m+1}^{(k)}$. The assumption can be weakened so that the values are vanishing, at the price of more complicated proofs and rates in Theorem 4.2. Define

$$\mathbf{\Sigma}_{y,m+1}^{(k)} := \breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k}\mathbf{A}_{-k}^{\mathrm{T}}\breve{\mathbf{q}}_{-k}^{(m)}\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}} + \sum_{j=1}^{d_{-k}}(\breve{\mathbf{q}}_{-k}^{(m)})_j^2\mathbf{\Sigma}_{\varepsilon,j}^{(k)} + \breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{e,-k}\mathbf{A}_{e,-k}^{\mathrm{T}}\breve{\mathbf{q}}_{-k}^{(m)}\mathbf{A}_{e,k}\mathbf{A}_{e,k}^{\mathrm{T}}. \quad (4.5)$$

The matrix $\mathbf{\Sigma}_{y,m+1}^{(k)}$ is in fact the expected value of $\widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)}$ in (4.2), pretending that $\breve{\mathbf{q}}_{-k}^{(m)}$ is a constant vector.

Theorem 4.1. *Let Assumption (E1), (F1) and (RE2) hold. Define the correlation matrix*

$$\mathbf{R}_{y,m+1}^{(k)} = \mathrm{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)})\mathbf{\Sigma}_{y,m+1}^{(k)}\mathrm{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)}), \quad k \in [K].$$

Then for large enough T, d_k, *we have in probability* $\lambda_j(\mathbf{R}_{y,m+1}^{(k)}) \succeq_P r_k^{-1}d_k^{1-\alpha_{k,1}+\alpha_{k,j}} > 1$ *for* $j \in [r_k]$, *whereas* $\lambda_j(\mathbf{R}_{y,m+1}^{(k)}) \leq 1$ *for* $j = r_k+1, \ldots, d_k$.

This theorem is in parallel to Theorem 1 of Fan et al. (2022). With this, we can write

$$r_k = \max\{j : \lambda_j(\mathbf{R}_{y,m+1}^{(k)}) > 1,\ j \in [d_k]\}.$$

In light of this, the estimator $\widehat{r}_k$ in (4.4) makes sense. The following theorem shows further that $\widehat{r}_k$ is in fact consistent for r_k for a suitable choice of η_T.

Theorem 4.2. *Let (RE2) and all the assumptions in Theorem 3.2 hold. Suppose*

$$d_k^{\alpha_{k,1}-\alpha_{k,r_k}}\left(\sqrt{\frac{r(r_e+d_k)}{Tg_s}} + \frac{Kr}{T}\sqrt{\frac{d}{g_s}}\right) = o(1), \quad k \in [K],$$

where g_s *is defined in Theorem 3.2. Then as* $T, d_k \to \infty$, *we have for each* $k \in [K]$,

$$\lambda_j(\widetilde{\mathbf{R}}_{y,m+1}^{(k)}) = \begin{cases} \succeq_P r_k^{-1}d_k^{1-\alpha_{k,1}+\alpha_{k,j}} & \\ \quad \cdot(1 + O_P\{r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}a_T(0) + a_T(\alpha_{k,1})\}), & j \in [r_k]; \\ \leq 1 + O_P\{b_T\}, & j \in [d_k]/[r_k], \end{cases}$$

where for $0 < \delta \leq 1$,

$$a_T(\delta) := \sqrt{\frac{r}{T}}\left[1 + d_k^{\delta/2} g_s^{-1/2}\left(r_e^{1/2} + d_k^{1/2} + K\sqrt{\frac{rd}{T}}\right) + d_k^{\delta} g_s^{-1}\frac{K^2 r^{1/2} d}{T^{3/2}}\right],$$
$$b_T := d_k^{\alpha_{k,1}} g_s^{-1}\left\{\sqrt{\frac{(r_e + d_k)d_k}{T}} + \frac{K\sqrt{r(r_e + d_k)d}}{T} + \frac{K^2 rd}{T^2}\right\},$$

with $r_k d_k^{2\alpha_{k,1} - \alpha_{k,r_k} - 1} a_T(0)$, $a_T(\alpha_{k,1})$ *and* b_T *assumed* $o(1)$. *Hence* $\widehat{r}_k$ *in (4.4) is a consistent estimator for* r_k *if we choose* $\eta_T = Cb_T$ *for some constant* $C > 0$.

To gain some insights from the theorem, we can compare the rates of $\lambda_j(\widetilde{\mathbf{R}}_{y,m+1}^{(k)})$ from Theorem 4.2 and $\lambda_j(\mathbf{R}_{y,m+1}^{(k)})$ from Theorem 4.1. In this comparison, b_T serves as the 'bias correction term' attributed to noise, indicating the extent of eigenvalue disturbance caused by noise series. Therefore, the rate of b_T helps determine the threshold η_T that we choose. Suppose the strongest factor for each mode-k unfolded matrix is pervasive, i.e., $\alpha_{j,1} = 1$ for each $j \in [K]$, and r_k and K are constants with $d_1 \asymp \cdots \asymp d_K \asymp T$. Then

$$r_k d_k^{2\alpha_{k,1} - \alpha_{k,j} - 1} a_T(0) + a_T(1) \asymp T^{-1/2}, \quad b_T = O(d_k^{1/2} d_{-k}^{-1} + d_{-k}^{-1/2} + T^{-1}).$$

This shows that the rate of convergence of b_T is at best $T^{-1/2}$ when $K = 2$, and T^{-1} when $K \geq 3$. It means that our search for η_T can be in the form $CT^{-1/2}$ when $K = 2$, and CT^{-1} when $K \geq 3$. The extra rate assumptions in the theorem may not be more stringent than those in Theorem 3.2 and (RE2). For instance, if K and each r_k for $k \in [K]$ are constants with $d_1 \asymp \cdots \asymp d_K \asymp T$ and all factors are pervasive, then the extra rate assumptions in Theorem 4.2 are satisfied automatically.

4.2.2 Practical implementation for core rank estimator

Since there is only one mode-k unfolding matrix from our data, we propose the following algorithm for Bootstrapping the mode-k fibres to facilitate the search for η_T.

Bootstrapping Algorithm for mode-k tensor fibres and projected data

1. Initialize an integer $B > 0$, and independent sequences of i.i.d. Bernoulli random variables $\{\xi_j^{(b)}\}_{j \in [d_{-k}]}$ for each $b \in [B]$.
2. For each b, create $\mathbf{W}_b \in \mathbb{R}^{d_{-k} \times d_{-k}}$, where the i-th column is $\mathbf{0}$ except its j-th zero is replaced by $\xi_i^{(b)}$, with j chosen uniformly from $[d_{-k}]$.

3. Define new projected data $\mathbf{y}_{t,m+1,b}^{(k)} := \mathrm{mat}_k(\mathscr{X}_t - \bar{\mathscr{X}})\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\check{\mathbf{q}}_{\text{-}k}^{(m)}$ for each $b \in [B]$.

Essentially, we Bootstrap the mode-k fibres by choosing them randomly with replacement, and augment the vector of projection $\check{\mathbf{q}}_{\text{-}k}^{(m)}$ accordingly by pre-multiplying it with $\mathbf{W}_b^{\mathrm{T}}$. Note that though the sample size $d_{\text{-}k}$ could be diverging, we control each row of $\mathbf{W}_b$ to contains at most c $\xi_i^{(b)}$'s with a finite c. We take $c = 8$ here as an example in practice, meaning that a fibre is at most chosen 8 times in each Bootstrap sample. This facilitates our proof of Theorem 4.3 to bound the norm of $\mathbf{W}_b$, although for all our simulations, a fibre is never chosen more than 8 times.

From here on, we drop the subscript $m+1$ for the ease of presentation. With the new projected data, we then create new covariance and correlation matrices:

$$\widetilde{\mathbf{\Sigma}}_{y,b}^{(k)} := T^{-1}\sum_{t=1}^{T}\mathbf{y}_{t,b}^{(k)}\mathbf{y}_{t,b}^{(k)\mathrm{T}}, \quad \widetilde{\mathbf{R}}_{y,b}^{(k)} := \mathrm{diag}^{-1/2}(\widetilde{\mathbf{\Sigma}}_{y,b}^{(k)})\widetilde{\mathbf{\Sigma}}_{y,b}^{(k)}\mathrm{diag}^{-1/2}(\widetilde{\mathbf{\Sigma}}_{y,b}^{(k)}), \ \ k \in [K], b \in [B].$$

Theorem 4.3. *Let all the assumptions in Theorem 4.2 hold. Recall $\mathbf{U}_k$ as defined in (3.5) to be the left singular vectors of $\mathbf{A}_k$, and $\mathbf{U}_{-k} := \mathbf{U}_K \otimes \cdots \mathbf{U}_{k+1} \otimes \mathbf{U}_{k-1} \otimes \cdots \otimes \mathbf{U}_1$. Suppose for each $k \in [K]$, the elements in the unit vector $\mathbf{U}_{\text{-}k,(1)} =: (u_j)_{j\in[d_{\text{-}k}]}$ have the same moment structure up to the 4th order, and $E(u_i^2u_j^2) = d_{\text{-}k}^{-2}(1+o(1))$ for $i \neq j$ as $d_{\text{-}k} \to \infty$. Then Theorem 4.1 holds for $\mathbf{R}_{y,m+1}^{(k)}$ defined there but with $\check{\mathbf{q}}_{\text{-}k}^{(m)}$ in $\mathbf{\Sigma}_{y,m+1}^{(k)}$ replaced by $\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\check{\mathbf{q}}_{\text{-}k}^{(m)}$. Theorem 4.2 holds also for $\widetilde{\mathbf{R}}_{y,b}^{(k)}$.*

The above theorem means that any procedures for finding the number of factors exploiting Theorem 4.1 and 4.2, should work for our Bootstrapped correlation matrix $\widetilde{\mathbf{R}}_{y,b}^{(k)}$ too. The assumption on $E(u_i^2u_j^2)$ is mild, since it is easily see that $E(u_i^2) = d_{\text{-}k}^{-1}$, so that at exact independence we have $E(u_i^2u_j^2) = d_{\text{-}k}^{-2}$. We are essentially assuming that the covariance among the u_i's are $o(d_{\text{-}k}^{-2})$, so that u_i and u_j are nearly uncorrelated.

For a constant C, we use $\eta_T = CT^{-1/2}$ for $K = 2$ and $\eta_T = CT^{-1}$ for $K \geq 3$, and calculate

$$\widehat{r}_k^{(b)}(C) := \max\{j : \lambda_j(\widetilde{\mathbf{R}}_{y,b}^{(k)}) > 1 + \eta_T, \ j \in [d_k]\}.$$

We propose to choose C with

$$\widehat{C} := \min_{C>0}\widehat{\mathrm{var}}(\{\widehat{r}_k^{(b)}(C)\}_{b\in[B]}),$$

where $\widehat{\text{var}}(\{x_t\}_{t\in\mathscr{T}})$ is the sample variance of $\{x_t\}_{t\in\mathscr{T}}$. Finally, our estimator for r_k is defined to be

$$\check{r}_k := \text{Mode of } \{\widehat{r}_k^{(b)}(\widehat{C})\}_{b\in[B]}. \tag{4.6}$$

The intuition of $\widehat{C}$ and $\check{r}_k$ is as follows. If there are r_k factors for $\mathbf{A}_k$, then the first r_k eigenvalues of $\widetilde{\mathbf{R}}_{y,b}^{(k)}$ for each $b \in [B]$ should be approximately well-separated. Setting a large C will create a large threshold $1+\eta_T$ that is almost always lying in between $\lambda_j(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ and $\lambda_{j+1}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ for some fixed $j \in [r_k]$ for each $b \in [B]$, so that $\widehat{\text{var}}(\{\widehat{r}_k^{(b)}(C)\}_{b\in[B]})$ will be small, or even equals 0.

However, if C is small such that $1+\eta_T$ is now in between $\lambda_j(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ and $\lambda_{j+1}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ for some $j \in [d_k]/[r_k]$ and some $b \in [B]$, then we expect that this particular threshold will lie in between $\lambda_{j'}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ and $\lambda_{j'+1}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ for some $j' \neq j$ and some others $b \in [B]$, since all these eigenvalues are less than or equal to 1 by Theorem 4.1, and their variability is originated from the noise series only, making them less stable compared to when $j \in [r_k]$. Hence for a small enough C, we expect $\widehat{\text{var}}(\{\widehat{r}_k^{(b)}(C)\}_{b\in[B]})$ to be large. The range of values of C such that $1+\eta_T$ lies in between $\lambda_{r_k}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ and $\lambda_{r_k+1}(\widetilde{\mathbf{R}}_{y,b}^{(k)})$ for the majority of $b \in [B]$ will then include $\widehat{C}$. The definition of $\check{r}_k$ in (4.6) allows for variability arises from the noises and the r_k-th factor which can be weak and hence may not be detected in all Bootstrap samples.

While we haven't developed an explicit theoretical framework for our final estimator $\check{r}_k$, our empirical findings in Section 4.3 illustrate that, with a suitable selection of η_T as recommended by Theorem 4.2, the bootstrapped procedure provides a reliable estimator of the rank of the core tensor across various scenarios. Finally, in all the simulation settings in Section 4.3, we use $B = 50$ Bootstrap samples. This is a safe number, since reducing it to 10 in fact hardly change the results in our simulation experiments.

4.3 Simulation Experiments

In this section, we conduct simulation experiments to test the performances of our proposed rank estimators (BCorTH) with bootstrapping of tensor fibres for tuning parameter selection, and compare it with other state-of-the-art methods.

4.3.1 Simulation settings

For generating our data, we use model (4.1), with elements in μ being i.i.d. standard normal in each repetition of experiment. For $k \in [K]$, each factor loading matrix $\mathbf{A}_k$ is generated

independently with $\mathbf{A}_k = \mathbf{B}_k\mathbf{R}_k$, where the elements in $\mathbf{B}_k \in \mathbb{R}^{d_k \times r_k}$ are i.i.d. $U(u_1,u_2)$, and $\mathbf{R}_k \in \mathbb{R}^{r_k \times r_k}$ is diagonal with the jth diagonal element being $d_k^{-\zeta_{k,j}}$, $0 \leq \zeta_{k,j} \leq 0.5$. Pervasive (strong) factors have $\zeta_{k,j} = 0$, while weak factors have $0 < \zeta_{k,j} \leq 0.5$.

The elements in $\mathscr{F}_t$ are independent standardized AR(5) with AR coefficients 0.7,0.3,-0.4,0.2 and -0.1. Same for the elements in $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$ in (3.2), but their AR coefficients are (-0.7,-0.3,-0.4,0.2,0.1) and (0.8,0.4, -0.4,0.2,-0.1) respectively. The standard deviation of each element of $\boldsymbol{\varepsilon}_t$ is randomly generated with i.i.d. $|\mathcal{N}(0,1)|$. Each entry of the matrices $\mathbf{A}_{e,k} \in \mathbb{R}^{d_k \times r_{e,k}}, k \in [K]$ is generated with i.i.d. standard normal, but has an independent probability of 0.7 being set exactly to 0. Each experiment is repeated 500 times. Similar to Section 3.4.1, we consider the simulation settings (I), (II), (III) and (IV), with sub-settings (a) and (b), detailed below:

(Ia) Two strong factors with $r_k = 2$, $\zeta_{k,j} = 0$ for all k, j, and $u_1 = -2$, $u_2 = 2$ (elements in $\mathbf{A}_k$ have mean 0)).

(IIa) One strong factor and one weak factor with $r_k = 2$, $\zeta_{k,1} = 0$ and $\zeta_{k,2} = 0.2$ for all k; $u_1 = -2$, $u_2 = 2$.

(IIIa) Two weak factors with $r_k = 2$, $\zeta_{k,1} = 0.1$ and $\zeta_{k,2} = 0.2$ for all k; $u_1 = -2$, $u_2 = 2$.

(IVa) Four strong factors with $r_k = 4$, $\zeta_{k,j} = 0$ for all k, j; $u_1 = -2$, $u_2 = 2$.

Setting (Ib) to (IVb) are the same as (Ia) to (IVa) respectively, except that $u_1 = 0$, $u_2 = 2$, so that the elements in $\mathbf{A}_k$ have non-zero mean, leading to larger signal accumulation for the initial pre-averaging procedure as introduced in Section 3.2.

Setting (I)(II)(III) and (IV) are designed to test the performance of rank estimators under different profiles of factor strengths. In Setting (I), we have two strong factors with $\alpha_{k,1} = \alpha_{k,2} = 1$ for each mode k, which is consistent with the pervasive factor assumptions of Barigozzi et al. (2023b); Chen and Fan (2021); He et al. (2023a, 2022); Yu et al. (2022). In Setting (II), $\alpha_{k,1} = 1$ and $\alpha_{k,2} = 0.6$, so the factor strengths differ and we have one strong factor and one weak factor. In Setting (III), even the strongest factor becomes weak, as $\alpha_{k,1} = 0.8$ and $\alpha_{k,2} = 0.6$. In Setting (IV), we may encounter what we refer to as 'pseudo weak factors'. This is because even if we generate four factors to be equally strong, the four population eigenvalues are likely to be separated, resulting in certain factors exhibiting a 'weaker' behavior compared to others.

In each Setting (I)-(IV), the distinction between sub-settings (a) and (b) is intended to highlight the impact of signal accumulation through pre-averaging procedure as introduced

in Section 3.2, which may subsequently affect the performance of our proposed rank estimators (BCorTH). In general, pre-averaging takes advantage in sub-setting (b) when all entries of $\mathbf{A}_k$ share the same sign, which leads to greater signal accumulation. For a more detailed analysis, please refer to the simulation results in Section 3.4.2.

To test the performance of different estimation methods under heavy-tailed distributions, we consider two distributions for the innovation processes of $\mathscr{F}_t$, $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$: 1) i.i.d. standard normal; 2) i.i.d. t_3. Thus, there are totally sixteen profiles considered. For all profiles above, we set $r_{e,k} = 2$ for all k.

4.3.2 Core tensor rank estimations

We compare the performance of our BCorTh with other competitors for estimating the rank of core tensors. All the twelve profiles in Section 4.3.1 and three settings of different dimensions are considered:

i. $K = 2$, $T = 100$, $d_1 = d_2 = 40$;

ii. $K = 2$, $T = 200$, $d_1 = d_2 = 80$;

iii. $K = 3$. $T = 200$, $d_k = 25$.

The methods we consider include iTIP-ER by Han et al. (2022), α-PCA-ER by Chen and Fan (2021), PE-ER by Yu et al. (2022) and Barigozzi et al. (2023b), RTFA-ER by He et al. (2022) and He et al. (2023a), and MRTS-ER by Barigozzi et al. (2023b). These methods are based on the spirit of eigenvalue-ratio (ER) criteria of the (adjusted) sample covariance matrices, which are defined differently in their corresponding processes of factor loading estimations. Among them, the robust tensor factor analysis (RTFA) proposed by He et al. (2022) and He et al. (2023a), as well as the Matrix Kendall's tau (MRTS) by He et al. (2022), are robust procedures. The α-PCA-ER method by Chen and Fan (2021) is implemented with $\alpha = 0$ (the performances for $\alpha \in \{-1, 0, 1\}$ are comparable according to Yu et al. (2022)).

Table 4.1 records the correct proportion over 500 repetitions of different rank estimators under different settings and dimensions. For BCorTh, we set the number of bootstrapped samples to be $B = 50$ for settings with $K = 2$, and $B = 10$ for settings with $K = 3$. We have tested that reducing B from 50 to 10 does not significantly change the results of BCorTh. Also, for $K = 3$, we do not report the results for α-PCA-ER and MRTS-ER since they are only designed for matrix time series ($K = 2$).

From Table 4.1, all rank estimators perform better when T, d_k or K increases, and BCorTh generally outperforms all competitors in every setting and dimension we consider. When $K = 2$, it is obvious that all methods, except BCorTh, perform quite poorly in Setting (II) and (III) when weak factors are present (especially in sub-setting (b)), while BCorTh can still give relatively good performances. In Setting (IV) with four strong factors (which may lead to some 'pseudo weak factors'), most estimators still perform quite well in detecting the correct number of four factors. MRTS-ER and α-PCA-ER give extremely poor estimates in all settings except (Ia). BCorTh is robust as well, since changing the error distribution from normal to t_3 does not have large effects in its estimation accuracy. When $K = 3$, the accuracy of all estimators increases, and BCorTh still gives the best performances among all competitors. However, as a general recommendation, we suggest using our method when d_{-k} is not excessively large. This is because, as demonstrated in Chapter 3.4.1, the computational cost of the initial pre-averaging and projection algorithm increases rapidly with d_{-k}, resulting in a significant increase in the total computational cost required for obtaining BCorTh in such scenario.

4.4 Real Data Analysis

In this section, we conduct a real data analysis using two distinct datasets. We estimate both the factor loading spaces and the number of factors in each dataset, presenting all the methods introduced in Chapter 3 and Chapter 4.

4.4.1 Fama-French portfolio returns

We analyse a set of Fama-French portfolio returns data formed on size and operating profitability. Stocks are categorized into 10 different sizes (market equity, using NYSE market equity deciles) and 10 different operating profitability (OP) levels (using NYSE OP deciles. OP is annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by book equity for the last fiscal year end). These levels, and hence the stocks in each category, are allocated at the end of June each year. Moreover, the stocks in each of the 10×10 categories form exactly two portfolios, one being value weighted, and the other of equal weight. Hence, there are two sets of 10×10 portfolios with their time series of returns observed. We use monthly data from July 1973 to June 2021, so that $T = 576$, and each data tensor we have thus has size $10 \times 10 \times 576$. For more details, please visit

Setting	BCorTh		iTIP-ER		PE-ER		α-PCA-ER		RTFA-ER		MRTS-ER	
	$\mathcal{N}$	t_3	$\mathcal{N}$	t_3	$\mathcal{N}$	t_3	$\mathcal{N}$	t_3	$\mathcal{N}$	t_3	$\mathcal{N}$	t_3
$K=2,\ T=100,\ d_1=d_2=40$												
(Ia)	.994	.988	.894	.866	.892	.878	.842	.810	.894	.884	.842	.810
(Ib)	.998	1.000	.830	.794	.908	.896	.014	.012	.906	.898	.040	.030
(IIa)	.994	.966	.754	.640	.762	.690	.010	.038	.768	.702	.026	.022
(IIb)	.954	.928	.070	.084	.126	.080	.014	.046	.092	.070	.026	.056
(IIIa)	.758	.684	.556	.482	.334	.408	.054	.092	.320	.372	.048	.068
(IIIb)	.772	.712	.108	.202	.052	.116	.122	.180	.048	.106	.128	.182
(IVa)	.998	.986	.930	.918	.934	.946	.946	.940	.954	.946	.932	.926
(IVb)	.834	.828	.876	.780	.906	.842	.010	.012	.886	.882	.024	.028
$K=2,\ T=200,\ d_1=d_2=80$												
(Ia)	.994	.990	.926	.910	.768	.854	.944	.904	.768	.842	.768	.804
(Ib)	.998	1.000	.966	.956	.730	.898	.006	.012	.686	.884	.020	.022
(IIa)	.992	.972	.814	.796	.438	.632	.000	.010	.406	.602	.000	.000
(IIb)	.998	.998	.258	.176	.332	.230	.004	.024	.296	.230	.004	.016
(IIIa)	.594	.620	.188	.292	.014	.092	.000	.010	.016	.092	.000	.010
(IIIb)	.978	.968	.096	.074	.078	.082	.060	.074	.070	.080	.028	.054
(IVa)	1.000	.998	.996	.930	.932	.886	.986	.920	.932	.926	.870	.882
(IVb)	.996	.998	.994	.932	.956	.948	.004	.008	.910	.938	.058	.032
$K=3,\ T=200,\ d_1=d_2=25$												
(Ia)	1.000	1.000	.996	.990	.992	.980	/	/	.776	.732	/	/
(Ib)	1.000	.998	.998	.986	.972	.982	/	/	.998	1.000	/	/
(IIa)	1.000	.988	.988	.968	.834	.854	/	/	.106	.072	/	/
(IIb)	.994	.992	.988	.968	.948	.940	/	/	.948	.920	/	/
(IIIa)	.930	.856	.910	.866	.522	.544	/	/	.100	.056	/	/
(IIIb)	.996	.996	.920	.868	.764	.738	/	/	.804	.760	/	/

Table 4.1 Correct Proportion ($(\widehat{r}_1,\widehat{r}_2)=(2,2)$ for $K=2$, $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3)=(2,2,2)$ for $K=3$) of rank estimation under different settings, dimensions and error distributions ($\mathcal{N}$ for normally distributed errors, t_3 for t_3 distributed errors).

https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_100_port_szme_op.html

Since the market factor is certainly pervasive in financial returns, we use the following CAPM to remove its effects and facilitate detection of potentially weaker factors:

$$\mathbf{y}_t = \bar{\mathbf{y}} + \boldsymbol{\beta}(x_t - \bar{x}) + \mathbf{e}_t,$$

where $\mathbf{y}_t \in \mathbb{R}^{100}$ contains the returns of the 100 portfolios at time t, x_t is the return for the NYSE composite index at time t, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100})^{\mathrm{T}}$ is the vector of β's for the 100

portfolios. Least squares estimation leads us to

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(\mathbf{y}_t - \bar{\mathbf{y}})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}.$$

Hence the data we analyse is $\{\mathbf{y}_t - \bar{\mathbf{y}} - \widehat{\boldsymbol{\beta}}(x_t - \bar{x})\}_{t=1,\ldots,576}$, with each observed vector reshaped into a 10×10 tensor.

	BCorTh		iTIP-ER		PE-ER		α-PCA-ER		RTFA-ER		MRTS-ER	
	$\widehat{r}_1$	$\widehat{r}_2$	$\widehat{r}_1$	$\widehat{r}_2$	$\widehat{r}_1$	$\widehat{r}_2$	$\widehat{r}_1$	$\widehat{r}_2$	$\widehat{r}_1$	$\widehat{r}_2$	$\widehat{r}_1$	$\widehat{r}_2$
Value Weighted	2	2	2	2	2	2	2	2	2	2	2	2
Equal Weight	2	2	2	2	2	2	2	2	2	2	2	2

Table 4.2 Rank estimators for Fama-French Portfolios.

Table 4.2 shows that all of the rank estimators we consider, including BCorTh, give $\widehat{r}_1 = \widehat{r}_2 = 2$ for both value weighted and equal weight portfolios. Using $(\widehat{r}_1, \widehat{r}_2) = (2,2)$, we estimate the factor loading spaces by our iterative projection (PROJ) method and compare to iTIPUP (Han et al., 2020) and PE (Yu et al., 2022) and Barigozzi et al. (2023b)). Similar to Wang et al. (2019), we show the estimated loading matrices after a varimax rotation that maximizes the variance of the squared factor loadings, scaled by 30 for a cleaner view. To save space, we only show the results for value weighted portfolios in Table 4.3 and 4.4.

Method	Factor	OP1	OP2	OP3	OP4	OP5	OP6	OP7	OP8	OP9	OP10
PROJ	1	8	-3	-6	-7	-8	-9	-12	-11	-12	-14
	2	26	10	7	5	4	3	1	2	1	0
iTIPUP	1	-25	-14	-9	-1	-2	-1	0	-3	1	-2
	2	12	-13	-5	-8	-7	-8	-10	-11	-9	-9
PE	1	28	8	4	2	1	0	0	0	0	2
	2	5	-5	-7	-9	-10	-10	-11	-11	-12	-12

Table 4.3 OP Factor Loading Matrices for Value Weighted Portfolios after rotation and scaling. Magnitudes larger than 8 are highlighted in red.

From Table 4.3, we can see that from PROJ and PE, OP1 itself (possibly OP2 as well) forms one group while OP6 to OP10 form another. iTIPUP also gives a clear grouping effect but is a bit different from the other two methods. Nevertheless, we can say that OP1 and OP2 possibly represents "low operating profitability", while OP6 to OP10 are "high operating profitability", and are governed by different factors. For Size, from the three

Method	Factor	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
PROJ	1	-11	-11	-11	-10	-11	-9	-9	-6	-4	10
	2	-2	1	0	3	4	6	6	8	10	25
iTIPUP	1	-11	-13	-13	-13	-11	-9	-3	-2	0	9
	2	6	2	-2	-2	-5	-9	-12	-12	-14	-16
PE	1	-14	-15	-12	-10	-9	-6	-4	-1	2	9
	2	5	1	-2	-5	-7	-10	-11	-13	-14	-15

Table 4.4 Size Factor Loading Matrices for Value Weighted Portfolios after rotation and scaling. Magnitudes larger than 8 are highlighted in red.

methods, S1 to S5 form one group ("small size") while another group contains at least S9 and S10 ("large size"), possibly S6 to S8 as well.

4.4.2 NYC taxi traffic

We analyse taxi traffic pattern in New York city. The data includes all individual taxi rides operated by Yellow Taxi within New York City, published at

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

To simplify the discussion, we only consider rides within Manhattan Island. The dataset contains 1.1 billion trip records within the period of January 1, 2011 to December 31, 2021. Each trip record includes fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Our study focuses on the pick-up and drop-off dates/times, and pick-up and drop-off locations of each ride.

The pick-up and drop-off locations in Manhattan are coded according to 69 predefined zones and we will use them to classify the pick-up and drop-off locations. Furthermore, we divide each day into 24 hourly periods, with the first hourly period from 0am to 1am. The total number of rides moving among the zones within each hour is recorded, yielding a $\mathcal{X}_t \in \mathbb{R}^{69\times 69\times 24}$ tensor for each day. More specifically, $x_{i1,i2,i3,t}$ is the number of trips from zone i_1 (the pick-up zone) to zone i_2 (the drop-off zone) and the pickup time is within the i_3-th hourly period on day t. We consider business day and non-business day separately. Hence we will analyse two tensor time series. The business-day series is 2770 days long, and the non-business-day series is 1248 days long, within the period of January 1, 2011 to December 31, 2021.

We first estimate the rank of the core tensors using BCorTh as well as other state-of-the-art methods. BCorTh gives $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3) = (3,3,2)$ for business-day series, and $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3) = (3,2,2)$ for non-business-day series, while $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3) = (1,1,1)$ for iTIP-ER, PE-ER and RTFA-ER. However, based on our common knowledge and previous analysis conducted by Chen et al. (2022), $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3) = (1,1,1)$ is obviously not a reasonable choice for the rank of the core tensor, since a single factor can hardly be sufficient to reveal all traffic patterns. It is very likely that all of iTIP-ER, PE-ER and RTFA-ER fail to detect the weak factors in both time series, since these methods are designed to analyse pervasive factors only. For ease of presentation and comparison, we use $(\widehat{r}_1,\widehat{r}_2,\widehat{r}_3) = (3,3,2)$ for both business-day and non-business-day series to estimate their factor loadings, and present the results of our iterative projection estimator.

Figure 4.1 and 4.2 show the heatmaps of the loading matrices $\mathbf{A}_1$ (pick-up locations) for the business day and non-business day series, respectively. It is seen that during business days, the midtown/Times square area (tourism and office buildings) is heavily loaded on Factor 1, east village/lower east (arts, music venues and restaurants) on Factor 2 and upper east side (affluent neighborhoods and museums) on Factor 3. For non-business days, the overall pattern for the three factors is generally similar, but with some non-negligible differences. The area around Penn Station (large transportation hub) loads extremely heavily in Factor 2, while its loading is much lighter than the midtown center and midtown east for business day series, where a lot of office buildings locate.

Figure 4.3 and 4.4 show the loading matrices $\mathbf{A}_2$ (drop-off locations) for the business day and non-business day series, respectively. For both business days and non-business days, the drop-off factor matrices are quite similar to their pick-up factors. Similarly, the area around Penn Station is heavily loaded in non-business days, but is overshadowed by midtown center in business days. In addition, in Factor 1 of non-business days series, west village (arts, music venues and theatres) loads heavily together with east village.

Table 4.5 and 4.6 show the loading matrices $\mathbf{A}_3$ (time of day) for business days and non-business days, respectively. For ease of presentation, we show the estimated loading matrices after a varimax rotation, scaled by 30 for a cleaner view. For business days, it can be seen that day-time business hour (9am to 4pm) and evening hours (7pm to 12am) load heavily on Factor 1, while morning rush-hours (6am to 9am), evening rush-hours (5pm to 7pm) and night life hours (0am to 2am) load heavily on Factor 2. For non-business days, the patterns of estimated factors are significantly different: Evening hours from 6pm to 1am load heavily on Factor 1, while late-night hours from 1am to 5am load heavily on Factor 2. The different factor loadings reveal the difference between people's travelling

habits in business days and non-business days. During non-business days, morning (and evening) rush-hours and day-time business hours no longer appear in the factors, while people tend to travel more frequently by taxi at evening and at night, and their night life lasts to much later hours than in the business days.

	0am		2		4		6		8		10		12pm		2		4		6		8		10		12am
1	6	3	2	1	1	0	1	4	7	9	8	7	7	7	7	7	6	4	6	8	9	8	8	7	
2	12	7	4	2	1	-2	-11	-16	-8	0	3	-1	-2	0	-2	0	3	-7	-9	-2	3	0	3	8	

Table 4.5 Estimated loading matrix $\mathbf{A}_3$ for hour of day fibre, business day, after rotation and scaling. Magnitudes larger than 7 are highlighted in red.

	0am		2		4		6		8		10		12pm		2		4		6		8		10		12am
1	9	-1	-1	-1	-2	-1	1	2	3	5	5	5	5	5	5	5	4	6	8	9	9	9	10	12	
2	0	17	14	13	10	5	1	-1	-1	-2	0	2	4	4	5	4	4	2	0	0	2	0	-2	-4	

Table 4.6 Estimated loading matrix $\mathbf{A}_3$ for hour of day fibre, non-business day, after rotation and scaling. Magnitudes larger than 7 are highlighted in red.

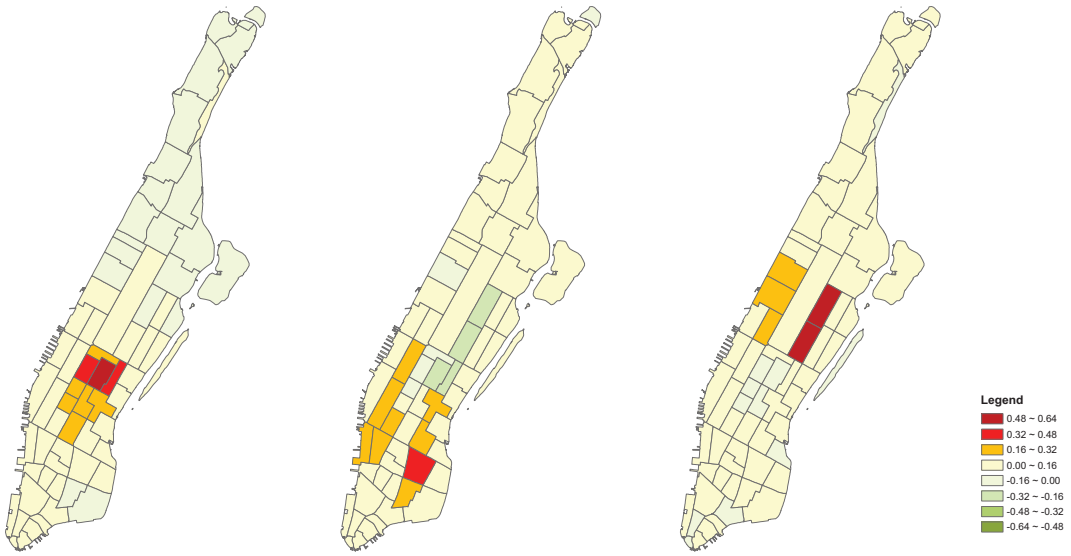


Fig. 4.1 Loadings on three pickup factors for business day series.

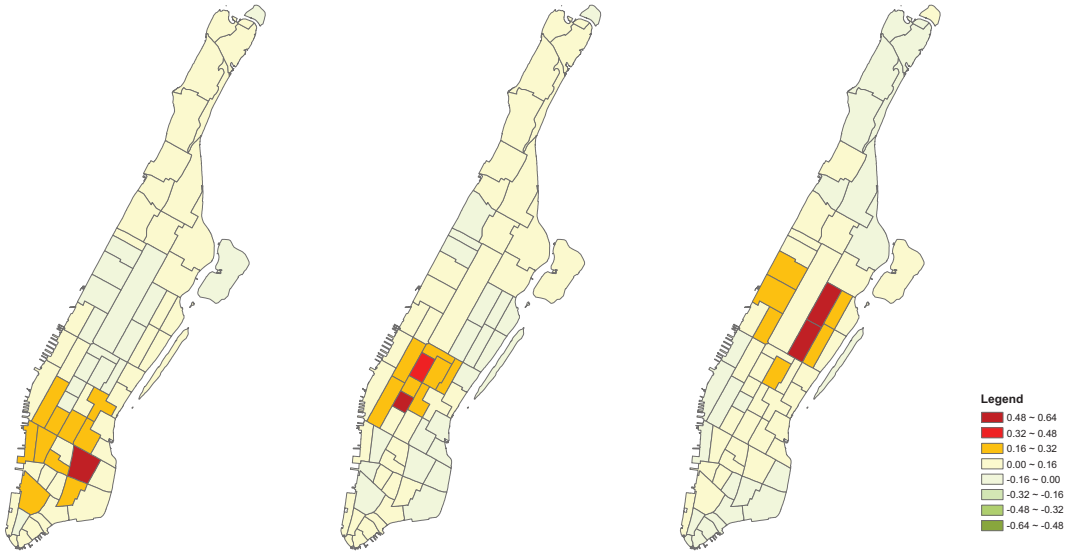


Fig. 4.2 Loadings on three pickup factors for non-business day series.

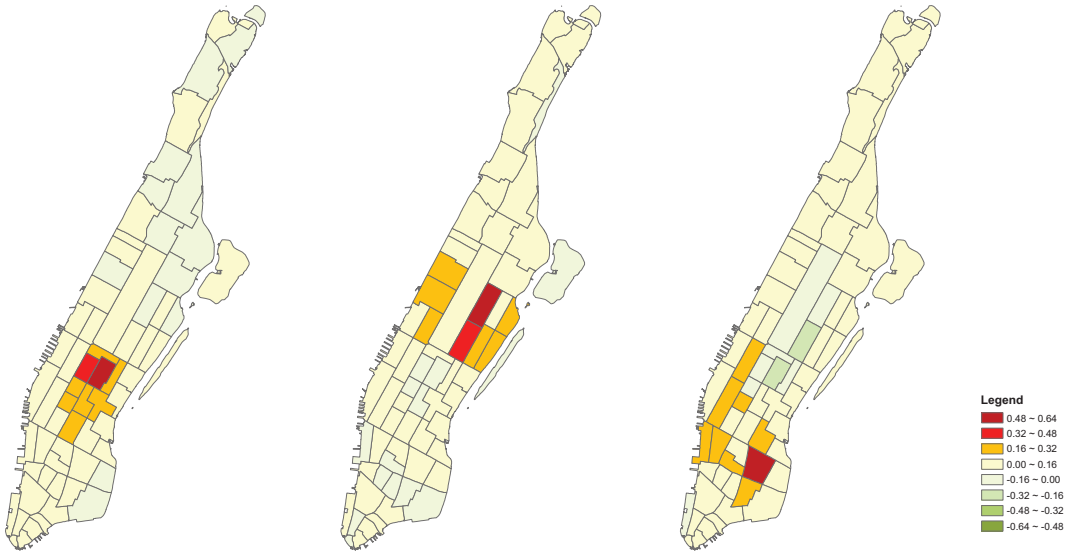


Fig. 4.3 Loadings on three dropoff factors for business day series.

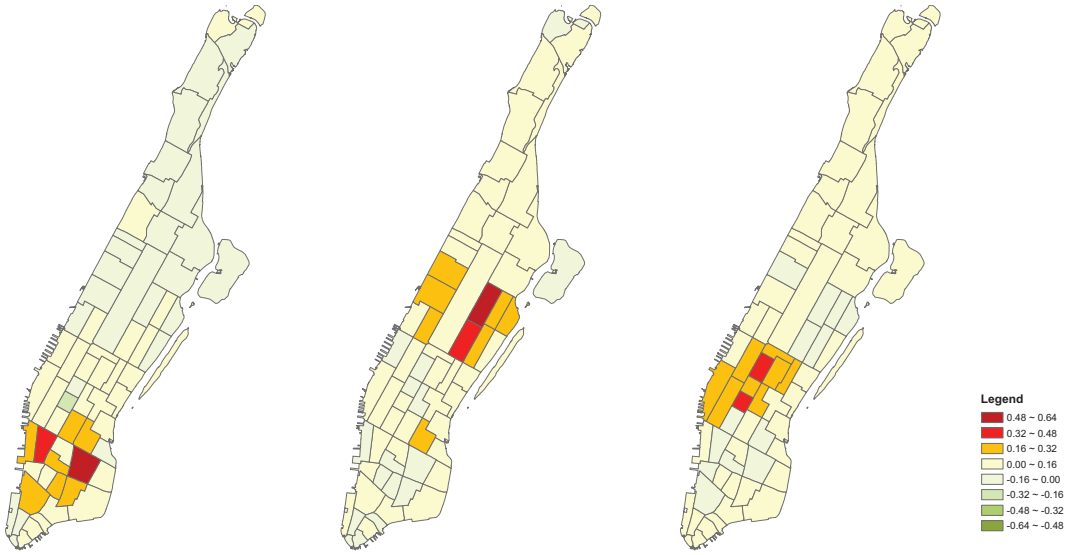


Fig. 4.4 Loadings on three dropoff factors for non-business day series.

4.5 A Brief Introduction to `TensorPreAve`

All our methods in Chapter 3 and Chapter 4 are written into an R package `TensorPreAve` published on CRAN and GitHub. For a given data of tensor time series, the function `rank_factors_est` can estimate both the rank of core tensors and factor loading matrices simultaneously.

Alternatively, there are individual functions designed for each step of our estimation procedure. The pre-averaging method is implemented in function `pre_est`, with default parameters chosen carefully to be applied in most scenarios and dimensions. We also provide an alternative way to tune the parameter j manually using the function `pre_eigenplot` to observe the eigenvalues of a random sample (see Chapter 3.2.3 for more details). With initial estimated directions of the strongest factors obtained by the function `pre_est`, users can feed them into the function `iter_proj` to get the iterative projection estimator. Finally, the rank estimator (BCorTh) can be obtained by the function `bs_cor_rank` by inputting the refined directions from `iter_proj`. The default number of bootstrapped samples is 50, but users can reduce it to 10 to save computational time when tensor dimension is large. Please refer to the vignettes and manual of `TensorPreAve` for more details.

4.6 Proof of Theorems

Proof of Theorem 4.1. First, it is easy to see that (RE2) implies that

$$\operatorname{diag}(\mathbf{\Sigma}_{y,m+1}^{(k)}) = \breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k}\mathbf{A}_{-k}^{\mathrm{T}}\breve{\mathbf{q}}_{-k}^{(m)}\operatorname{diag}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})(1+o(1)). \tag{4.7}$$

Define

$$\begin{aligned}
\mathbf{Qm}_1^{(k)} &:= \operatorname{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)})\mathbf{A}_k(\mathbf{I}_{r_k}\otimes\breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{-k}),\\
\mathbf{Qm}_2^{(k)} &:= \operatorname{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)})(\breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\otimes\mathbf{I}_{d_k})\operatorname{diag}^{1/2}(\mathbf{\Sigma}_{\varepsilon,1}^{(k)},\ldots,\mathbf{\Sigma}_{\varepsilon,d_{-k}}^{(k)}),\\
\mathbf{Qm}_3^{(k)} &:= \operatorname{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)})\mathbf{A}_{e,k}(\mathbf{I}_{r_{e,k}}\otimes\breve{\mathbf{q}}_{-k}^{(m)\mathrm{T}}\mathbf{A}_{e,-k}),
\end{aligned} \tag{4.8}$$

where $\mathbf{\Sigma}_{y,m+1}^{(k)}$ is defined in (4.5). Then we have

$$\mathbf{R}_{y,m+1}^{(k)} = \sum_{j=1}^{3}\mathbf{Qm}_j^{(k)}\mathbf{Qm}_j^{(k)\mathrm{T}}.$$

Consider

$$\begin{aligned}
\operatorname{tr}(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) &= \operatorname{tr}(\operatorname{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)})\cdot\breve{g}_{-k}^{(m)}\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\cdot\operatorname{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)}))\\
&= \operatorname{tr}[(\breve{g}_{-k}^{(m)})^{-1/2}\operatorname{diag}^{-1/2}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})(1+o(1))^{-1}\cdot\breve{g}_{-k}^{(m)}\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\\
&\quad\cdot(\breve{g}_{-k}^{(m)})^{-1/2}\operatorname{diag}^{-1/2}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})(1+o(1))^{-1}]\\
&= d_k(1+o(1)),
\end{aligned} \tag{4.9}$$

where the second equality used (4.7). At the same time, for $j\in[r_k]$,

$$\begin{aligned}
\lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) &= \lambda_j(\operatorname{diag}^{-1/2}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})(1+o(1))\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\operatorname{diag}^{-1/2}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})(1+o(1))\\
&= \lambda_j(\mathbf{U}_k^{\mathrm{T}}\operatorname{diag}^{-1}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})\mathbf{U}_k\mathbf{G}_k(1+o(1))),\ \text{ with}\\
\lambda_j(\mathbf{G}_k)\lambda_{\min}(\operatorname{diag}^{-1}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}))(1+o(1)) &\le \lambda_j(\mathbf{U}_k^{\mathrm{T}}\operatorname{diag}^{-1}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}})\mathbf{U}_k\mathbf{G}_k(1+o(1)))\\
&\le \lambda_j(\mathbf{G}_k)\lambda_{\max}(\operatorname{diag}^{-1}(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}))(1+o(1)),
\end{aligned}$$

so that by Assumption (RE2), there are generic constants $c,C>0$ such that

$$d_k^{\alpha_{k,j}} \asymp_P c\lambda_j(\mathbf{G}_k) \le \lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) \le C\lambda_j(\mathbf{G}_k) \asymp_P d_k^{\alpha_{k,j}} \tag{4.10}$$

in probability, where $\lambda_j(\mathbf{G}_k)$ being asymptotic in probability to $d_k^{\alpha_{k,j}}$ is given by (3.31) in Lemma 3.2. Using (4.10), there exists a constant $C>0$ (generic, different from the above)

so that for $j \in [r_k]$,

$$\frac{\lambda_1(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}})}{\lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}})} \leq C d_k^{\alpha_{k,1}-\alpha_{k,j}}.$$

in probability. Hence using (4.9), in probability, we have

$$d_k(1+o(1)) \leq \mathrm{tr}(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) \leq r_k \lambda_1(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) \leq r_k \cdot C d_k^{\alpha_{k,1}-\alpha_{k,j}} \lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}),$$

implying that, in probability, there exists a constant $C > 0$ such that as T, d_k are large enough,

$$\lambda_j(\mathbf{R}_{y,m+1}^{(k)}) \geq \lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) \geq C d_k^{1-\alpha_{k,1}+\alpha_{k,j}}/r_k > 1, \tag{4.11}$$

since $r_k = o(d_k^{1-\alpha_{k,1}+\alpha_{k,j}})$ by (RE2) for $j \in [r_k]$. For $j \in [d_k]/[r_k]$,

$$\begin{aligned}
\lambda_j(\mathbf{R}_{y,m+1}^{(k)}) &\leq \lambda_j(\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}) + \lambda_1(\mathbf{Qm}_2^{(k)}\mathbf{Qm}_2^{(k)\mathrm{T}}) + \lambda_1(\mathbf{Qm}_3^{(k)}\mathbf{Qm}_3^{(k)\mathrm{T}}) \\
&\leq 0 + \left\|\mathrm{diag}^{-1}(\mathbf{\Sigma}_{y,m+1}^{(k)})\right\| \left(\sum_{j=1}^{d_{-k}} (\check{\mathbf{q}}_{-k}^{(m)})_j^2 \left\|\mathbf{\Sigma}_{\varepsilon,j}^{(k)}\right\| + \check{\mathbf{q}}_{-k}^{(m)\mathrm{T}} \mathbf{A}_{e,-k} \mathbf{A}_{e,-k}^{\mathrm{T}} \check{\mathbf{q}}_{-k}^{(m)} \left\|\mathbf{A}_{e,k}\right\|^2 \right) \\
&\leq (\check{g}_{-k}^{(m)})^{-1} \left\|\mathrm{diag}^{-1}(\mathbf{A}_k \mathbf{A}_k^{\mathrm{T}})(1+o(1))\right\| \cdot O(1) \\
&= O_P(d_k^{\alpha_{k,1}} g_s^{-1}) = o_P(1),
\end{aligned} \tag{4.12}$$

where the third inequality is in probability after using (4.7), and the last line used statement I(m) in the proof of Theorem 3.2, that $\check{g}_{-k}^{(m)} \asymp_P g_s d_k^{-\alpha_{k,1}}$. This completes the proof of the theorem. □

Before proving Theorem 4.2, define

$$\mathbf{S}_{y,m+1}^{(k)} := \mathrm{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)}) \widetilde{\mathbf{\Sigma}}_{y,m+1}^{(k)} \mathrm{diag}^{-1/2}(\mathbf{\Sigma}_{y,m+1}^{(k)}).$$

It is then easy to see that

$$\widehat{\mathbf{R}}_{y,m+1}^{(k)} := \mathrm{diag}^{-1/2}(\mathbf{S}_{y,m+1}^{(k)}) \mathbf{S}_{y,m+1}^{(k)} \mathrm{diag}^{-1/2}(\mathbf{S}_{y,m+1}^{(k)}).$$

We state and prove the following lemma.

Lemma 4.1. *Assume all the assumptions in Theorem 4.2 hold. Then for $k \in [K]$,*

$$\max_{j\in[r_k]}\left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\mathbf{R}^{(k)}_{y,m+1})}-1\right| = O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}\sqrt{\frac{r}{T}}\left(1+K\sqrt{\frac{rd}{Tg_s}}+\frac{K^2r^{1/2}d}{T^{3/2}g_s}\right)\right),$$

$$\max_{j\in[d_k]}\left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1})}-1\right| = O_P\left(\sqrt{\frac{r}{T}}\left[1+d_k^{\alpha_{k,1}/2}g_s^{-1/2}\left(r_e^{1/2}+d_k^{1/2}+K\sqrt{\frac{rd}{T}}\right)+d_k^{\alpha_{k,1}}g_s^{-1}\frac{K^2r^{1/2}d}{T^{3/2}}\right]\right),$$

$$\max_{j\in[d_k]/[r_k]}|\lambda_j(\mathbf{S}^{(k)}_{y,m+1})-\lambda_j(\mathbf{R}^{(k)}_{y,m+1})| = O_P\left(d_k^{\alpha_{k,1}}g_s^{-1}\left\{\sqrt{\frac{(r_e+d_k)d_k}{T}}+\frac{K\sqrt{r(r_e+d_k)d}}{T}+\frac{K^2rd}{T^2}\right\}\right).$$

Proof of Lemma 4.1. Define, using the definition of $\mathbf{S}_{ij,m}$ defined in Lemma 3.7,

$$\mathbf{S}_{ij,r} := \mathrm{diag}^{-1/2}(\mathbf{\Sigma}^{(k)}_{y,m+1})\mathbf{S}_{ij,m}\mathrm{diag}^{-1/2}(\mathbf{\Sigma}^{(k)}_{y,m+1}),\quad i,j=1,2,3.$$

Then using (4.8), for $j\in[r_k]$, we can decompose

$$\left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\mathbf{R}^{(k)}_{y,m+1})}-1\right| \le \frac{1}{\lambda_j(\mathbf{R}^{(k)}_{y,m+1})}\left(\sum_{i=1}^{3}\left\|\mathbf{S}_{ii,r}-\mathbf{Qm}_i^{(k)}\mathbf{Qm}_i^{(k)\mathrm{T}}\right\|+2\sum_{i<j}\left\|\mathbf{S}_{ij,r}\right\|+\sum_{i\neq j}\left\|\check{\mathbf{S}}_{ij}\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})\right\|\right).$$

From (4.7) and Assumption (RE2), we have that

$$\lambda_1(\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})),\quad \lambda_{d_k}(\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1}))\asymp(\check{g}^{(m)}_{\text{-}k})^{-1}\asymp_P d_k^{\alpha_{k,1}}g_s^{-1},$$

where the last order is from statement I(m) in the proof of Theorem 3.2, that $\check{g}^{(m)}_{\text{-}k}d_k^{\alpha_{k,1}}\asymp_P g_s$. Hence coupled with the results from Lemma 3.7 and (4.11), and the fact that from an argument similar to (3.73) and the result of Theorem 3.2 that for large enough m,

$$\left\|\check{\mathbf{q}}^{(m)}_{\text{-}k}-\mathbf{U}_{\text{-}k,(1)}\right\| = O_P\left(K\sqrt{\frac{r}{T}}\right),$$

we have, using the notations in Lemma 3.7,

$$\begin{aligned}
\frac{\left\|\mathbf{S}_{11,r}-\mathbf{Qm}_1^{(k)}\mathbf{Qm}_1^{(k)\mathrm{T}}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &\le \frac{\left\|\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_{y,m+1}^{(k)})\right\|\left\|\mathbf{S}''_{11,m}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} = O_P\left(r_k d_k^{\alpha_{k,1}-\alpha_{k,j}-1}(\breve{g}_{-k}^{(m)})^{-1}\cdot\breve{g}_{-k}^{(m)}d_k^{\alpha_{k,1}}\sqrt{\frac{r}{T}}\right)\\
&= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}\sqrt{\frac{r}{T}}\right),\\
\frac{\left\|\mathbf{S}_{22,r}-\mathbf{Qm}_2^{(k)}\mathbf{Qm}_2^{(k)\mathrm{T}}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &\le \frac{\left\|\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_{y,m+1}^{(k)})\right\|\left\|\mathbf{S}_{22,m}-\mathbf{Q}_{2,m}^{(k)}\mathbf{Q}_{2,m}^{(k)\mathrm{T}}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})}\\
&= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1}\cdot\left\{\frac{K^2rd}{T^2}+\frac{d_k}{\sqrt{T}}+\frac{K\sqrt{rd_kd}}{T}\right\}\right),\\
\frac{\left\|\mathbf{S}_{33,r}-\mathbf{Qm}_3^{(k)}\mathbf{Qm}_3^{(k)\mathrm{T}}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1}\cdot\sqrt{\frac{r_e}{T}}\right),\\
\frac{\left\|\mathbf{S}_{12,r}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1/2}\left(\sqrt{\frac{rd_k}{T}}+\frac{Kr\sqrt{d}}{T}\right)\right),\\
\frac{\left\|\mathbf{S}_{13,r}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1/2}\sqrt{\frac{rr_e}{T}}\right),\\
\frac{\left\|\mathbf{S}_{23,r}\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1}\left(\sqrt{\frac{r_ed_k}{T}}+\frac{K\sqrt{rr_ed}}{T}\right)\right),\\
\frac{\left\|\check{\mathbf{S}}_{ij}\mathrm{diag}^{-1}(\boldsymbol{\Sigma}_{y,m+1}^{(k)})\right\|}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} &= O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}g_s^{-1}T^{-1}\right).
\end{aligned}$$

These implies that

$$\max_{j\in[r_k]}\left|\frac{\lambda_j(\mathbf{S}_{y,m+1}^{(k)})}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})}-1\right| = O_P\left(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1}\sqrt{\frac{r}{T}}\left(1+K\sqrt{\frac{rd}{Tg_s}}+\frac{K^2r^{1/2}d}{T^{3/2}g_s}\right)\right),$$

which is the first statement in the lemma.

For the second statement, for $j \in [d_k]$ and $k \in [K]$, defining $\mathbf{u}_j$ to be a unit vector with the j-th position being 1 and 0 elsewhere,

$$\begin{aligned}\left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1})}-1\right| &= \left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\mathbf{S}^{(k)}_{y,m+1}\mathrm{diag}^{-1}(\mathbf{S}^{(k)}_{y,m+1}))}-1\right| \leq \max_{j\in[d_k]}|\lambda_j(\mathrm{diag}(\mathbf{S}^{(k)}_{y,m+1}))-1| \\ &= \max_{j\in[d_k]}|\lambda_j(\mathrm{diag}(\mathbf{S}^{(k)}_{y,m+1}-\mathbf{R}^{(k)}_{y,m+1}))| = \max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}(\mathbf{S}^{(k)}_{y,m+1}-\mathbf{R}^{(k)}_{y,m+1})\mathbf{u}_j| \\ &\leq \max_{j\in[d_k]}\left|\mathbf{u}_j^{\mathrm{T}}\left(\mathbf{S}^{(k)}_{11,r}-\mathbf{Q}\mathbf{m}_1^{(k)}\mathbf{Q}\mathbf{m}_1^{(k)\mathrm{T}}\right)\mathbf{u}_j\right| + 2\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}^{(k)}_{12,r}\mathbf{u}_j| + 2\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}^{(k)}_{13,r}\mathbf{u}_j| \\ &+ \sum_{i=2}^{3}|\mathbf{u}_j^{\mathrm{T}}(\mathbf{S}^{(k)}_{ii,r}-\mathbf{Q}\mathbf{m}_i^{(k)}\mathbf{Q}\mathbf{m}_i^{(k)\mathrm{T}})\mathbf{u}_j| + 2|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}^{(k)}_{23,r}\mathbf{u}_j| + \sum_{i\neq j}\left\|\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})\check{\mathbf{S}}_{ij}\right\|.\end{aligned}$$

But similar to the above,

$$\begin{aligned}&\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}(\mathbf{S}_{11,r}-\mathbf{Q}\mathbf{m}_1^{(k)}\mathbf{Q}\mathbf{m}_1^{(k)\mathrm{T}})\mathbf{u}_j| = O_P(\left\|T^{-1}\mathbf{Z}_f^{(k)}\mathscr{A}_{f,T}\mathbf{M}_T\mathscr{A}_{f,T}^{\mathrm{T}}\mathbf{Z}_f^{(k)\mathrm{T}}-\mathbf{I}_r\right\|) = O_P\left(\sqrt{\frac{r}{T}}\right), \\ &\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}^{(k)}_{12,r}\mathbf{u}_j| = O_P\left((\check{g}^{(m)}_{-k})^{-1/2}\left(\sqrt{\frac{rd_k}{T}}+\frac{Kr\sqrt{d}}{T}\right)\right), \quad \max_{j\in[d_{-k}]}|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}_{13,r}\mathbf{u}_j| = O_P\left((\check{g}^{(m)}_{-k})^{-1/2}\sqrt{\frac{rr_e}{T}}\right), \\ &\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}(\mathbf{S}_{22,r}-\mathbf{Q}\mathbf{m}_2^{(k)}\mathbf{Q}\mathbf{m}_2^{(k)\mathrm{T}})\mathbf{u}_j| = O_P\left((\check{g}^{(m)}_{-k})^{-1}\left\{\frac{K^2rd}{T^2}+\frac{d_k}{\sqrt{T}}+\frac{K\sqrt{rd_kd}}{T}\right\}\right), \\ &\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}\mathbf{S}^{(k)}_{23,r}\mathbf{u}_j| = O_P\left((\check{g}^{(m)}_{-k})^{-1}\left(\sqrt{\frac{r_ed_k}{T}}+\frac{K\sqrt{rr_ed}}{T}\right)\right), \\ &\max_{j\in[d_k]}|\mathbf{u}_j^{\mathrm{T}}(\mathbf{S}_{33,r}-\mathbf{Q}\mathbf{m}_3^{(k)}\mathbf{Q}\mathbf{m}_3^{(k)\mathrm{T}})\mathbf{u}_j| = O_P\left((\check{g}^{(m)}_{-k})^{-1}\sqrt{\frac{r_e}{T}}\right), \quad \left\|\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})\check{\mathbf{S}}_{ij}\right\| = O_P((\check{g}^{(m)}_{-k})^{-1}T^{-1}).\end{aligned}$$

Hence the above implies that

$$\begin{aligned}\max_{j\in[d_k]}\left|\frac{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}{\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1})}-1\right| &= O_P\left(\sqrt{\frac{r}{T}}\left[1+(\check{g}^{(m)}_{-k})^{-1/2}\left(r_e^{1/2}+d_k^{1/2}+K\sqrt{\frac{rd}{T}}\right)+(\check{g}^{(m)}_{-k})^{-1}\frac{K^2r^{1/2}d}{T^{3/2}}\right]\right) \\ &= O_P\left(\sqrt{\frac{r}{T}}\left[1+d_k^{\alpha_{k,1}/2}g_s^{-1/2}\left(r_e^{1/2}+d_k^{1/2}+K\sqrt{\frac{rd}{T}}\right)+d_k^{\alpha_{k,1}}g_s^{-1}\frac{K^2r^{1/2}d}{T^{3/2}}\right]\right),\end{aligned}$$

which is the second statement of the lemma.

Finally, consider for $j \in [d_k]/[r_k]$,

$$\begin{aligned}|\lambda_j(\mathbf{S}^{(k)}_{y,m+1}) - \lambda_j(\mathbf{R}^{(k)}_{y,m+1})| &\leq \left|\lambda_j\left(\mathbf{S}_{11,r} + \sum_{j=2}^{3}(\mathbf{S}_{1j,r} + \mathbf{S}_{j1,r} + \mathbf{Qm}^{(k)}_j \mathbf{Qm}^{(k)\mathrm{T}}_j)\right) - \lambda_j(\mathbf{R}^{(k)}_{y,m+1})\right| \\ &\quad + 2\|\mathbf{S}_{23,r}\| + \sum_{j=2}^{3}\|\mathbf{S}_{jj,r} - \mathbf{Qm}^{(k)}_j \mathbf{Qm}^{(k)\mathrm{T}}_j\| + \sum_{i\neq j}\|\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})\check{\mathbf{S}}^{(k)}_{ij}\| \\ &= 2\|\mathbf{S}_{23,r}\| + \sum_{j=2}^{3}\|\mathbf{S}_{jj,r} - \mathbf{Qm}^{(k)}_j \mathbf{Qm}^{(k)\mathrm{T}}_j\| + \sum_{i\neq j}\|\mathrm{diag}^{-1}(\mathbf{\Sigma}^{(k)}_{y,m+1})\check{\mathbf{S}}^{(k)}_{ij}\|, \end{aligned} \tag{4.13}$$

if we can show that for large enough T, d_k, in probability, for $j \in [d_k]/[r_k]$,

$$\lambda_j\left(\mathbf{S}_{11,r} + \sum_{j=2}^{3}(\mathbf{S}_{1j,r} + \mathbf{S}_{j1,r} + \mathbf{Qm}^{(k)}_j \mathbf{Qm}^{(k)\mathrm{T}}_j)\right) = \lambda_j(\mathbf{R}^{(k)}_{y,m+1}). \tag{4.14}$$

With (4.13), from the rates obtained in previous arguments, we then have

$$|\lambda_j(\mathbf{S}^{(k)}_{y,m+1}) - \lambda_j(\mathbf{R}^{(k)}_{y,m+1})| = O_P\left(d_k^{\alpha_{k,1}} g_s^{-1}\left\{\sqrt{\frac{(r_e+d_k)d_k}{T}} + \frac{K\sqrt{r(r_e+d_k)d}}{T} + \frac{K^2 rd}{T^2}\right\}\right),$$

which is the third statement of the lemma. Hence it remains to show (4.14). To this end, define the kernel of $\mathbf{A}^{\mathrm{T}}_k \mathrm{diag}^{-1/2}(\mathbf{\Sigma}^{(k)}_{y,m+1})$ to be

$$\mathscr{N} := \{\mathbf{w} : \mathbf{A}^{\mathrm{T}}_k \mathrm{diag}^{-1/2}(\mathbf{\Sigma}^{(k)}_{y,m+1})\mathbf{w} = \mathbf{0}\} \equiv \{\mathbf{w} : \mathbf{A}^{\mathrm{T}}_k \mathbf{w} = \mathbf{0}\}$$

in probability by (4.7) and assumption (RE2), and define $\mathscr{N}^{\perp}$ the corresponding row space, which has $\mathscr{N}^{\perp} = \mathrm{span}(\mathbf{A}_k)$ in probability by the above argument. Hence by Assumption (L1) that $\mathbf{A}_k$ is of full rank, the dimension of $\mathscr{N}^{\perp}$ is exactly r_k (in probability), and the rank-nullity theorem implies that, in probability, the dimension of $\mathscr{N}$ is exactly $d_k - r_k$. Consider, by (3.80) and Assumption (L1), for some constant $c > 0$, we have in probability that

$$\lambda_{r_k}(\mathbf{S}_{11,r}) \geq \check{g}^{(m)}_{-k} \min_{\|\mathbf{w}\|=1:\mathbf{w}\in\mathscr{N}^{\perp}} \left\|\mathbf{w}^{\mathrm{T}}\mathrm{diag}^{-1/2}(\mathbf{\Sigma}^{(k)}_{y,m+1})\mathbf{A}_k\right\|^2 \lambda_r(T^{-1}\mathbf{Z}^{(k)}_f \mathscr{A}_{f,T}\mathbf{M}_T \mathscr{A}^{\mathrm{T}}_{f,T}\mathbf{Z}^{(k)\mathrm{T}}_f) \geq c d_k^{\alpha_{k,r_k}}.$$

At the same time, we have

$$\left\|\mathbf{Qm}^{(k)}_2 \mathbf{Qm}^{(k)\mathrm{T}}_2 + \mathbf{Qm}^{(k)}_3 \mathbf{Qm}^{(k)\mathrm{T}}_3\right\| = O_P((\check{g}^{(m)}_{-k})^{-1}) = O_P(d_k^{\alpha_{k,1}} g_s^{-1}),$$

and earlier calculations show that

$$\|\mathbf{S}_{12,r}\| = O_P\left(d_k^{\alpha_{k,1}} g_s^{-1/2}\left(\sqrt{\frac{rd_k}{T}} + \frac{Kr\sqrt{d}}{T}\right)\right), \quad \|\mathbf{S}_{13,r}\| = O_P\left(d_k^{\alpha_{k,1}} g_s^{-1/2}\sqrt{\frac{rr_e}{T}}\right).$$

Hence in probability, the eigenvalues of $\mathbf{S}_{11,r}$ are dominating those in

$$\mathbf{G} := \sum_{j=2}^{3}(\mathbf{S}_{1j,r} + \mathbf{S}_{j1,r} + \mathbf{Qm}_j^{(k)}\mathbf{Qm}_j^{(k)\mathrm{T}}).$$

Hence the r_k eigenvectors corresponding to the largest r_k eigenvalues of $\mathbf{G}+\mathbf{S}_{11,r}$ coincide with those of $\mathbf{S}_{11,r}$'s for large enough d_k, which are all necessarily in $\mathscr{N}^\perp$. This means that the (r_k+1)-th largest eigenvalue of $\mathbf{G}+\mathbf{S}_{11,r}$ and beyond will have eigenvectors in $\mathscr{N}$ since the dimension of $\mathscr{N}^\perp$ is r_k. Take $\mathbf{w} \in \mathscr{N}$, then it is easy to see that

$$\mathbf{w}^{\mathrm{T}}(\mathbf{G}+\mathbf{S}_{11,r})\mathbf{w} = \mathbf{w}^{\mathrm{T}}(\mathbf{Qm}_2^{(k)}\mathbf{Qm}_2^{(k)\mathrm{T}} + \mathbf{Qm}_3^{(k)}\mathbf{Qm}_3^{(k)\mathrm{T}})\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{R}_{y,m+1}^{(k)}\mathbf{w},$$

showing that $\lambda_j(\mathbf{G}+\mathbf{S}_{11,r}) = \lambda_j(\mathbf{R}_{y,m+1}^{(k)})$ for $j \in [d_k]/[r_k]$, which is (4.14). This completes the proof of the lemma. □

Proof of Theorem 4.2. For $j \in [r_k]$ and $k \in [K]$, we can decompose

$$\begin{aligned}
\lambda_j(\widehat{\mathbf{R}}_{y,m+1}^{(k)}) &\geq \lambda_j(\mathbf{R}_{y,m+1}^{(k)})\Bigg\{1 - \left|\frac{\lambda_j(\widehat{\mathbf{R}}_{y,m+1}^{(k)})}{\lambda_j(\mathbf{S}_{y,m+1}^{(k)})} - 1\right| - \left|\frac{\lambda_j(\mathbf{S}_{y,m+1}^{(k)})}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} - 1\right| \\
&\quad - \left|\frac{\lambda_j(\widehat{\mathbf{R}}_{y,m+1}^{(k)})}{\lambda_j(\mathbf{S}_{y,m+1}^{(k)})} - 1\right| \cdot \left|\frac{\lambda_j(\mathbf{S}_{y,m+1}^{(k)})}{\lambda_j(\mathbf{R}_{y,m+1}^{(k)})} - 1\right|\Bigg\} \\
&= \lambda_j(\mathbf{R}_{y,m+1}^{(k)})(1 + O_P(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1} a_T(0) + a_T(\alpha_{k,1}))) \\
&\succeq_P r_k d_k^{1-\alpha_{k,1}+\alpha_{k,j}}(1 + O_P(r_k d_k^{2\alpha_{k,1}-\alpha_{k,j}-1} a_T(0) + a_T(\alpha_{k,1}))),
\end{aligned}$$

where the second last line used the results in Lemma 4.1, and the last line used the result from Theorem 4.1, noting that

$$\begin{aligned}
a_T(0) &= \sqrt{\frac{r}{T}}\left[1 + g_s^{-1/2}\left(r_e^{1/2} + d_k^{1/2} + K\sqrt{\frac{rd}{T}}\right) + g_s^{-1}\frac{K^2 r^{1/2} d}{T^{3/2}}\right] \\
&= O\left(\sqrt{\frac{r}{T}}\left[1 + K\sqrt{\frac{rd}{Tg_s}} + \frac{K^2 r^{1/2} d}{T^{3/2} g_s}\right]\right).
\end{aligned}$$

Similarly, for $j \in [d_k]/[r_k]$,

$$\begin{aligned}\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1}) &\leq \lambda_j(\mathbf{R}^{(k)}_{y,m+1})\left(1+\left|\frac{\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1})}{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}-1\right|\right)+|\lambda_j(\mathbf{S}^{(k)}_{y,m+1})-\lambda_j(\mathbf{R}^{(k)}_{y,m+1})| \\ &\quad +|\lambda_j(\mathbf{S}^{(k)}_{y,m+1})-\lambda_j(\mathbf{R}^{(k)}_{y,m+1})|\cdot\left|\frac{\lambda_j(\widehat{\mathbf{R}}^{(k)}_{y,m+1})}{\lambda_j(\mathbf{S}^{(k)}_{y,m+1})}-1\right| \\ &= o_P(1)\cdot(1+o_P(1))+O_P(b_T) \leq 1+O_P(b_T),\end{aligned}$$

where the equality used the results of Lemma 4.1 and (4.12). This completes the proof of the theorem. □

Proof of Theorem 4.3. To prove the theorem, we note that the proofs of Theorem 4.1 and 4.2 will go through for the Bootstrapped fibres and a redefined $\mathbf{\Sigma}^{(k)}_{y,m+1}$ in (4.5) (with $\check{\mathbf{q}}^{(m)}_{\text{-}k}$ replaced by $\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\check{\mathbf{q}}^{(m)}_{\text{-}k}$) if the following hold:

I. $\left\|\mathbf{W}_b\right\| < \infty$;

II. $\check{\mathbf{q}}^{(m)\mathrm{T}}_{\text{-}k}\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\mathbf{A}_{\text{-}k}\mathbf{A}_{\text{-}k}^{\mathrm{T}}\mathbf{W}_b\mathbf{W}_b\check{\mathbf{q}}^{(m)}_{\text{-}k} \asymp_P g_d/d_k^{\alpha_{k,1}}$.

I. holds because we restrict the number of times a fibre can be chosen in a Bootstrap sample to be at most 8, so that $\left\|\mathbf{W}_b\right\| \leq 8 < \infty$.

Using the fact that $\left\|\check{\mathbf{q}}^{(m)}_{\text{-}k} - \mathbf{U}_{\text{-}k,(1)}\right\| = o_P(1)$, similar to (3.72), II. can be proved if we can show that

$$\mathbf{U}^{\mathrm{T}}_{\text{-}k,(1)}\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\mathbf{U}_{\text{-}k}\mathbf{G}_{\text{-}k}\mathbf{U}^{\mathrm{T}}_{\text{-}k}\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} \asymp_P g_s/d_k^{\alpha_{k,1}}.$$

With Assumption (L1'), this boils down to showing that

$$Q := \mathbf{U}^{\mathrm{T}}_{\text{-}k,(1)}\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}\mathbf{U}_{\text{-}k,(1)} > c > 0 \tag{4.15}$$

with high probability for some constant $c > 0$ as $T, d_k \to \infty$. To do this, define $\mathbf{U}_{\text{-}k,(1)} =: (u_j)_{j\in[d_{\text{-}k}]}$, and

$\mathscr{P}$:= Information on the position of $\xi_i^{(b)}$, $i \in [d_{\text{-}k}]$ on each column of $\mathbf{W}_b$;

$\mathscr{V}$:= Information on the values of the $\xi_i^{(b)}$, $i \in [d_{\text{-}k}]$;

$\mathscr{D}$:= Information on the values of u_j.

Let X_j be the number of $\xi_i^{(b)}$'s assigned at the jth row of $\mathbf{W}_b$ (say at column number $i_1,\ldots,i_{X_j}$). Hence the jth diagonal value of $\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}}$ is

$$(\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}})_{jj} = \sum_{\ell=1}^{X_j} \xi_{i_\ell}^{(b)}, \text{ meaning that } E[(\mathbf{W}_b\mathbf{W}_b)_{jj}|\mathscr{P}] = \frac{X_j}{2}, \ \operatorname{var}[(\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}})_{jj}|\mathscr{P}] = \frac{X_j}{4}.$$

But since $\sum_{j=1}^{d_{\text{-}k}} X_j = d_{\text{-}k}$, and each X_j is identically distributed, we have

$$E(X_j) = 1, \quad \operatorname{corr}(X_i, X_j) = -\frac{1}{d_{\text{-}k}-1} \ \text{ for } i \neq j.$$

Hence

$$E(Q|\mathscr{D}) = E[E(Q|\mathscr{P},\mathscr{D})|\mathscr{D}] = E\left[\sum_{j=1}^{d_{\text{-}k}} u_j^2 E[(\mathbf{W}_b\mathbf{W}_b^{\mathrm{T}})_{jj}|\mathscr{P}]|\mathscr{D}\right] = \frac{E(X_j)}{2}\sum_{j=1}^{d_{\text{-}k}} u_j^2 = \frac{1}{2},$$

$$\begin{aligned}
\operatorname{var}(Q|\mathscr{D}) &= E[\operatorname{var}(Q|\mathscr{D},\mathscr{P})|\mathscr{D}] + \operatorname{var}(E(Q|\mathscr{D},\mathscr{P})|\mathscr{D}), \\
&= E\left(\sum_{j=1}^{d_{\text{-}k}} u_j^4 \cdot \frac{X_j}{4}|\mathscr{D}\right) + \operatorname{var}\left(\sum_{j=1}^{d_{\text{-}k}} u_j^2 \cdot \frac{X_j}{2}|\mathscr{D}\right) \\
&= \frac{1}{4}\sum_{j=1}^{d_{\text{-}k}} u_j^4 + \sum_{j=1}^{d_{\text{-}k}} \frac{u_j^4}{4}\operatorname{var}(X_j) + \frac{1}{2}\sum_{i<j} u_i^2 u_j^2\left(-\frac{\operatorname{var}(X_j)}{d_{\text{-}k}-1}\right) \\
&= \frac{1}{4}\sum_{j=1}^{d_{\text{-}k}} u_j^4 + \frac{\operatorname{var}(X_j)}{4(d_{\text{-}k}-1)}\left(d_{\text{-}k}\sum_{j=1}^{d_{\text{-}k}} u_j^4 - 1\right).
\end{aligned}$$

But $\sum_{i=1}^{d_{\text{-}k}} u_i^2 = 1$ and the same moment structure up to the 4th order imply that

$$Eu_j^2 = \frac{1}{d_{\text{-}k}}, \quad Eu_j^4 = \frac{1}{d_{\text{-}k}} - (d_{\text{-}k}-1)E(u_i^2u_j^2) = \frac{1}{d_{\text{-}k}^2} + o(d_{\text{-}k}^{-1}) = o(d_{\text{-}k}^{-1}).$$

Hence $E(Q) = E(E(Q|\mathscr{D})) = 1/2$, and

$$\operatorname{var}(Q) = E(\operatorname{var}(Q|\mathscr{D})) = \frac{d_{\text{-}k}}{4}Eu_j^4 + \frac{\operatorname{var}(X_j)}{4(d_{\text{-}k}-1)}(d_{\text{-}k}^2Eu_j^4 - 1) = o(1),$$

since $0 \le X_j \le 8$ means that $\operatorname{var}(X_j) < \infty$. With $E(Q) = 1/2$ and $\operatorname{var}(Q) = o(1)$, we have established (4.15). This completes the proof of the theorem. □

Chapter 5

Factor Strengths Estimation in Time Series Factor Models

5.1 Introduction

Factor modelling has become an increasingly important tool for analyzing high dimensional data across various academic fields, including finance, economics, psychology, and biology. In high dimensional vector or tensor time series, it is generally assumed that a small number of factors drive the dynamics of all variables, leading to significant dimension reduction. Traditional factor models primarily focus on vector time series, exploring various assumptions regarding cross-correlation and serial dependence structures (Bai, 2003; Bai and Ng, 2002, 2007, 2023; Chamberlain and Rothschild, 1983; Fan et al., 2013, 2019; Forni et al., 2000; Lam and Yao, 2012; Lam et al., 2011; Pan and Yao, 2008; Stock and Watson, 2005, 2002). More recently, studies have extended their scope to include matrix factor models (Chen and Fan, 2021; He et al., 2022; Wang et al., 2019; Yu et al., 2022) and tensor factor models (Barigozzi et al., 2023a,b; Chen et al., 2022; Han et al., 2020, 2022), incorporating emerging data in more complex matrix or tensor formats (for further details on the assumptions of factor models under different data structures, please refer to Chapter 2).

In factor modelling, a crucial assumption pertains to the strengths of factors. In the early studies of standard vector factor models (Bai, 2003; Bai and Ng, 2002; Stock and Watson, 2002), it is typically assumed that all r factors are strong, commonly referred to as

pervasive. Specifically, in the model

$$\mathbf{x}_t = \mathbf{A}\mathbf{f}_t + \mathbf{e}_t, \quad t \in [T],$$

where $\mathbf{x}_t \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d\times r}$, the pervasive factor assumption implies that all r eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ diverge proportionally to d, i.e., $\lambda_j(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \asymp d$ for $j \in [r]$. This results in a clear partition of the eigenvalues of the observed covariance matrix into two sets: large eigenvalues representing factor-related variation and small eigenvalues representing idiosyncratic variation. Such a clear partition is also crucial for validating the procedure to estimate the number of factors by analyzing the empirical behaviors of eigenvalues (Ahn and Horenstein, 2013; Bai, 2003; Onatski, 2010).

However, a clear separation of the eigenvalues into one set of large eigenvalues and a second set of small eigenvalues is typically not found in practice. Empirical studies in economics and finance indicate that eigenvalues often diverge at varying rates (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986). In response, models introducing weak factors have been proposed for analyzing vector time series (Bai and Ng, 2023; Freyaldenhoven, 2022; Hallin and Liška, 2011; Lam and Yao, 2012; Lam et al., 2011; Onatski, 2012; Uematsu and Yamagata, 2022). For the j-th column $\mathbf{a}_j$ of $\mathbf{A}$, its factor strength α_j, ranging between 0 and 1, is defined such that

$$\|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}, \quad j \in [r],$$

ensuring that

$$\lambda_j(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \asymp d^{\alpha_j}, \quad j \in [r].$$

Thus, a strong (or pervasive) factor has $\alpha_j = 1$, while a weak factor has $\alpha_j < 1$. Theoretically, a weak factor can result from two scenarios: (i) the factor has a weak effect on some or all observables, or (ii) it affects only a subset of observables, referred to as a "local" factor by Freyaldenhoven (2022).

Building on assumptions about weak factors for vector time series, the literature has developed studies focusing on the estimation of the factor loading space and the number of factors when weak factors are present in the model (Bai and Ng, 2023; Freyaldenhoven, 2022; Lam and Yao, 2012; Lam et al., 2011; Onatski, 2012; Uematsu and Yamagata, 2022). Lam et al. (2011) and Bai and Ng (2023) demonstrate that the estimation accuracy of factor

loadings depends on the strength of the factors. Therefore, it may suffer in the presence of weak factors compared to the classical setting where all factors are pervasive.

In recent years, there has been a growing body of research in matrix and tensor factor modelling. While many studies assume pervasive factors (Barigozzi et al., 2023b; Chen and Fan, 2021; He et al., 2023a, 2022; Yu et al., 2022), matrix and tensor factor models with assumptions about weak factors have also emerged (Chen et al., 2022; Han et al., 2020, 2022), including the model we proposed in Chapter 3. Similar to the vector case, the presence of weak factors can diminish the effectiveness of estimation methods developed for pervasive factors only (Barigozzi et al., 2023b; Chen and Fan, 2021; He et al., 2023a, 2022; Yu et al., 2022). Moreover, the strengths of the factors in tensor factor models could also affect the estimation accuracy of the pre-averaging and iterative projection estimators introduced in Chapter 3, as well as the TIPUP procedure by Chen et al. (2022). Therefore, estimating the factor strengths can help us detect the presence of weak factors, which can be informative for comparing the accuracy of different estimation procedures and assisting in model selection. Additionally, factor strength estimation is crucial for deriving the asymptotic normality of these estimators, enabling inference procedures that are particularly useful for forecasting purposes.

There is limited research on estimating factor strengths. Uematsu and Yamagata (2022) assume sparsity in the factor loading matrix $\mathbf{A}$ and employ techniques akin to adaptive LASSO for factor selection. They calculate the estimated factor strengths by counting the number of nonzero elements in the estimated factor loading matrix. Another study with similar sparsity assumptions, Bailey et al. (2021), proposes estimating factor strengths based on the proportion of statistically significant factor loadings, but it concentrates on cases where factors are observed, while our primary emphasis is on latent factor models.

The sparsity assumptions in these works specifically address scenarios akin to case (ii) mentioned earlier, i.e., when factors are weak due to being "local". In many applications especially when analyzing economic and financial time series, the sparsity assumption provides convenient interpretation of the factors (Freyaldenhoven, 2022, 2023; Uematsu and Yamagata, 2022). On the other hand, if a factor is weak because it has a 'weak' impact on some or all observables, then it may not be detectable by sparsity assumptions. Connor and Korajczyk (2022) consider such a structure with observed factors, and demonstrate its application when modelling U.S. equity return series. Without the sparsity assumption, estimating the factor strengths in such scenarios remains an open problem. Additionally, with emerging literature extending the factor model approach to modeling matrix and tensor time series, no method has been provided so far to estimate factor strengths in

matrix or tensor factor models. Thus, the challenge of estimating factor strengths persists, especially when not relying on sparsity assumptions, and particularly when extending to matrix and tensor factor models.

In this chapter, we propose a novel method to estimate factor strengths in factor models for vector and matrix time series. Our method does not assume the factor loading matrix is sparse. Instead, we make use of covariance information and the estimated factor loading matrices to extract factor strengths directly. To the best of our knowledge, this is the first method to estimate factor strengths that can be applied in general settings when the factor loading matrices are not necessarily sparse. Furthermore, we extend this method to estimate factor strengths in matrix factor models, i.e., tensor factor models with $K = 2$. The factor strengths on the row loading matrices and column loading matrices are estimated with specific identifiability conditions provided. Numerical experiments show that our method performs well in various settings of vector and matrix factor models, shedding light on future research directions in this field.

The rest of this chapter is organized as follows. Section 5.2 discusses the definition and identification of factor strengths in our model. Section 5.3 introduces our propose method for estimating factor strengths based on extracting covariance information on vector factor models, and Section 5.4 extends this approach to matrix factor models. Section 5.5 presents our simulation studies, demonstrating the performance of our method in various settings.

5.2 Definition and Identification of Factor Strengths

The models we consider in this chapter are time series factor models in vector or matrix formats. We start with a vector factor model, which takes the form

$$\mathbf{x}_t = \mathbf{A}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \quad t \in [T], \tag{5.1}$$

where $\mathbf{x}_t \in \mathbb{R}^d$, $\boldsymbol{\varepsilon}_t \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{d\times r}$ is the factor loading matrix, and $\mathbf{f}_t \in \mathbb{R}^r$ are the latent factors.

To discuss the intuition for defining factor strengths, consider $\mathbf{a}_j$ be the j-th column of $\mathbf{A}$. If the j-th (observed or latent) factor f_{tj} has a pervasive effect on all units of $\mathbf{x}_t$, then it means $\mathbf{a}_j$ is dense, and thus $\sum_{i=1}^d a_{ij}^2 \asymp d$, i.e., $\|\mathbf{a}_j\|^2 \asymp d$. This is called a pervasive factor. In the studies of classical high dimensional factor models (see Bai (2003); Bai and Ng (2002); Stock and Watson (2002) for examples), it is typically assumed that all r factors are

pervasive, which implies that $\lambda_j(\mathbf{\Sigma}_x) \asymp \lambda_j(\mathbf{A}\mathbf{\Sigma}_f\mathbf{A}^\top) \asymp d$ for $j \in [r]$ and $\lambda_{r+1}(\mathbf{\Sigma}_x) = O(1)$, where $\mathbf{\Sigma}_f = \mathbb{E}[\mathbf{f}_t\mathbf{f}_t^\top]$ is positive definite, and $\mathbf{\Sigma}_x = \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]$.

However, empirical studies in economics and finance have demonstrated that eigenvalues of observed covariance matrices often diverge at different rates (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986). Furthermore, model with solely pervasive factors implies a large gap between the values of $\lambda_r(\mathbf{\Sigma}_x)$ and $\lambda_{r+1}(\mathbf{\Sigma}_x)$, but it is often missing in financial and economic datasets (Fan et al., 2013). This suggests that the assumption of all pervasive factors may be too restrictive, and some factors may have 'weaker' effects than others. Factor strengths are thus introduced to explain such behavior.

In the context of observed factors, Bailey et al. (2021) defines the strength of a factor as the degree of pervasiveness of its effects. For a weak (non-pervasive) factor, the factor loading is sparse, and the factor strength α_j can be seen as a measure of the number of non-zero factor loadings a_{ij}, such that $(\sum_{i=1}^d a_{ij}^2)^{1/2} = \|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}$. The sparsity assumption on $\mathbf{A}$ is also imposed by many literature dealing with latent factors (Freyaldenhoven, 2022, 2023; Uematsu and Yamagata, 2022), and can be referred to as a 'local' factor. On the other hand, in the context of observed factors, Connor and Korajczyk (2022) consider another scenario where the diverging eigenvalues can be induced by nonsparse factor loadings. In this scenario, a factor can have a weak effect because it affects all the variables at similar strengths, but thinly. For example, if $\mathbf{a}_j$ is not sparse but composed of non-zero values of order $d^{(\alpha_j-1)/2}$, then $\|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}$. Both scenarios result in $\|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}$, which accounts for the empirical observation that eigenvalues diverge at a slower rate than d when $\alpha_j < 1$. With $\lambda_j(\mathbf{\Sigma}_x) \asymp d^{\alpha_j}$, we can also interpret α_j as a measure of the strength of signal which can be observed from the model.

The definition of factor strengths can be similarly applied to the literature on latent factor models. However, when factors are not observed, more assumptions need to be imposed to identify the model. This is because the model (2.3) remains unchanged if we replace the pair $(\mathbf{A};\mathbf{f}_t)$ on the right-hand side with $(\mathbf{AH};\mathbf{H}^{-1}\mathbf{f}_t)$ for any invertible $\mathbf{H}$. In models with only pervasive factors, Bai (2003) and Bai and Ng (2002) propose to impose the normalization of either $\frac{\mathbf{A}^\top\mathbf{A}}{d} = \mathbf{I}_r$ or $\mathbf{E}[\mathbf{f}_t\mathbf{f}_t^\top] = \mathbf{I}_r$ when employing the principal component estimator (see Williams (2019) for example for more discussions on the implications of these strategies). On the other hand, when defining factor strengths, it is more common to assume $\mathbf{E}[\mathbf{f}_t\mathbf{f}_t^\top] = \mathbf{I}_r$ and relax the assumption on $\mathbf{A}$. Intuitively, we can think of the factors as 'primitive' exogenous forces, which do not have common causes, and it is natural to treat them as orthogonal (Bernanke, 1986). In this sense, factor strengths α_j can be defined and interpreted similarly to those of observed factors, as $\|\mathbf{a}_j\|^2 \asymp d^{\alpha_j}$

(Bai and Ng, 2023; Freyaldenhoven, 2022; Lam et al., 2011; Uematsu and Yamagata, 2022).

With the assumption of sparse factor loadings (i.e., local factors), in order to estimate the factor loadings, factor strengths, and the number of factors, Freyaldenhoven (2022) and Uematsu and Yamagata (2022) impose the assumption that $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{D}$, where $\mathbf{D}$ is diagonal, with diagonal elements $d_{jj} \asymp d^{\alpha_j}$. This assumption is more restrictive but necessary to separately identify the factors for estimation purposes. Otherwise, different rotations could mix weak factors with stronger ones, making the factor strengths unidentifiable and impossible to estimate. Freyaldenhoven (2022) show that with local factors, $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{D}$ will be approximately true if the groups of outcomes affected by different factors are sufficiently distinct. In a more recent work, Freyaldenhoven (2023) proposes a rotation criterion that minimizes the l_1-norm to recover the loading if the true loading is sparse.

Similar to the assumptions of Freyaldenhoven (2022) and Uematsu and Yamagata (2022), but without sparsity constraints, we present the following assumptions to identify the factor loadings and factor strengths in our model (2.3).

(V1) (Factor strengths) $\mathbf{A}$ *is of full rank, and* $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{D}$*, where* $\mathbf{D}$ *is a diagonal matrix. Define the diagonal entries of* $\mathbf{D}$ *as* $d_{jj} := (\mathbf{D})_{jj}$*, then* $d_{jj} \asymp d^{\alpha_j}$ *for* $j \in [r]$*, and* $0 < \alpha_r \leq \cdots \leq \alpha_1 \leq 1$.

(V2) (Latent factors) *There is* $\mathbf{z}_{f,t}$ *the same dimension as* $\mathbf{f}_t$*, such that* $\mathbf{f}_t = \sum_{q\geq 0} a_{f,q}\mathbf{z}_{f,t-q}$*. The time series* $\{\mathbf{z}_{f,q}\}$ *has i.i.d. elements with mean 0, variance 1, and* $\mathbb{E}|z_{f,t,i}|^4 \leq c$ *for some* $c < \infty$ *independent of* $t \in [T]$ *and* $i \in [r]$*. The coefficients* $a_{f,q}$ *are so that* $\sum_{q\geq 0} a_{f,q}^2 = 1$.

Assumption (V1) defines factor strengths in the model. As in the earlier discussion, though the concept of factor strengths itself does not require the orthogonal loadings, $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ being diagonal is necessary to identify and estimate a spectrum of different factor strengths (Freyaldenhoven, 2022; Uematsu and Yamagata, 2022). It's also important to note that in Assumption (V1), we do not impose any sparsity assumptions on the factor loading matrix $\mathbf{A}$, in contrast to other recent literature dealing with weak factors (Bailey et al., 2021; Freyaldenhoven, 2022; Uematsu and Yamagata, 2022). Consequently, a factor in our model can be weak if either (i) the factor has a weak effect on some or all observables, or (ii) it affects only a subset of observables. Such a relaxed assumptions provide more flexibility for our approach to be used in practice, if no prior knowledge of the sparse factor loadings is assumed.

Assumption (V2) states that $\mathbf{f}_t$ has uncorrelated elements, which is standard in the litarature as previously discussed. Define $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_T]^{\mathrm{T}}$, then Assumption (V2) implies $\mathbf{E}[\frac{\mathbf{F}^{\mathrm{T}}\mathbf{F}}{T}] = \mathbf{I}_r$ and $\left\|\frac{\mathbf{F}^{\mathrm{T}}\mathbf{F}}{T}\right\| = O_{\mathbb{P}}(1)$, facilitating the rationality of our method as described in Section 5.3.

Remark: While Assumption (V1) relaxes the assumption of weak factors being 'local', this broader framework also means that interpreting the α_j as a measure of sparsity in the loading becomes less straightforward. In general, it is hard to determine whether a weak factor with $\alpha_j < 1$ is caused by local or pervasive scenarios. Therefore, while estimating α_j in Section 5.3 and 5.4 can help us identify the presence of weak factors and approximate their strengths, it may not fully unveil the true structure of the factor loading matrix. Additional techniques are required to uncover such complexities. Consequently, we suggest that our method can be used in conjunction with other existing methods that estimate the sparsity of factor loadings (Freyaldenhoven, 2023; Johnstone and Silverman, 2004; Uematsu and Yamagata, 2022).

For example, based on the estimated factor strengths from our method, we may identify some weak factors and wish to impose sparsity assumptions on some of them. Then, methods for estimating the sparsity of loadings (Freyaldenhoven, 2023; Johnstone and Silverman, 2004; Uematsu and Yamagata, 2022) could help reveal the potentially true sparse structure. Consequently, we can estimate the sparsity level of local factors and quantify their effects. By comparing such effects with our estimated strengths, we might be able to address any unexplained effects not attributed to local factors (i.e. weak effect for some factors on some or all observables). We leave this as an interesting avenue for future research.

5.3 Factor strengths estimation in vector factor models

To introduce our method for estimating factor strengths, let's first focus on the vector factor model (5.1). To estimate factor strengths α_j, $j \in [r]$, note that from Assumption (V1), the factor loading matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{Q}\mathbf{D}^{\frac{1}{2}}$, where $\mathbf{Q} \in \mathbb{R}^{d\times r}$ has orthogonal columns such that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_r$, and $\mathbf{D} \in \mathbb{R}^{r\times r}$ is a diagonal matrix. Since $\mathbf{Q}$ is orthogonal, the information about factor strengths in $\mathbf{A}$ is fully encapsulated in $\mathbf{D}$, given that $d_{jj} \asymp d^{\alpha_j}$. Consequently, we can estimate factor strengths by estimating the diagonal elements of $\mathbf{D}$. To achieve this, we define $\widehat{\mathbf{S}} = \widehat{\mathbf{Q}}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_x\widehat{\mathbf{Q}}$, where $\widehat{\mathbf{Q}}$ is an estimator of $\mathbf{Q}$, and $\widehat{\mathbf{\Sigma}}_x = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}$.

Then

$$\begin{aligned}\widehat{\mathbf{S}} =& \widehat{\mathbf{Q}}^{\mathrm{T}}\mathbf{Q}\mathbf{D}^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t^{\mathrm{T}}\right)\mathbf{D}^{\frac{1}{2}}\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{Q}} \\ &+\widehat{\mathbf{Q}}^{\mathrm{T}}\mathbf{Q}\mathbf{D}^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{e}_t^{\mathrm{T}}\right)\widehat{\mathbf{Q}}+\widehat{\mathbf{Q}}^{\mathrm{T}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{f}_t^{\mathrm{T}}\right)\mathbf{D}^{\frac{1}{2}}\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{Q}}+\widehat{\mathbf{Q}}^{\mathrm{T}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t^{\mathrm{T}}\right)\widehat{\mathbf{Q}}. \quad (5.2)\end{aligned}$$

If the error terms $\boldsymbol{\varepsilon}_t$ are appropriately bounded, with proper assumptions on its cross-correlation and serial dependence, the last three terms in (5.2) become small in comparison to the first term. Moreover, considering that $\mathbf{E}[\mathbf{f}_t\mathbf{f}_t^{\mathrm{T}}] = \mathbf{I}_r$ by Assumption (V2), and assuming we have an estimator $\widehat{\mathbf{Q}}$ that is close to $\mathbf{Q}$, we can make the following approximation:

$$\widehat{\mathbf{S}} \approx \widehat{\mathbf{Q}}^{\mathrm{T}}\mathbf{Q}\mathbf{D}^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t^{\mathrm{T}}\right)\mathbf{D}^{\frac{1}{2}}\mathbf{Q}^{\mathrm{T}}\widehat{\mathbf{Q}} \approx \mathbf{D}.$$

In practice, depending on more specific model assumptions, the estimated $\widehat{\mathbf{Q}}$ can be obtained through various approaches, such as PCA of the sample covariance matrix or sample autocovariance matrix (see Bai (2003); Bai and Li (2012); Bai and Liao (2016); Lam and Yao (2012); Lam et al. (2011) for examples). Now, given that $\mathbf{D}$ is diagonal, we can directly derive the estimator for d_{jj}, $j \in [r]$, by using the diagonal entries of $\widehat{\mathbf{S}}$, such that $\widehat{d}_{jj} := \widehat{s}_{jj}$, where $\widehat{d}_{jj}$ and $\widehat{s}_{jj}$ represent the j-th diagonal entries of $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{S}}$, respectively. Thus, the factor strengths on $\mathbf{A}$ can be estimated as

$$\widehat{\alpha}_j = \frac{\log\left(\widehat{d}_{jj}\right)}{\log(d)}, \quad j \in [r], \quad (5.3)$$

and we can further obtain $\widehat{\mathbf{A}}$ as $\widehat{\mathbf{A}} = \widehat{\mathbf{Q}}\widehat{\mathbf{D}}^{\frac{1}{2}}$, where $\widehat{\mathbf{D}}$ is a diagonal matrix with diagonal entries given by $\widehat{d}_{jj}$.

To assess the estimator $\widehat{\alpha}_j$ obtained by (5.3), note that the true factor strength α_j for the model (5.1) is defined as

$$\left\|\mathbf{a}_j\right\|^2 = Cd^{\alpha_j}, \quad j \in [r], \quad (5.4)$$

where C is a constant that may vary across different j. Additionally, introduce the realized factor strength $\widetilde{\alpha}_j$ as

$$\widetilde{\alpha}_j := \frac{\log(\left\|\mathbf{a}_j\right\|^2)}{\log(d)} = \alpha_j + \frac{\log(C)}{\log(d)}. \quad (5.5)$$

It is important to note that our estimator $\widehat{\alpha}_j$ is, in fact, an estimator for $\widetilde{\alpha}_j$ rather than the true α_j, as C and α_j are not identifiable. However, given that C is a constant, when the dimension d grows large, we expect $\frac{\log(C)}{\log(d)} \to 0$, leading to a negligible difference between $\widetilde{\alpha}_j$ and α_j. In general when C is unknown, $\widetilde{\alpha}_j$ converges to the true α_j with a rate of $\log(d)^{-1}$, and the rate of $|\widehat{\alpha}_j - \alpha_j|$ cannot surpass this bound. However, in the special case where we assume $C = 1$, then $\widetilde{\alpha}_j = \alpha_j$, making the factor strength α_j exactly identifiable. In such case, we can achieve a better rate of convergence for $|\widehat{\alpha}_j - \alpha_j|$ potentially. Thus, in practical situations with finite samples and a moderately sized d, it is desirable for C to be close to 1, ensuring that $\widetilde{\alpha}_j$ does not significantly differ from α_j. In such cases, $\widehat{\alpha}_j$ serves as a reliable approximation to the true α_j.

5.4 Extension to matrix factor models

In Section 5.3, we discuss our method to estimate factor strengths in a vector factor model. The similar approach can be extended to a matrix factor model, which is developed for analyzing time series observations recorded in matrix form (Chen and Fan, 2021; He et al., 2023a; Wang et al., 2019; Yu et al., 2022). Consider the matrix factor model:

$$\mathbf{X}_t = \mathbf{A}_1 \mathbf{F}_t \mathbf{A}_2^{\mathrm{T}} + \mathbf{E}_t, \quad t \in [T], \tag{5.6}$$

where $\mathbf{X}_t \in \mathbf{R}^{d_1 \times d_2}$, $\mathbf{E}_t \in \mathbf{R}^{d_1 \times d_2}$, $\mathbf{F}_t \in \mathbf{R}^{r_1 \times r_2}$, and $\mathbf{A}_k \in \mathbf{R}^{d_k \times r_k}$ for $k = 1, 2$. The following assumptions for matrix factor models are direct extensions of Assumptions (V1) and (V2) for vector factor models:

(M1) (Factor strengths) *For $k = 1, 2$, $\mathbf{A}_k$ is of full rank, and $\mathbf{A}_k^{\mathrm{T}} \mathbf{A}_k = \mathbf{D}_k$, where $\mathbf{D}_k$ is a diagonal matrix. Define the diagonal entries of $\mathbf{D}_k$ as $d_{k,jj} := (\mathbf{D}_k)_{jj}$, then $d_{k,jj} \asymp d^{\alpha_{k,j}}$ for $j \in [r_k]$, and $0 < \alpha_{k,r_k} \leq \cdots \leq \alpha_{k,1} \leq 1$.*

(M2) (Latent factors) *There is $\mathbf{Z}_{f,t}$ the same dimension as $\mathbf{F}_t$, such that $\mathbf{F}_t = \sum_{q \geq 0} a_{f,q} \mathbf{Z}_{f,t-q}$. The time series $\{\mathbf{Z}_{f,q}\}$ has i.i.d. elements with mean 0, variance 1 and $\mathbb{E}|z_{f,t,i,j}|^4 \leq c$ for some $c < \infty$ independent of $t \in [T]$ and $i \in [r_1], j \in [r_2]$. The coefficients $a_{f,q}$ are so that $\sum_{q \geq 0} a_{f,q}^2 = 1$.*

Assumptions (M1) and (M2) are parallel to the tensor factor model assumptions made in Chapter 3.2.1 when $K = 2$, except that we assume $\mathbf{A}_k^{\mathrm{T}} \mathbf{A}_k$ to be diagonal here, in order to separately identify and estimate the factor strengths. Assumption (M2) also implies

$\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t^{\mathrm{T}}]=\mathbf{I}_{r_1}$ and $\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^{\mathrm{T}}\mathbf{F}_t]=\mathbf{I}_{r_2}$. With Assumption (M1), we can write $\mathbf{A}_1=\mathbf{Q}_1\mathbf{D}_1^{\frac{1}{2}}$ and $\mathbf{A}_2=\mathbf{Q}_2\mathbf{D}_2^{\frac{1}{2}}$, where $\mathbf{Q}_k\in\mathbf{R}^{d_k\times r_k}$ has orthogonal columns for $k=1,2$. Then (5.6) can be written as

$$\mathbf{X}_t=\mathbf{Q}_1\mathbf{D}_1^{\frac{1}{2}}\mathbf{F}_t\mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_2^{\mathrm{T}}+\mathbf{E}_t,\ \ t\in[T].$$

To estimate the factor strengths on $\mathbf{A}_1$, similar to the vector case, we can create $\widehat{\mathbf{S}}_1=\widehat{\mathbf{Q}}_1^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{1x}\widehat{\mathbf{Q}}_1$, where $\widehat{\mathbf{Q}}_1$ is an estimator of $\mathbf{Q}_1$, and $\widehat{\mathbf{\Sigma}}_{1x}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}$. Then

$$\begin{aligned}\widehat{\mathbf{S}}_1=&\widehat{\mathbf{Q}}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{D}_1^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{D}_2\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_1^{\mathrm{T}}\widehat{\mathbf{Q}}_1\\&+\widehat{\mathbf{Q}}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{D}_1^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_2^{\mathrm{T}}\mathbf{E}_t^{\mathrm{T}}\right)\widehat{\mathbf{Q}}_1+\widehat{\mathbf{Q}}_1^{\mathrm{T}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}_t\mathbf{Q}_2\mathbf{D}_2^{\frac{1}{2}}\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_1^{\mathrm{T}}\widehat{\mathbf{Q}}_1+\widehat{\mathbf{Q}}_1^{\mathrm{T}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}_t\mathbf{E}_t^{\mathrm{T}}\right)\widehat{\mathbf{Q}}_1.\end{aligned}\tag{5.7}$$

The last three terms in (5.7) will become small compared to the first term if the error terms $\mathbf{E}_t$ are small with proper assumptions on its cross-correlation and serial dependence. For matrix factor models, $\widehat{\mathbf{Q}}_1$ can be obtained using the pre-averaging and iterative projection algorithm developed in Chapter 3. Alternatively, literature has been developed to obtain $\widehat{\mathbf{Q}}_1$ using different approaches under various model assumptions (see Barigozzi et al. (2023b); Chen and Fan (2021); Chen et al. (2022); He et al. (2023a); Wang et al. (2019) for examples). If $\widehat{\mathbf{Q}}_1$ is close to $\mathbf{Q}_1$, then

$$\begin{aligned}\widehat{\mathbf{S}}_1&\approx\widehat{\mathbf{Q}}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{D}_1^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{D}_2\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{D}_1^{\frac{1}{2}}\mathbf{Q}_1^{\mathrm{T}}\widehat{\mathbf{Q}}_1\\&\approx\mathbf{D}_1^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{D}_2\mathbf{F}_t^{\mathrm{T}}\right)\mathbf{D}_1^{\frac{1}{2}}\\&\approx\mathbf{D}_1^{\frac{1}{2}}\mathrm{tr}(\mathbf{D}_2)\mathbf{D}_1^{\frac{1}{2}}\\&=\mathrm{tr}(\mathbf{D}_2)\mathbf{D}_1.\end{aligned}\tag{5.8}$$

If $\mathbf{D}_2$ is known, or we have an estimate for it, we can then estimate the diagonal entries of $\mathbf{D}_1$ by using the diagonal entries of $\widehat{\mathbf{S}}_1/\mathrm{tr}(\mathbf{D}_2)$.

Similarly, to estimate the factor strengths on $\mathbf{A}_2$, if we define $\widehat{\mathbf{S}}_2 = \widehat{\mathbf{Q}}_2^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{2x}\widehat{\mathbf{Q}}_2$, where $\widehat{\mathbf{Q}}_2$ is an estimator of $\mathbf{Q}_2$, and $\widehat{\mathbf{\Sigma}}_{2x} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t$, then by similar arguments,

$$\begin{aligned}
\widehat{\mathbf{S}}_2 &\approx \widehat{\mathbf{Q}}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{D}_2^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^{\mathrm{T}}\mathbf{D}_1\mathbf{F}_t\right)\mathbf{D}_2^{\frac{1}{2}}\mathbf{Q}_2^{\mathrm{T}}\widehat{\mathbf{Q}}_2 \\
&\approx \mathbf{D}_2^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^{\mathrm{T}}\mathbf{D}_1\mathbf{F}_t\right)\mathbf{D}_2^{\frac{1}{2}} \\
&\approx \mathbf{D}_2^{\frac{1}{2}}\mathrm{tr}(\mathbf{D}_1)\mathbf{D}_2^{\frac{1}{2}} \\
&= \mathrm{tr}(\mathbf{D}_1)\mathbf{D}_2. \qquad (5.9)
\end{aligned}$$

Thus, if $\mathbf{D}_1$ is known or if we have an estimate of it, then we can estimate the diagonal entries of $\mathbf{D}_2$ by using the diagonal entries of $\widehat{\mathbf{S}}_2/\mathrm{tr}(\mathbf{D}_1)$.

However, in most practical scenarios, neither $\mathbf{D}_1$ nor $\mathbf{D}_2$ is known. Consequently, we aim to estimate both $\mathbf{D}_1$ and $\mathbf{D}_2$ simultaneously from (5.8) and (5.9). In such situations, it is crucial to note that due to matrix multiplication, the factor strengths on $\mathbf{A}_1$ and $\mathbf{A}_2$ are not identifiable in matrix factor models. This lack of identifiability is reflected in the relationships derived from (5.8) and (5.9):

$$\mathrm{tr}(\mathbf{D}_1)\mathrm{tr}(\mathbf{D}_2) \approx \mathrm{tr}(\widehat{\mathbf{S}}_1). \qquad (5.10)$$

and

$$\mathrm{tr}(\mathbf{D}_1)\mathrm{tr}(\mathbf{D}_2) \approx \mathrm{tr}(\widehat{\mathbf{S}}_2). \qquad (5.11)$$

Therefore, to estimate the factor strengths on $\mathbf{D}_1$ and $\mathbf{D}_2$ simultaneously, it is necessary to define the identifiability condition as:

$$\frac{\mathrm{tr}(\mathbf{D}_1)}{r_1 d_1} = \frac{\mathrm{tr}(\mathbf{D}_2)}{r_2 d_2}. \qquad (5.12)$$

Note that the identifiability condition is not unique. However, we choose to define (5.12) as it is convenient for interpretation. The intuition behind (5.12) is that, in general, larger factor strengths will be "assigned" to larger dimensions. For instance, consider $r_1 = r_2 = 1$ and $\alpha_{1,1} = \alpha_{2,1} = 1$. In this case, $\mathrm{tr}(\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1) = \mathrm{tr}(\mathbf{D}_1) \approx r_1 d_1$ and $\mathrm{tr}(\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2) = \mathrm{tr}(\mathbf{D}_2) \approx r_2 d_2$. Consequently, by (5.12), the estimated factor strengths will recover the true ones, i.e., $\widehat{\alpha}_{1,1} \approx \widehat{\alpha}_{2,1} \approx 1$. On the other hand, if $\mathbf{A}_1$ and $\mathbf{A}_2$ have the exact same dimensions ($r_1 = r_2$

and $d_1 = d_2$), they will be "assigned" the same factor strengths, as the factor strengths on them are completely symmetric and indistinguishable from each other.

With identifiability condition (5.12), together with (5.10) and (5.11), we can allocate the proper factor strengths on $\mathbf{D}_1$ and $\mathbf{D}_2$ accordingly. For more accuracy and consistency in calculation, we can use the average of $\mathrm{tr}(\widehat{\mathbf{S}}_1)$ and $\mathrm{tr}(\widehat{\mathbf{S}}_2)$ as an estimate of $\mathrm{tr}(\mathbf{D}_1)\mathrm{tr}(\mathbf{D}_2)$ and solve for $\mathrm{tr}(\mathbf{D}_1)$ and $\mathrm{tr}(\mathbf{D}_2)$, respectively. This leads to the following approximations:

$$\mathrm{tr}(\mathbf{D}_1) \approx \left(\frac{\mathrm{tr}(\widehat{\mathbf{S}}_1 + \widehat{\mathbf{S}}_2)}{2} \cdot \frac{r_1 d_1}{r_2 d_2} \right)^{\frac{1}{2}}, \tag{5.13}$$

and

$$\mathrm{tr}(\mathbf{D}_2) \approx \left(\frac{\mathrm{tr}(\widehat{\mathbf{S}}_1 + \widehat{\mathbf{S}}_2)}{2} \cdot \frac{r_2 d_2}{r_1 d_1} \right)^{\frac{1}{2}}. \tag{5.14}$$

By substituting (5.14) and (5.13) back into (5.8) and (5.9), we can estimate the diagonal entries of $\mathbf{D}_1$ and $\mathbf{D}_2$ by taking the corresponding diagonal entries in $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$, respectively, normalized to specific magnitudes, which leads to:

$$\widehat{d}_{1,jj} := \frac{\widehat{s}_{1,jj}}{\left(\frac{\mathrm{tr}(\widehat{\mathbf{S}}_1 + \widehat{\mathbf{S}}_2)}{2} \cdot \frac{r_2 d_2}{r_1 d_1} \right)^{\frac{1}{2}}}, \quad j \in [r_1],$$

where $\widehat{d}_{1,jj}$ and $\widehat{s}_{1,jj}$ are the j-th diagonal entries of $\widehat{\mathbf{D}}_1$ and $\widehat{\mathbf{S}}_1$, respectively, and

$$\widehat{d}_{2,jj} := \frac{\widehat{s}_{2,jj}}{\left(\frac{\mathrm{tr}(\widehat{\mathbf{S}}_1 + \widehat{\mathbf{S}}_2)}{2} \cdot \frac{r_1 d_1}{r_2 d_2} \right)^{\frac{1}{2}}}, \quad j \in [r_2],$$

where $\widehat{d}_{2,jj}$ and $\widehat{s}_{2,jj}$ are the j-th diagonal entries of $\widehat{\mathbf{D}}_2$ and $\widehat{\mathbf{S}}_2$, respectively. Finally, the factor strengths on $\mathbf{A}_1$ and $\mathbf{A}_2$ can be estimated as:

$$\widehat{\alpha}_{1,j} = \frac{\log\left(\widehat{d}_{1,jj}\right)}{\log(d_1)}, \quad j \in [r_1],$$

and

$$\widehat{\alpha}_{2,j} = \frac{\log\left(\widehat{d}_{2,jj}\right)}{\log(d_2)}, \quad j \in [r_2],$$

and we can further obtain $\widehat{\mathbf{A}}_k = \widehat{\mathbf{Q}}_k \widehat{\mathbf{D}}_k^{\frac{1}{2}}$, where $\widehat{\mathbf{D}}_k$ is a diagonal matrix with diagonal entries given by $\widehat{d}_{k,jj}$, for $k = 1, 2$.

5.5 Simulation Experiments

In this section, we conduct simulation experiments to test the performances of our proposed method to estimate factor strengths in vector and matrix factor models.

5.5.1 Simulation settings

For generating our data, we use model (5.1) for vector time series ($K = 1$), and (5.6) for matrix time series ($K = 2$), both of which are special cases of the general tensor factor model (2.7). For $k \in [K]$, each factor loading matrix $\mathbf{A}_k$ is generated independently with $\mathbf{A}_k = \mathbf{B}_k \mathbf{D}_k$, where the elements in $\mathbf{B}_k \in \mathbb{R}^{d_k \times r_k}$ are i.i.d. $U(-\sqrt{3}, \sqrt{3})$, and $\mathbf{D}_k \in \mathbb{R}^{r_k \times r_k}$ is diagonal with the j-th diagonal element being $d_k^{-\zeta_{k,j}}$, $0 \leq \zeta_{k,j} \leq 0.5$. Pervasive (strong) factors have $\zeta_{k,j} = 0$, while weak factors have $0 < \zeta_{k,j} \leq 0.5$. In this way, the constant C in (5.4) will be close to 1, so that $\widetilde{\alpha}_{k,j} \approx \alpha_{k,j}$ and $\|\mathbf{a}_{k,j}\| \approx d_k^{\alpha_{k,j}}$ for $j \in [r_k]$, $k \in [K]$.

The elements in $\mathbf{f}_t$ for vector time series (or $\mathbf{F}_t$ for matrix time series) are independent standardized AR(1) with AR coefficients 0.8. The elements in $\mathbf{e}_t$ (or $\mathbf{E}_t$) are generated based on Assumptions (E1)(E2) in Chapter 3.2.1, where the elements in $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$ are independent standardized AR(5) with coefficients (-0.7,-0.3,-0.4,0.2,0.1) and (0.8,0.4, -0.4,0.2,-0.1) respectively. The innovation processes of $\mathscr{F}_t$, $\mathscr{F}_{e,t}$ and $\boldsymbol{\varepsilon}_t$ are all i.i.d. standard normal. The errors $\mathbf{e}_t$ (or $\mathbf{E}_t$) are then normalized based on the signal-to-noise ratio δ, defined as the average ratio of standard errors of $\mathbf{f}_t$ and $\mathbf{e}_t$ (or $\mathbf{F}_t$ and $\mathbf{E}_t$). This normalization ensures that $\frac{1}{d}\sum_{j=1}^{d} \text{var}(\mathbf{e}_{t,j}) = \frac{1}{\delta^2}$ (or $\frac{1}{d_1 d_2}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \text{var}(\mathbf{e}_{t,i,j}) = \frac{1}{\delta^2}$). We assume $\delta = 2$ for all simulation experiments in this chapter.

We set $r_k = 2$ for all k, and consider two settings of factor strengths, detailed below:

(I) One strong factor and one weak factor with $\zeta_{k,1} = 0$ and $\zeta_{k,2} = 0.2$ for all k, so that $\alpha_{k,1} = 1$, $\alpha_{k,2} = 0.6$.

(II) Two weak factors with different strengths, with $\zeta_{k,1} = 0.1$ and $\zeta_{k,2} = 0.2$ for all k, so that $\alpha_{k,1} = 0.8$, $\alpha_{k,2} = 0.6$.

Note that in the simulation settings, we generate $\mathbf{A}_k$ in a way such that the constant C in (5.4) will be close to 1. As discussed previously, the rate of $|\widehat{\alpha}_j - \alpha_j|$ cannot be

faster than $\log(d)^{-1}$ if C is not close to 1. Hence, we set C to be close to 1 to ensure $\widetilde{\alpha}_j \approx \alpha_j$, which facilitates the potential observation of a faster rate of $|\widehat{\alpha}_j - \alpha_j|$ in finite samples. Additionally, assuming C is close to 1, we exclude the setting where $\alpha_{k,1} = \alpha_{k,2}$ (i.e., two factors in $\mathbf{A}_k$ having exactly the same strength). This exclusion is based on the fact that when $\alpha_{k,1} = \alpha_{k,2}$ and $C = 1$, it implies that the two eigenvalues of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k$, and thus the two population eigenvalues, become nearly identical. However, such behavior is not supported by empirical observations in economics and finance (Freyaldenhoven, 2022; Ross, 1976; Trzcinka, 1986), as population eigenvalues usually diverge at different rates. In practice, observing two very close eigenvalues is rare, since factor loading space is of finite dimension with $r_k \ll d_k$. In real data scenarios, it's more realistic to consider factor loading entries as random. With random entries, the probability of having very close eigenvalues is significantly small, especially given that the dimension d_k is much larger than the number of factors r_k. Finally, in the literature dealing with factor strengths, it is also typically necessary in theory to assume that the eigenvalues of $\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k$ are distinct (see Freyaldenhoven (2022); Lam et al. (2011); Uematsu and Yamagata (2022) for examples, and Assumption (L1') in Chapter 3.2.1 as well), which facilitates the use of perturbation theory for estimating a well-defined eigenvector estimator. In fact, when two eigenvalues are very close, an estimated eigenvector can be vastly different from a particular target. Therefore, we only consider distinct factor strengths when assuming $C \approx 1$ in our simulations.

5.5.2 Results

For vector factor models ($K = 1$), we consider all combinations of dimensions $d = 50, 100, 200, 400, 800$ and $T = 50, 100, 200, 400, 800$ for each of the two settings outlined in Section 5.5.1. We estimate $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ following the process described in Section 5.3, where $\widehat{\mathbf{Q}}$ is estimated using PCA of the sample covariance matrix (Bai, 2003). Tables 5.1 and 5.2 record the mean and standard deviation over 100 repetitions of factor strengths estimations under different settings and dimensions.

Based on the results presented in Tables 5.1 and 5.2, our factor strengths estimators demonstrate good performance across all settings in vector factor models. Both $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ converge to the true factor strengths α_1 and α_2, with a particularly notable improvement as T increases. Furthermore, the standard deviation of the estimators decreases with the increase in T or d. It's essential to note that the standard deviation of estimation is influenced not only by errors in the estimation procedure but also by the fact that $\widetilde{\alpha}_j$ is not

generated to be exactly α_j but with some small variance (i.e., the constant C in (5.4) is not exactly 1). Nevertheless, given that $\widetilde{\alpha}_j \approx \alpha_j$, the estimated $\widehat{\alpha}_j$ still serves as a good approximation of α_j.

	d	$T=50$	$T=100$	$T=200$	$T=400$	$T=800$
$\widehat{\alpha}_1$	50	1.00(0.10)	0.99(0.07)	1.00(0.06)	1.00(0.05)	1.00(0.04)
	100	0.99(0.09)	0.98(0.06)	0.98(0.05)	1.00(0.04)	1.00(0.03)
	200	0.99(0.08)	0.98(0.06)	0.99(0.04)	1.00(0.03)	1.00(0.02)
	400	0.98(0.07)	0.99(0.04)	1.00(0.04)	1.00(0.03)	1.00(0.02)
	800	0.98(0.07)	1.00(0.04)	0.99(0.03)	1.00(0.02)	1.00(0.02)
$\widehat{\alpha}_2$	50	0.56(0.08)	0.59(0.08)	0.59(0.06)	0.59(0.05)	0.59(0.05)
	100	0.58(0.07)	0.59(0.07)	0.59(0.05)	0.59(0.04)	0.60(0.03)
	200	0.59(0.06)	0.60(0.05)	0.60(0.04)	0.60(0.03)	0.60(0.02)
	400	0.62(0.05)	0.59(0.05)	0.60(0.04)	0.60(0.03)	0.60(0.02)
	800	0.64(0.04)	0.61(0.04)	0.60(0.03)	0.60(0.02)	0.60(0.02)

Table 5.1 The mean and standard deviation (in brackets) of the estimated factor strengths for Setting (I) under vector factor models. The true factor strengths are $\alpha_1 = 1$, $\alpha_2 = 0.6$.

	d	$T=50$	$T=100$	$T=200$	$T=400$	$T=800$
$\widehat{\alpha}_1$	50	0.82(0.09)	0.81(0.08)	0.82(0.05)	0.81(0.05)	0.81(0.04)
	100	0.80(0.09)	0.80(0.07)	0.80(0.05)	0.80(0.03)	0.80(0.03)
	200	0.79(0.07)	0.79(0.05)	0.80(0.04)	0.81(0.03)	0.80(0.03)
	400	0.80(0.07)	0.80(0.05)	0.80(0.04)	0.80(0.02)	0.80(0.02)
	800	0.81(0.05)	0.80(0.05)	0.80(0.03)	0.80(0.03)	0.80(0.02)
$\widehat{\alpha}_2$	50	0.56(0.09)	0.56(0.08)	0.59(0.06)	0.60(0.05)	0.59(0.05)
	100	0.57(0.07)	0.59(0.05)	0.59(0.05)	0.59(0.03)	0.59(0.03)
	200	0.59(0.05)	0.59(0.05)	0.60(0.04)	0.59(0.03)	0.60(0.02)
	400	0.61(0.05)	0.60(0.05)	0.60(0.03)	0.60(0.03)	0.60(0.02)
	800	0.63(0.03)	0.61(0.04)	0.60(0.03)	0.60(0.02)	0.60(0.02)

Table 5.2 The mean and standard deviation (in brackets) of the estimated factor strengths for Setting (II) under vector factor models. The true factor strengths are $\alpha_1 = 0.8$, $\alpha_2 = 0.6$.

For matrix factor models ($K = 2$), we consider the following five settings of different dimensions for d_1 and d_2:

i. $d_1 = d_2 = 25$;

ii . $d_1 = d_2 = 50$;

iii . $d_1 = d_2 = 100$;

iv . $d_1 = 25, \quad d_2 = 50$;

v . $d_1 = 50, \quad d_2 = 100$.

We consider all combinations of $T = 50, 100, 200, 400, 800$, and the above five settings of dimensions for each of the two settings outlined in Section 5.5.1. We estimate $\widehat{\alpha}_{1,1}$, $\widehat{\alpha}_{1,2}$, $\widehat{\alpha}_{2,1}$, and $\widehat{\alpha}_{2,2}$ following the process described in Section 5.4, where $\widehat{\mathbf{Q}}_1$ and $\widehat{\mathbf{Q}}_2$ are estimated using the pre-averaging and iterative projection algorithm developed in Chapter 3. Tables 5.3 and 5.4 record the mean and standard deviation over 100 repetitions of factor strengths estimations under different settings and dimensions.

From Tables 5.3 and 5.4, our estimation procedure performs effectively across all settings in matrix factor models. The identifiability condition (5.12) efficiently allocates factor strengths between $\mathbf{A}_1$ and $\mathbf{A}_2$. When $d_1 = d_2$, we estimate relatively similar factor strengths for $\mathbf{A}_1$ and $\mathbf{A}_2$ since they are indistinguishable. Moreover, all estimated factor strengths converge to the true values as T and d increase. In cases where $d_1 \neq d_2$, the estimated factor strengths on $\mathbf{A}_1$ and $\mathbf{A}_2$ are allocated based on the relative magnitudes of d_1 and d_2, contributing to the recovery of true factor strengths. This tendency is particularly pronounced in Setting (I), where the strongest factors on $\mathbf{A}_1$ and $\mathbf{A}_2$ are pervasive.

	(d_1, d_2)	$T = 50$	$T = 100$	$T = 200$	$T = 400$	$T = 800$
$\widehat{\alpha}_{1,1}$	$(25, 25)$	1.00(0.07)	1.00(0.06)	0.99(0.05)	0.99(0.05)	1.00(0.04)
	$(50, 50)$	0.99(0.06)	1.00(0.05)	1.00(0.04)	1.00(0.03)	1.00(0.03)
	$(100, 100)$	0.99(0.04)	0.99(0.03)	1.00(0.02)	1.00(0.02)	1.00(0.02)
	$(25, 50)$	1.01(0.08)	0.99(0.06)	0.99(0.05)	0.99(0.04)	0.99(0.04)
	$(50, 100)$	0.98(0.06)	1.00(0.04)	0.99(0.04)	0.99(0.03)	0.99(0.02)
$\widehat{\alpha}_{1,2}$	$(25, 25)$	0.55(0.11)	0.57(0.09)	0.58(0.09)	0.58(0.06)	0.59(0.06)
	$(50, 50)$	0.57(0.08)	0.58(0.07)	0.59(0.07)	0.59(0.05)	0.60(0.04)
	$(100, 100)$	0.57(0.07)	0.59(0.06)	0.59(0.04)	0.59(0.04)	0.59(0.03)
	$(25, 50)$	0.54(0.11)	0.55(0.10)	0.57(0.08)	0.57(0.06)	0.57(0.06)
	$(50, 100)$	0.57(0.09)	0.56(0.07)	0.59(0.06)	0.59(0.05)	0.59(0.04)
$\widehat{\alpha}_{2,1}$	$(25, 25)$	1.00(0.07)	1.00(0.05)	0.99(0.05)	0.99(0.05)	1.00(0.04)
	$(50, 50)$	0.99(0.06)	0.99(0.05)	1.00(0.04)	1.00(0.03)	1.00(0.03)
	$(100, 100)$	0.99(0.04)	0.99(0.03)	1.00(0.02)	1.00(0.02)	1.00(0.02)
	$(25, 50)$	1.01(0.06)	1.00(0.04)	1.00(0.04)	1.00(0.03)	1.00(0.03)
	$(50, 100)$	0.99(0.05)	1.01(0.03)	1.00(0.03)	1.00(0.02)	1.00(0.02)
$\widehat{\alpha}_{2,2}$	$(25, 25)$	0.55(0.12)	0.56(0.09)	0.58(0.08)	0.58(0.06)	0.58(0.06)
	$(50, 50)$	0.59(0.09)	0.59(0.07)	0.58(0.06)	0.59(0.04)	0.59(0.04)
	$(100, 100)$	0.58(0.08)	0.58(0.06)	0.59(0.04)	0.59(0.04)	0.59(0.03)
	$(25, 50)$	0.58(0.10)	0.59(0.07)	0.59(0.06)	0.59(0.04)	0.60(0.05)
	$(50, 100)$	0.59(0.08)	0.58(0.06)	0.59(0.05)	0.60(0.04)	0.60(0.03)

Table 5.3 The mean and standard deviation (in brackets) of the estimated factor strengths for Setting (I) under matrix factor models. The true factor strengths are $\alpha_{1,1} = \alpha_{2,1} = 1$, $\alpha_{1,2} = \alpha_{2,2} = 0.6$.

	(d_1,d_2)	$T=50$	$T=100$	$T=200$	$T=400$	$T=800$
$\widehat{\alpha}_{1,1}$	$(25,25)$	0.82(0.07)	0.81(0.05)	0.80(0.05)	0.81(0.05)	0.81(0.05)
	$(50,50)$	0.80(0.06)	0.81(0.04)	0.80(0.03)	0.80(0.03)	0.81(0.03)
	$(100,100)$	0.79(0.05)	0.80(0.03)	0.80(0.02)	0.80(0.02)	0.80(0.02)
	$(25,50)$	0.79(0.07)	0.79(0.05)	0.78(0.05)	0.78(0.05)	0.78(0.04)
	$(50,100)$	0.78(0.06)	0.78(0.05)	0.78(0.03)	0.78(0.03)	0.78(0.02)
$\widehat{\alpha}_{1,2}$	$(25,25)$	0.56(0.10)	0.55(0.08)	0.58(0.06)	0.58(0.06)	0.58(0.06)
	$(50,50)$	0.57(0.06)	0.58(0.06)	0.58(0.05)	0.59(0.04)	0.59(0.04)
	$(100,100)$	0.58(0.08)	0.60(0.05)	0.59(0.03)	0.60(0.03)	0.59(0.02)
	$(25,50)$	0.53(0.10)	0.54(0.08)	0.55(0.07)	0.55(0.06)	0.55(0.06)
	$(50,100)$	0.56(0.08)	0.56(0.06)	0.57(0.04)	0.57(0.04)	0.57(0.04)
$\widehat{\alpha}_{2,1}$	$(25,25)$	0.83(0.07)	0.80(0.06)	0.81(0.05)	0.81(0.04)	0.81(0.05)
	$(50,50)$	0.81(0.05)	0.81(0.04)	0.80(0.04)	0.81(0.03)	0.81(0.03)
	$(100,100)$	0.79(0.04)	0.80(0.03)	0.80(0.03)	0.80(0.02)	0.80(0.02)
	$(25,50)$	0.83(0.06)	0.83(0.04)	0.81(0.04)	0.82(0.03)	0.82(0.03)
	$(50,100)$	0.82(0.04)	0.82(0.04)	0.82(0.03)	0.82(0.02)	0.82(0.02)
$\widehat{\alpha}_{2,2}$	$(25,25)$	0.55(0.09)	0.57(0.08)	0.57(0.06)	0.57(0.06)	0.58(0.06)
	$(50,50)$	0.57(0.06)	0.57(0.06)	0.59(0.05)	0.59(0.04)	0.59(0.04)
	$(100,100)$	0.60(0.06)	0.59(0.05)	0.60(0.04)	0.60(0.03)	0.59(0.02)
	$(25,50)$	0.58(0.08)	0.60(0.06)	0.62(0.05)	0.61(0.04)	0.61(0.04)
	$(50,100)$	0.60(0.06)	0.61(0.05)	0.61(0.03)	0.62(0.03)	0.62(0.02)

Table 5.4 The mean and standard deviation (in brackets) of the estimated factor strengths for Setting (II) under matrix factor models. The true factor strengths are $\alpha_{1,1}=\alpha_{2,1}=0.8$, $\alpha_{1,2}=\alpha_{2,2}=0.6$.

Chapter 6

A New Form of Consistency for Large Covariance Matrix Estimators

6.1 Introduction

The estimation of the covariance matrix $\mathbf{\Sigma}_p$ and its inverse $\mathbf{\Omega}_p = \mathbf{\Sigma}_p^{-1}$, referred to as the precision matrix, plays a crucial role in statistical analysis across various fields. Examples include risk estimation and portfolio allocation in finance, multiple hypotheses testing in general statistical analysis, graphical modelling and clustering for gene discovery in bioinformatics, and factor analysis in economics. With the increase in computational power and significant technological developments, obtaining relatively large datasets has become easier than ever. Many of these datasets are high dimensional, where the number of variables p is comparable to or even larger than the sample size n. This poses challenges for traditional covariance matrix estimators, such as the sample covariance matrix.

A significant issue arises from the fact that the sample covariance matrix is ill-conditioned in high dimensional settings, in the sense that its eigenvalues are more extreme than their population counterparts. In random matrix theory, it can be shown that when $\mathbf{\Sigma}_p = \mathbf{I}_p$ and $p/n \to c \in (0, \infty)$, the distribution of the eigenvalues of the sample covariance matrix, called the empirical spectral density, does not converge to a single mass at 1. Instead, it converges to a markedly different distribution known as the Marchenko-Pastur distribution (Marčenko and Pastur, 1967). Furthermore, the eigenvectors of the sample covariance matrix can differ significantly from those of $\mathbf{\Sigma}_p$ (Johnstone and Lu, 2009; Ledoit and Péché, 2011). Bai and Yin (1993) and Bai and Silverstein (2010) further demonstrated that when $\mathbf{\Sigma}_p = \mathbf{I}_p$ and $p/n \to c > 0$, the smallest and largest eigenvalues of the sample

covariance matrix converge to $\max(0,(1-\sqrt{c})^2)$ and $(1+\sqrt{c})^2)$, respectively. Moreover, when $p > n$, the sample covariance matrix is not invertible, preventing the direct estimation of $\mathbf{\Omega}_p = \mathbf{\Sigma}_p^{-1}$.

To overcome the above mentioned problems, various regularization methods have been studied to estimate the covariance or precision matrices in different applications. Two major branches of regularization methods have been proposed. The first branch assumes special structures on the covariance matrix $\mathbf{\Sigma}_p$ or the precision matrix $\mathbf{\Omega}_p$. Example methods include thresholding (Bickel and Levina, 2008b; Cai and Liu, 2011), banding (Bickel and Levina, 2008a), tapering (Cai et al., 2010; Furrer et al., 2006), penalization (Huang et al., 2006; Lam and Fan, 2009; Ravikumar et al., 2011; Rothman et al., 2008), factor modelling (Fan et al., 2008, 2013; Guo et al., 2017), modified cholesky decomposition (Pan and Mackenzie, 2003; Pourahmadi, 2007; Rothman et al., 2010), Lasso (Peng et al., 2009), adaptive Lasso (Kock and Callot, 2015) and graphical Lasso (Friedman et al., 2008; Mazumder and Hastie, 2012), among others. In general, structural assumptions, such as sparsity, bandedness or a factor structure, are needed for consistent estimation.

Despite the effectiveness of the above mentioned methods under their assumed structures, the estimators may deviate considerably from the true matrix when prior information is inaccurate. In practical scenarios, determining the true covariance matrix structure beforehand is typically unfeasible, introducing the challenge of selecting the appropriate matrix structure or estimation method.

Hence, the second branch of regularization methods concentrates on shrinking the eigenvalues of sample covariance matrices without assuming specific structures on the true covariance or precision matrices. The concept of shrinkage estimators for covariance matrices was initially introduced by Stein (1975, 1986). Modern developments began with Ledoit and Wolf (2004), who proposed a linear shrinkage estimator that shrinks the eigenvalues of the sample covariance matrix towards the identity matrix. Schäfer and Strimmer (2005) extended this idea to shrink towards different target matrices. Moreover, Won et al. (2013) introduced a condition-number regularized estimator, maintaining the middle portion of the sample eigenvalues while winsorizing the more extreme eigenvalues at specific constants. Additionally, Ledoit and Wolf (2012) suggested a nonlinear shrinkage estimator, and Abadir et al. (2014) proposed a model-free regularized estimator using a data splitting scheme. Recently, Lam (2016) introduced nonparametric eigenvalue–regularized covariance matrix estimator (NERCOME) through data splitting, providing a theoretically supported data splitting scheme for asymptotic efficiency. Lam and Feng (2018) also

applied similar ideas to derive a nonparametrically eigenvalue-regularized integrated volatility matrix estimator (NERIVE) for high-frequency data.

While these eigenvalue-regularized estimators demonstrate strong empirical performance in practice, the consistency of these estimators has not been well studied thus far. Though other theoretical properties of these estimators, such as the optimality under expected Frobenius loss (Lam, 2016; Ledoit and Wolf, 2004, 2012) or expected quadratic loss (Schäfer and Strimmer, 2005) have been shown, the proof of the widely used spectral norm consistency usually relies on structural assumptions; thus, it is challenging to establish for non-structured estimators. The study of the convergence properties of these estimators remains a challenge.

Recently, Xu et al. (2015) introduced a new matrix convergence criterion called *normalized consistency*. It is closely related to, but different from, the commonly used spectral norm convergence. Let $\widehat{\mathbf{\Sigma}}_p$ be any estimator of $\mathbf{\Sigma}_p$. We recall the classical definition of spectral norm consistency, which is

$$\rho(\widehat{\mathbf{\Sigma}}_p - \mathbf{\Sigma}_p) = o_{\mathbb{P}}(1), \tag{6.1}$$

where $\rho(\mathbf{B})$ is the spectral radius of a matrix $\mathbf{B}$, which coincides with the operator norm $\|\mathbf{B}\|$ when $\mathbf{B}$ is symmetric positive definite. Meanwhile, let $f = \left\|\mathbf{\Sigma}_p\right\|_F$ and $\widehat{f} = \left\|\widehat{\mathbf{\Sigma}}_p\right\|_F$. Then we say that the estimate $\widehat{\mathbf{\Sigma}}_p$ of $\mathbf{\Sigma}_p$ is normalized consistent if

$$\rho(\widehat{\mathbf{\Sigma}}_p/\widehat{f} - \mathbf{\Sigma}_p/f) = o_{\mathbb{P}}(1). \tag{6.2}$$

Xu et al. (2015) introduced normalized consistency to assist inferences and hypothesis testing for high dimensional data. In general, normalized consistency does not imply spectral norm consistency and vice versa. The relationship between these two types of convergences was discussed briefly in Xu et al. (2015). Xu et al. (2015) also showed that the sample covariance matrix, which does not exploit any structural assumptions, can be normalized consistent under certain conditions. The conditions required are closely related to the dependence structure of $\mathbf{\Sigma}_p$. For example, if $cov(y_i) = \mathbf{\Sigma}_p$ and the entries of y_i are strongly dependent in the sense that $\left\|\mathbf{\Sigma}_p\right\|_F \asymp p$ (which can occur in data from a factor model, for example), then the normalized consistency of the sample covariance matrix is satisfied (see Theorem 3.3 from Xu et al. (2015) for details, also Forni et al. (2009) for its implications for factor models). However, its rate of convergence is still unclear. Also, it remains an open and interesting question as to whether other state-of-the-art covariance matrix estimators are normalized consistent.

Inspired by the research gaps discussed above, it is natural to consider the question of whether normalized consistency can be explored for the class of non-structured covariance matrix estimators such as NERCOME (Lam, 2016). Since traditional spectral norm consistency remains difficult to study for these estimators, if new forms of consistency can be established, they will provide more solid theoretical guarantees that can be applied to these estimators. In this chapter, we prove normalized consistency for NERCOME and the corresponding precision matrix estimator under some mild conditions. Applications for the normalized consistency of NERCOME are also investigated. However, it should also be noted that the definition of normalized consistency in (6.2) is intended to normalize the effect of high dimensionality (which can only be applied when $p \to \infty$), which means that it does not offer the same desirable properties and applications as spectral norm consistency or other consistency measure. As a result, normalized consistency is currently only applicable in specific scenarios, such as aiding in high dimensional hypothesis testing where Central Limit Theorems may not necessarily apply (Xu et al., 2015).

The rest of this chapter is organized as follows. Section 6.2 briefly reviews the estimation procedures and properties of NERCOME as an eigenvalue-regularized covariance matrix estimator. Section 6.3 provides theoretical results for proving normalized consistency for NERCOME and its precision matrix estimator. Section Section 6.4 presents simulation experiments to demonstrate the empirical evidence of normalized consistency for NERCOME. Additionally, it illustrates the performance of NERCOME in a high dimensional hypothesis test, showcasing an application of normalized consistency.

6.2 A Brief Review of NERCOME

In this section, we briefly review the estimation principles and properties of NERCOME as proposed by Lam (2016). Let $\mathbf{y}_i \in \mathbb{R}^p$, $i \in [n]$ be an independent and identically distributed (i.i.d) sample with mean 0 and the covariance matrix $\mathbf{\Sigma}_p = \mathbb{E}(\mathbf{y}_i \mathbf{y}_i^{\mathrm{T}})$. We assume that $p = p_n$ and $p/n \to c > 0$ as $n \to \infty$. Let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$, and the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$.

Lam (2016) proposed to utilise the sample splitting idea as introduced by Abadir et al. (2014), who split the data into two parts, say $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, where $\mathbf{Y}_1$ has size $p \times n_1$ and $\mathbf{Y}_2$ has size $p \times n_2$ with $n_1 + n_2 = n$. Define the sample covariance for $\mathbf{Y}_i$ as

$$\widetilde{\mathbf{\Sigma}}_i = \frac{1}{n_i}\mathbf{Y}_i \mathbf{Y}_i^{\mathrm{T}}, \quad i = 1, 2.$$

Let $m = n_1$ be the split location, and $\widetilde{\mathbf{\Sigma}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$ be the eigenvalue decomposition of $\widetilde{\mathbf{\Sigma}}_1$, where $\mathbf{P}_1$ is the matrix of eigenvectors of $\mathbf{Y}_1$. Then NERCOME can be calculated as

$$\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1 \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}, \tag{6.3}$$

with the corresponding precision matrix estimator

$$\widehat{\mathbf{\Omega}}_m = \widehat{\mathbf{\Sigma}}_m^{-1}. \tag{6.4}$$

To understand the idea behind NERCOME, we can consider a class of rotation-equivalent estimators $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{P}$ is the matrix of eigenvectors for the sample covariance matrix $\mathbf{S}$, and $\mathbf{D}$ is a diagonal matrix. For the optimization problem

$$\min_{\mathbf{D}} \left\| \mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_p \right\|_F, \tag{6.5}$$

Ledoit and Péché (2011) showed that the optimal solution is given by $\mathbf{D} = \mathrm{diag}(d_1, \cdots, d_p)$, where $d_i = \mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$, and $\mathbf{p}_i$ is the i-th column of $\mathbf{P}$. Thus, the "ideal" estimator, also called the finite-sample optimal estimator by Ledoit and Wolf (2012), can be defined as

$$\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal}} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{P})\mathbf{P}^{\mathrm{T}}. \tag{6.6}$$

It may be tempting to estimate the ideal estimator by directly using $\mathbf{P}$ and $\mathbf{S}$ (as an estimate of $\mathbf{\Sigma}_p$) from the whole sample set. However, this may lead to poor performance in practice, since $\mathbf{D}$ should not reuse the data that has already been used for calculating $\mathbf{P}$, as they worsen the estimate of $\mathbf{D}$. That is why we need the splitting of the whole data set and use independent observations $\mathbf{Y}_1$ and $\mathbf{Y}_2$ to regularize the eigenvalues.

If we replace $\mathbf{P}$ by $\mathbf{P}_1$ in (6.5), then similarly, the optimal solution is given by $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$. Lam (2016) actually showed that $\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ is asymptotically the same as $d_i = \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}$, and proved that $\widehat{\mathbf{\Sigma}}_m$ converge to $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal},1} = \mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$, the ideal estimator with $\mathbf{P}_1$ replacing $\mathbf{P}$. Moreover, to assess the quality of estimator, Lam (2016) defined the efficiency loss of an estimator $\widehat{\mathbf{\Sigma}}$ as

$$EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = 1 - \frac{L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\mathrm{Ideal}})}{L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})}, \tag{6.7}$$

where $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}})$ is a loss function for estimating $\mathbf{\Sigma}_p$ using $\widehat{\mathbf{\Sigma}}$. If $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) \leq 0$, it means that the estimator $\widehat{\mathbf{\Sigma}}$ is performing at least as good as the ideal estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal}}$ in terms

of the loss function L. They further showed that $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_m) \xrightarrow{a.s.} 0$ with respect to the Frobenius loss when $m/n \to 1$ and $n - m \to \infty$, so that $\widehat{\mathbf{\Sigma}}_m$ is asymptotically as efficient as the ideal estimator (6.6) in estimating $\mathbf{\Sigma}_p$.

While using $\mathbf{P}_1$ as the matrix of eigenvectors does not fully utilize all the data, as opposed to using $\mathbf{P}$, the performance of $\widehat{\mathbf{\Sigma}}_m$ can be improved by defining an averaged estimator that averages over all $\widehat{\mathbf{\Sigma}}_m$'s calculated with different choices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Recall that each vector $\mathbf{y}_i$ in $\mathbf{Y}$ is independent of each other and identically distributed. We can permute the data and form a different data matrix $\mathbf{Y}^{(j)}$, with the data split into $\mathbf{Y}_1^{(j)}$ and $\mathbf{Y}_2^{(j)}$ accordingly. For the j-th permutation, let $\mathbf{P}_{1j}$ and $\widetilde{\mathbf{\Sigma}}_2^{(j)}$ be defined similarly to $\mathbf{P}_1$ and $\widetilde{\mathbf{\Sigma}}_2$. Then, the corresponding covariance matrix estimator is given by

$$\widehat{\mathbf{\Sigma}}_m^{(j)} = \mathbf{P}_{1j}\text{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}.$$

Suppose we have M permutations of the data, then the final averaged estimator can be calculated as

$$\widehat{\mathbf{\Sigma}} = \frac{1}{M}\sum_{j=1}^{M}\widehat{\mathbf{\Sigma}}_m^{(j)}. \tag{6.8}$$

Lam (2016) showed that $\widehat{\mathbf{\Sigma}}$ is also asymptotically as efficient as $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ in estimating $\mathbf{\Sigma}_p$ with respect to the Frobenius loss, with empirically good performance demonstrated in a variety of settings.

For the precision matrix estimator $\widehat{\mathbf{\Omega}}_m$ as defined in (6.4), Lam (2016) showed that it is asymptotically optimal with respect to minimizing the inverse Stein's loss function (James and Stein, 1961; Ledoit and Wolf, 2013) for estimating $\mathbf{\Omega}_p$. After the averaging procedure to obtain $\widehat{\mathbf{\Sigma}}$ in (6.8), the corresponding precision matrix estimator can be defined as $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Sigma}}^{-1}$, and the asymptotic efficiency still holds for $\widehat{\mathbf{\Omega}}$.

Finally, Lam (2016) also showed that some asymptotic results for NERCOME still holds even when the data follows a factor model. Thus, we do not need to explicitly estimate the factor loading matrix, the number of factors, and the unknown factor series, if factor analysis is not the final goal and estimating the covariance or precision matrix is just an intermediate step. In such sense, NERCOME can be robust to changes in the structure of the data. For more details on the theoretical properties and empirical performances of NERCOME under various model assumptions, please refer to Lam (2016).

6.3 Normalized Consistency of NERCOME

To provide theoretical guarantees for NERCOME, Lam (2016) demonstrated its asymptotic efficiency compared to $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ in estimating $\mathbf{\Sigma}_p$ with respect to certain loss functions. However, the efficiency loss defined in (6.7) is still not an ideal measure: it only compares an estimator $\widehat{\mathbf{\Sigma}}$ with the ideal estimator $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ as defined in (6.6), which is restricted to be a rotation-equivalent estimator. The direct distance between $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}_p$, usually measured in terms of matrix norms, is still unknown. The proof of the traditional spectral norm consistency of covariance matrix estimators, as defined in (6.1), relies on structural assumptions, which are difficult to apply for NERCOME, a non-structured estimator. In this section, we establish a new type of matrix convergence criterion called normalized consistency, as defined in (6.2) (Xu et al., 2015), for NERCOME and the corresponding precision matrix estimator. The establishment of normalized consistency provides NERCOME with more theoretical guarantees for application in certain contexts.

We recall some general assumptions of NERCOME (Lam, 2016) as follows:

(A1) *Each observation can be written as* $\mathbf{y}_i = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_i$ *for* $i \in [n]$*, where each* $\mathbf{z}_i$ *is a* $p \times 1$ *vector of independent and identically distributed random variables* z_{ij}*. Each* z_{ij} *has mean 0 and unit variance, and* $\mathbb{E}|z_{ij}|^k \leq B < \infty$ *for some constant* B *and* $2 < k \leq 20$*.*

(A2) *The population covariance matrix* $\mathbf{\Sigma}_p$ *is non-random and of size* $p \times p$*. Furthermore,* $\left\|\mathbf{\Sigma}_p\right\| = O(p^{1/2})$*.*

(A3) *Let* $\tau_{n,1} \geq \cdots \geq \tau_{n,p}$ *be the* p *eigenvalues of* $\mathbf{\Sigma}_p$*, with corresponding eigenvectors* $\mathbf{v}_{n,1}, \cdots, \mathbf{v}_{n,p}$*. Define* $H_n(\tau) = p^{-1}\sum_{i=1}^{p} \mathbf{1}_{\tau_{n,i} \leq \tau}$ *the empirical distribution function (e.d.f.) of the population eigenvalues, where* $\mathbf{1}_A$ *is the indicator function of the set* A*. We assume* $H_n(\tau)$ *converges to some non-random limit* H *at every point of continuity of* H*.*

(A4) *The support of* H *defined above is the union of a finite number of compact intervals bounded away from zero and infinity. Also, there exists a compact interval in* $(0, \infty)$ *that contains the support of* H_n *for each* n*.*

Or, if the data follows a factor model $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$, the assumptions are:

(F1) *The series* $\{\boldsymbol{\varepsilon}_i\}$ *has* $\boldsymbol{\varepsilon}_i = \mathbf{\Sigma}_{\varepsilon}^{1/2}\boldsymbol{\xi}_i$*, where* $\boldsymbol{\xi}_i$ *is a* $p \times 1$ *vector of independent and identically distributed random variables* ξ_{ij} *. Each* ξ_{ij} *has mean 0 and unit variance,*

and $\mathbb{E}|\xi_{ij}|^k \leq B < \infty$ *for some constant* B *and* $k \leq 20$. *The factor series* $\{\mathbf{x}_t\}$ *has a constant dimension* r, *and* $\mathbf{x}_t = \mathbf{\Sigma}_x^{1/2}\mathbf{x}_t^*$, *where* $\mathbf{x}_t^*$ *is a* $r \times 1$ *vector of independent and identically distributed random variables* x_{ti}^*. *Also,* $\mathbb{E}|x_{ti}^*|^k \leq B < \infty$ *for some constant* B *and* $2 < k \leq 20$.

(F2) *The covariance matrix* $\mathbf{\Sigma}_x = var(\mathbf{x}_i)$ *is such that* $\left\|\mathbf{\Sigma}_x\right\| = O(1)$. *The covariance matrix* $\mathbf{\Sigma}_\varepsilon = var(\boldsymbol{\varepsilon}_i)$ *also has* $\left\|\mathbf{\Sigma}_\varepsilon\right\| = O(1)$. *Both covariance matrices are non-random. The factor loading matrix* $\mathbf{A}$ *is such that* $\left\|\mathbf{A}\right\|_F^2 = tr(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = O(p)$.

Note that for ease of presentation, Assumptions (A1) and (F1) here are actually the Assumptions (A1') and (F1') in Lam (2016). Please refer to Lam (2016) for more detailed explanations of these assumptions. Assumption (F2) with $\left\|\mathbf{A}\right\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}}) = O(p)$ entails both strong and weak factors as defined Lam and Yao (2012), as the population eigenvalues of $\mathbf{\Sigma}_y$ diverge with rate $O(p)$. Based on the above assumptions, Lam (2016) derived several theoretical results for NERCOME. We quote some of these results below, and they will be utilised to prove normalized consistency.

Lemma 6.1. *Let Assumptions (A1) to (A4) be satisfied. Or, if the data follows a factor model, let Assumptions (F1) and (F2) be satisfied. Suppose* $p/n_1 \to c_1 > 0$ *and* $\sum_{n\geq 1} n_2^{-3} < \infty$, *then for almost all* $x \in \mathbb{R}$,

$$\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} - \frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} \xrightarrow{a.s.} 0,$$

where $\lambda_{11} \geq \lambda_{12} \geq \cdots \geq \lambda_{1p}$ *are the eigenvalues of* $\widetilde{\mathbf{\Sigma}}_1$. *Furthermore, if the split location* m *is such that* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$, *then*

$$\max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i}}\right| \xrightarrow{a.s.} 0.$$

Lemma 6.1 summarizes the results from Theorem 1, Theorem 3 and Lemma 1 from Lam (2016). Moreover, Lam (2016) demonstrated that $\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}}$ converges to the non-random limit $\int_{-\infty}^{x}\delta(\lambda)dF(\lambda)$, where $\delta(\lambda)$ is the asymptotic nonlinear transform of $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$ as defined in (2.7) of Lam (2016), and the distribution function $F(\lambda)$ is the non-random limit of the empirical distribution the sample eigenvalues.

Based on the above results, we show normalized consistency for both NERCOME $\widehat{\mathbf{\Sigma}}_m$ and the precision matrix estimator $\widehat{\mathbf{\Omega}}_m$.

Theorem 6.1. *Let all assumptions in Lemma 6.1 be satisfied. Define* $f_s = \left\|\mathbf{\Sigma}_p\right\|_F$ *and* $\widehat{f}_s = \left\|\widehat{\mathbf{\Sigma}}_m\right\|_F$. *Then, if* $\left\|\mathbf{\Sigma}_p\right\| = o(\min(p^{1/4}, f_s))$, *the covariance matrix estimator* $\widehat{\mathbf{\Sigma}}_m$ *defined in (6.3) satisfies*

$$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_p/f_s) = o_{\mathbb{P}}(1). \tag{6.9}$$

Similarly, let $f_o = \left\|\mathbf{\Omega}_p\right\|_F$ *and* $\widehat{f}_o = \left\|\widehat{\mathbf{\Omega}}_m\right\|_F$. *If* $\left\|\mathbf{\Omega}_p\right\| = o(\min(p^{1/4}, f_o))$, *then the corresponding precision matrix estimator* $\widehat{\mathbf{\Omega}}_m$, *as defined in (6.4), satisfies*

$$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o) = o_{\mathbb{P}}(1). \tag{6.10}$$

Theorem 6.1 asserts that normalized consistency for NERCOME is achieved when $\left\|\mathbf{\Sigma}_p\right\| = o(\min(p^{1/4}, f_s))$. Note that the rate requirement $\left\|\mathbf{\Sigma}_p\right\| = o(p^{1/4})$ is more relaxed than Theorem 5 of Lam (2016), which shows asymptotic efficiency of NERCOME relative to the ideal estimator $\widehat{\mathbf{\Sigma}}_{\text{Ideal}}$ under the assumption that $\left\|\mathbf{\Sigma}_p\right\| = O(1)$. However, if the data follows a factor model, then the rate $\left\|\mathbf{\Sigma}_p\right\| = o(p^{1/4})$ correspond to a very weak factor, which excludes most practical applications of factor models. In fact, Theorem 5 of Lam (2016) excludes the data from a factor model as well. Hence, we recommend using NERCOME in general scenarios when a factor model is not a prior belief, though we also show in Section 6.4 that the empirical performance of NERCOME is robust to such data structure. On the other hand, the condition $\left\|\mathbf{\Omega}_p\right\| = o(p^{1/4})$ for achieving normalized consistency of the precision matrix estimator $\widehat{\mathbf{\Omega}}_m$ is mild, since the largest eigenvalue of a precision matrix typically remains bounded, even in a factor model with strong factors.

Proof of Theorem 6.1 From the proof of Theorem 3.3 of Xu et al. (2015), for a matrix $\mathbf{\Sigma}$ (which can be the covariance matrix $\mathbf{\Sigma}_p$ or the precision matrix $\mathbf{\Omega}_p$ in our case), let $\widehat{\mathbf{\Sigma}}$ be an estimate of $\mathbf{\Sigma}$. If the following condition is satisfied:

$$\text{tr}(\widehat{\mathbf{\Sigma}}^4)/\widehat{f}^4 = o_{\mathbb{P}}(1), \tag{6.11}$$

where $\widehat{f} = \left\|\widehat{\mathbf{\Sigma}}\right\|_F$, then normalized consistency holds if and only if $\rho(\mathbf{\Sigma}) = o(f)$, where $f = \left\|\mathbf{\Sigma}\right\|_F$. Thus, we only need to prove that condition (6.11) holds for $\widehat{\mathbf{\Sigma}}_m$ and $\widehat{\mathbf{\Omega}}_m$.

For the covariance matrix estimator $\widehat{\mathbf{\Sigma}}_m$, the left hand side of (6.11) can be written as

$$\frac{\text{tr}[(\widehat{\mathbf{\Sigma}}_m)^4]}{\text{tr}^2(\widehat{\mathbf{\Sigma}}_m)^2} = \frac{\frac{1}{p}\sum_{i=1}^p (\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^4}{\frac{1}{p}\left(\sum_{i=1}^p (\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^2\right)^2}. \tag{6.12}$$

We deal with the numerator and denominator of (6.12) separately. For the numerator, note that

$$\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^4 \leq 8\left(\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}-\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i})^4+\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i})^4\right).$$

For the first term on the right hand side,

$$\begin{aligned}\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}-\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i})^4 &\leq \left(\max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}-\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i}}\right|\right)^4 \cdot \max_{1\leq i\leq p}(\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i})^4\\ &= o_{\mathbb{P}}(1)*o(p)\\ &= o_{\mathbb{P}}(p),\end{aligned}$$

where the second line follows from Lemma 6.1. For the second term,

$$\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i})^4 \leq \|\mathbf{\Sigma}_p\|^4 = o(p).$$

Thus, $\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^4 = o_{\mathbb{P}}(p)$, which gives the upper bound of the numerator of (6.12). For the denominator, note that by Arithmetic-Quadratic Means inequality,

$$\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^2 \geq p\left(\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\right)^2.$$

Then, by Lemma 6.1, we have that $\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}$ converges to $\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^T\mathbf{\Sigma}_p\mathbf{p}_{1i}$ almost surely, which again converges to a non-random limit of a non-zero constant M. Thus,

$$\frac{1}{p}\left(\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^2\right)^2 \geq \frac{1}{p}\left(p\left(\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\right)^2\right)^2 = p\left(\frac{1}{p}\sum_{i=1}^{p}\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i}\right)^4 \to pM^4,$$

which gives an lower bound for the numerator of (6.12). Therefore, we finally have that

$$\frac{\mathrm{tr}[(\widehat{\mathbf{\Sigma}}_m)^4]}{\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_m)^2} = \frac{\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^4}{\frac{1}{p}\left(\sum_{i=1}^{p}(\mathbf{p}_{1i}^T\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i})^2\right)^2} \leq \frac{o_{\mathbb{P}}(p)}{pM^4} = o_{\mathbb{P}}(1),$$

which implies normalized consistency of $\widehat{\mathbf{\Sigma}}_m$ iff $\|\mathbf{\Sigma}_p\| = o(f_s)$. This proves (6.9), the normalized consistency of $\widehat{\mathbf{\Sigma}}_m$. To show normalized consistency of the precision matrix

estimator, recall that

$$\widehat{\mathbf{\Omega}}_m = \widehat{\mathbf{\Sigma}}_m^{-1} = \mathbf{P}_1 \left(\text{diag}(\mathbf{P}_1^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{P}_1) \right)^{-1} \mathbf{P}_1^{\mathrm{T}}.$$

Thus, we want to check condition (6.11) for $\widehat{\mathbf{\Omega}}_m$ by showing that

$$\frac{\text{tr}[(\widehat{\mathbf{\Omega}}_m)^4]}{\text{tr}^2(\widehat{\mathbf{\Omega}}_m)^2} = \frac{\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^{-4}}{\frac{1}{p}\left(\sum_{i=1}^{p}(\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^{-2}\right)^2} = o_{\mathbb{P}}(1).$$

Similarly, for the numerator, we have

$$\begin{aligned}
\frac{1}{p}\sum_{i=1}^{p}(\frac{1}{\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}})^4 &\leq \left(\max_{1\leq i\leq p} \left| \frac{\mathbf{p}_{1i}^T \mathbf{\Sigma}_p \mathbf{p}_{1i}}{\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}} \right| \right)^4 . \max_{1\leq i\leq p} (\frac{1}{\mathbf{p}_{1i}^T \mathbf{\Sigma}_p \mathbf{p}_{1i}})^4 \\
&\leq \left(\max_{1\leq i\leq p} \left| \frac{\mathbf{p}_{1i}^T \mathbf{\Sigma}_p \mathbf{p}_{1i}}{\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}} \right| \right)^4 . \left\| \mathbf{\Omega}_p \right\|^4 \\
&= o_{\mathbb{P}}(p),
\end{aligned}$$

since $\max_{1\leq i\leq p} \left| \frac{p_{1i}^T \mathbf{\Sigma}_p p_{1i}}{p_{1i}^T \widetilde{\mathbf{\Sigma}}_2 p_{1i}} \right|$ is bounded. For the denominator, by the Harmonic-Arithmetic-Quadratic Means inequality,

$$\begin{aligned}
\frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)^2 &\geq \left(\frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)\right)^2 \\
&\geq \left(\frac{p}{\sum_{i=1}^{p} \mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)^2 \\
&= \left(\frac{1}{\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)^2 \to \frac{1}{M^2},
\end{aligned}$$

where M is a constant. Thus, the denominator

$$\frac{1}{p}\left(\sum_{i=1}^{p}(\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^{-2}\right)^2 \geq \frac{1}{p}\left(p\left(\frac{1}{\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)^2\right)^2 = p\left(\frac{1}{\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}}\right)^4 \to \frac{p}{M^4}.$$

Therefore,

$$\frac{\text{tr}[(\widehat{\mathbf{\Omega}}_m)^4]}{\text{tr}^2(\widehat{\mathbf{\Omega}}_m)^2} = \frac{\frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^{-4}}{\frac{1}{p}\left(\sum_{i=1}^{p}(\mathbf{p}_{1i}^T \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i})^{-2}\right)^2} \leq \frac{o_{\mathbb{P}}(p)}{\frac{p}{M^4}} = o_{\mathbb{P}}(1),$$

which implies normalized consistency of $\widehat{\mathbf{\Omega}}_m$ iff $\left\|\mathbf{\Omega}_p\right\| = o(f_o)$. This proves (6.10), the normalized consistency of $\widehat{\mathbf{\Omega}}_m$, and completes the proof of Theorem 6.1. □

6.4 Simulation Experiments

In this section, we conduct numerical studies to examine the empirical evidence of normalized consistency for NERCOME under various data structures. Additionally, we explore an application of normalized consistency in high dimensional hypothesis testing and demonstrate the effectiveness and superiority of using NERCOME for the test.

6.4.1 Simulation settings

To investigate the performance of NERCOME under various covariance structures, we create five profiles of covariance matrices as follows:

(i) $\mathbf{\Sigma}_p = \mathbf{I}_p$.

(ii) The data comes from the factor model $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{A}$ has size $p \times 3$ and elements independently generated from the $N(0, 2^2)$ distribution. Moreover, $\mathbf{x}_i \overset{iid}{\sim} N(0, 2\mathbf{I}_r)$, and $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \mathbf{I}_p)$.

(iii) Generate $\mathbf{\Sigma}_p$ via the Cholesky decomposition $\mathbf{\Sigma}_p = \mathbf{S}\mathbf{S}^{\mathrm{T}}$, where $\mathbf{S}$ is a $p \times p$ random matrix with elements independently drawn from the Uniform$[0, 1]$ distribution. Next, standardize $\mathbf{\Sigma}_p$ into a correlation matrix so that the diagonal elements are all equal to 1

(iv) $\mathbf{\Sigma}_p = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}}$. The orthogonal matrix $\mathbf{Q}$ is randomly generated each time, and $\mathbf{D}$ is diagonal, with 40% of its values equal to 3, and 60% equal to 7.

(v) Identical to (iii), except that the elements of matrix $\mathbf{S}$ are independently drawn from the Uniform$[-1, 1]$ distribution.

In each profile, we generate $\mathbf{y}_i$, $i \in [n]$ as i.i.d. standard normal, i.e. $\mathbf{y}_i \overset{iid}{\sim} N(0, \mathbf{\Sigma}_p)$. The above five profiles can be categorized into three groups based on the dependence structure of $\mathbf{y}_i$. Profile (i) represents the "no dependence" setting, as the entries of $\mathbf{y}_i$ are mutually independent. Profiles (ii) and (iii) represent the "strong dependence" setting, where both profiles satisfy $\left\|\mathbf{\Sigma}_p\right\| \asymp \left\|\mathbf{\Sigma}_p\right\|_F \asymp p$. This condition corresponds to condition

(i) of Theorem 3.3 in Xu et al. (2015), indicating that the sample covariance matrix can achieve normalized consistency. On the other hand, profiles (iv) and (v) represent the "weak dependence" setting, where both profiles satisfy $\left\|\mathbf{\Sigma}_p\right\| \asymp 1$ and $\left\|\mathbf{\Sigma}_p\right\|_F \asymp p^{0.5}$. We have proven that NERCOME satisfies normalized consistency in these profiles.

The five profiles are simulated 100 times under all combinations of $n = 50, 100, 200, 500$ and $p = 50, 100, 200$. We consider the split location $m = 0.8n$ when calculating NERCOME. Let $\mathbf{\Omega}_p = \mathbf{\Sigma}_p^{-1}$ be the true precision matrix, and f_s and f_o be the Frobenius norms of $\mathbf{\Sigma}_p$ and $\mathbf{\Omega}_p$, respectively. We denote the NERCOME estimates of the covariance matrix and precision matrix as $\widehat{\mathbf{\Sigma}}_m$ and $\widehat{\mathbf{\Omega}}_m$, respectively, and their corresponding Frobenius norms as $\widehat{f}_s$ and $\widehat{f}_o$, respectively. For each simulation, $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_s/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o)$ are computed to assess the normalized consistency of the covariance and precision matrix estimates. Tables 6.1 through 6.5 report the averages of these quantities over 100 repetitions for each profile.

From the results, both $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o)$ converge to 0 for all profiles as n (or p) increases. Thus, empirical evidence shows that normalized consistency holds for NERCOME in all the profiles we considered. The empirical rate of convergence may depend on the profile setting. In profile (i), where all entries of $\mathbf{y}_i$ are mutually independent, $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o)$ converge very rapidly with relatively small n and p. The convergence rate may be slower when $\mathbf{\Sigma}_p$ are generated from more complicated structures, and it may vary for $\widehat{\mathbf{\Sigma}}_m$ and $\widehat{\mathbf{\Omega}}_m$ as well. However, normalized consistency still holds in general. Note that although we have not proven the normalized consistency of NERCOME when $\left\|\mathbf{\Sigma}_p\right\| \asymp \left\|\mathbf{\Sigma}_p\right\|_F$ as in profiles (ii) and (iii), empirical studies show that NERCOME still exhibits normalized consistency in this case.

$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_p/f_s)$				$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o)$			
n	$p = 50$	$p = 100$	$p = 200$	n	$p = 50$	$p = 100$	$p = 200$
50	0.038	0.028	0.021	50	0.038	0.024	0.016
100	0.028	0.019	0.014	100	0.029	0.018	0.012
200	0.019	0.015	0.009	200	0.021	0.014	0.009
500	0.013	0.009	0.006	500	0.014	0.009	0.006

Table 6.1 Profile (i). Mean of $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f}_s - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f}_o - \mathbf{\Omega}_p/f_o)$ under different dimensions n and p.

$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$				$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$			
n	$p=50$	$p=100$	$p=200$	n	$p=50$	$p=100$	$p=200$
50	0.154	0.141	0.141	50	0.043	0.029	0.025
100	0.111	0.109	0.106	100	0.028	0.019	0.014
200	0.086	0.085	0.081	200	0.022	0.015	0.009
500	0.058	0.056	0.058	500	0.014	0.010	0.006

Table 6.2 Profile (ii). Mean of $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$ under different dimensions n and p.

$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$				$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$			
n	$p=50$	$p=100$	$p=200$	n	$p=50$	$p=100$	$p=200$
50	0.082	0.082	0.082	50	0.627	0.839	0.899
100	0.059	0.059	0.059	100	0.077	0.730	0.883
200	0.042	0.041	0.041	200	0.046	0.050	0.811
500	0.026	0.026	0.026	500	0.018	0.026	0.030

Table 6.3 Profile (iii). Mean of $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$ under different dimensions n and p.

$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$				$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$			
n	$p=50$	$p=100$	$p=200$	n	$p=50$	$p=100$	$p=200$
50	0.086	0.064	0.046	50	0.089	0.057	0.038
100	0.079	0.057	0.041	100	0.090	0.058	0.037
200	0.072	0.054	0.039	200	0.083	0.060	0.039
500	0.053	0.049	0.038	500	0.061	0.056	0.041

Table 6.4 Profile (iv). Mean of $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$ under different dimensions n and p.

$\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$				$\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$			
n	$p=50$	$p=100$	$p=200$	n	$p=50$	$p=100$	$p=200$
50	0.179	0.151	0.117	50	0.651	0.857	0.899
100	0.149	0.130	0.108	100	0.081	0.747	0.892
200	0.120	0.110	0.094	200	0.038	0.046	0.827
500	0.084	0.081	0.073	500	0.026	0.028	0.036

Table 6.5 Profile (v). Mean of $\rho(\widehat{\mathbf{\Sigma}}_m/\widehat{f_s} - \mathbf{\Sigma}_p/f_s)$ and $\rho(\widehat{\mathbf{\Omega}}_m/\widehat{f_o} - \mathbf{\Omega}_p/f_o)$ under different dimensions n and p.

6.4.2 A high dimensional hypothesis test

We consider an application of normalized consistency for NERCOME, which is a high dimensional hypothesis test proposed by Xu et al. (2015). They developed an asymptotic theory for the L_2 norms of sample mean vectors in high dimensional data. An invariance principle for the L_2 norms is derived, and normalized consistency is proposed to guarantee the validity of resampling procedures for computing the cutoff value of a high dimensional hypothesis test. Here, we utilize NERCOME as an estimate of the covariance matrix and illustrate its superior performance compared to the sample covariance matrix in various settings.

Consider $\mathbf{x}_1, ..., \mathbf{x}_n$ as independent and identically distributed (i.i.d.) observations with a mean $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}_p$. To test the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \mathbf{0}$, Xu et al. (2015) proposed the test statistic $n\bar{\mathbf{x}}_n^{\mathrm{T}}\bar{\mathbf{x}}_n$, where $\bar{\mathbf{x}}_n = \sum_{i=1}^n \mathbf{x}_i/n$. They demonstrated that its distribution is asymptotically close to that of $\mathbf{y}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$ is the Gaussian analogue of $\mathbf{x}$. Specifically, define

$$R_n = \frac{n|\bar{\mathbf{x}}_n|_2^2 - f_1}{f},$$

where $f_1 = \mathrm{tr}(\boldsymbol{\Sigma}_p)$, and $f = \left\|\boldsymbol{\Sigma}_p\right\|_F$. Let

$$V = \sum_{j=1}^{p} \frac{\lambda_j}{f}(\eta_j - 1),$$

where λ_j's are the eigenvalues of $\boldsymbol{\Sigma}_p$, and η_j's are i.i.d. χ_1^2 random variables. In their Theorem 2.2, Xu et al. (2015) showed that

$$\sup_t |\mathbb{P}(R_n \le t) - \mathbb{P}(V \le t)| \to 0. \tag{6.13}$$

Thus, given the significance level $\alpha \in (0,1)$, let $u_{1-\alpha}$ be the $(1-\alpha)$th quantile of V, namely $\mathbb{P}(V \le u_{1-\alpha}) = 1-\alpha$. Then H_0 is rejected if $R_n > u_{1-\alpha}$. In (6.13), if $\boldsymbol{\Sigma}_p$ is known, then the distribution of V can be easily computed, either numerically or analytically. However, $\boldsymbol{\Sigma}_p$ is usually unknown in most applications. In such cases, Xu et al. (2015) showed that if we have a normalized consistent estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ with corresponding eigenvalues $\widehat{\lambda}_j$ and Frobenious norm $\widehat{f}$, then

$$\max_{j \le p} |f^{-1}\lambda_j - \widehat{f}^{-1}\widehat{\lambda}_j| \xrightarrow{p} 0.$$

Thus, with probability converging to 1, we have

$$\sup_t |\mathbb{P}(V \le t) - \mathbb{P}(\widehat{V} \le t)| \to 0, \tag{6.14}$$

where $\widehat{V} = \sum_{j=1}^{p} \frac{\widehat{\lambda}_j}{\widehat{f}}(\eta_j - 1)$. With (6.14), the distribution of V can be approximated by that of $\widehat{V}$ via simulations, and the critical values of testing H_0 can be computed accordingly.

With normalized consistency established for NERCOME, it offers theoretical assurances for its utilization as an estimate of the covariance matrix when simulating $\widehat{V}$. In the following, we conduct simulations to compare the performances of using NERCOME and the sample covariance matrix in the aforementioned high dimensional hypothesis test.

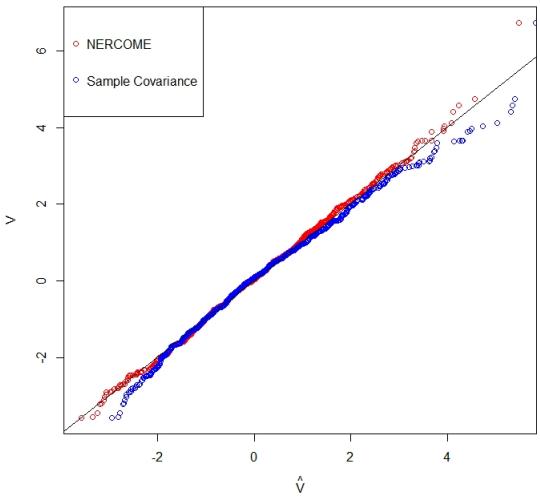
To investigate various data structures, we generate five profiles of covariance matrices as defined in Section 6.4.1. To test the performance of NERCOME and the sample covariance matrix under heavy-tailed distributions, we consider two distributions for $\mathbf{x}_i$, $i \in [n]$, for each profile:

(1) i.i.d. standard normal, i.e. $\mathbf{x}_i \overset{iid}{\sim} N(0, \mathbf{\Sigma}_p)$.

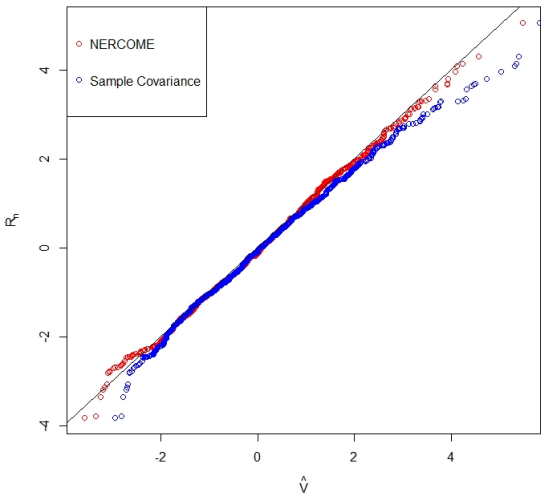
(2) i.i.d. t_5, i.e. $\mathbf{x}_i \overset{iid}{\sim} t_5(0, \mathbf{\Sigma}_p)$.

Thus, there are totally ten profiles considered. For each profile, we let $p = 50$ and $n = 50$ and generate $\mathbf{x}_1, \cdots, \mathbf{x}_n$ accordingly. We then calculate NERCOME $\widehat{\mathbf{\Sigma}}_m$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}}_s = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$. Next, we simulate the distribution of $\widehat{V}$ based on both $\widehat{\mathbf{\Sigma}}_m$ and $\widehat{\mathbf{\Sigma}}_s$, respectively. We compare the distribution of $\widehat{V}$ with V and $\widehat{R}_n$, where V is simulated based on the true covariance matrix $\mathbf{\Sigma}_p$, and $\widehat{R}_n = \frac{n|\mathbf{x}_n|_2^2 - \widehat{f}_1}{f^\dagger}$, where $f^\dagger$ is the ratio consistent estimate of f, namely $f^\dagger/f - 1 = o_{\mathbb{P}}(1)$ (Bai and Saranadasa, 1996; Chen and Qin, 2010; Xu et al., 2015). In Figure 6.1 to 6.5, we draw QQ-plots to measure the closeness of distributions between V and $\widehat{V}$, as well as $\widehat{R}_n$ and $\widehat{V}$, under different profiles of covariance matrices. We expect a "good" normalized consistent covariance matrix estimator should give $\widehat{V}$ that closely approximates V.

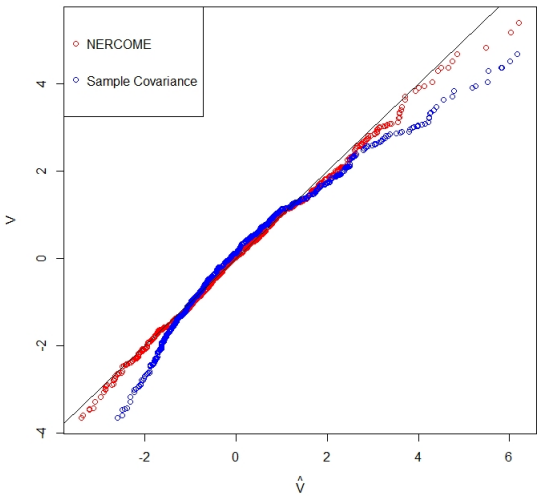
From Figure 6.1 to 6.5, when the data are normally distributed, in all profiles, both NERCOME and the sample covariance matrix provide good estimates of V by $\widehat{V}$. Also, $\widehat{R}_n$ converges to $\widehat{V}$ quickly. However, when the data are heavy-tailed distributed, differences occur among different profiles. In profiles (i), (iv), and (v), where $\left\|\mathbf{\Sigma}_p\right\| = o(f)$, NERCOME provides better approximations of V by $\widehat{V}$ compared to the sample covariance matrix (see, for example, Figures 6.1(c), 6.4(c), 6.5(c)); and $\widehat{R}_n$ converges to $\widehat{V}$ at a slower rate. It's worth noting that we have proven the normalized consistency of NERCOME in these



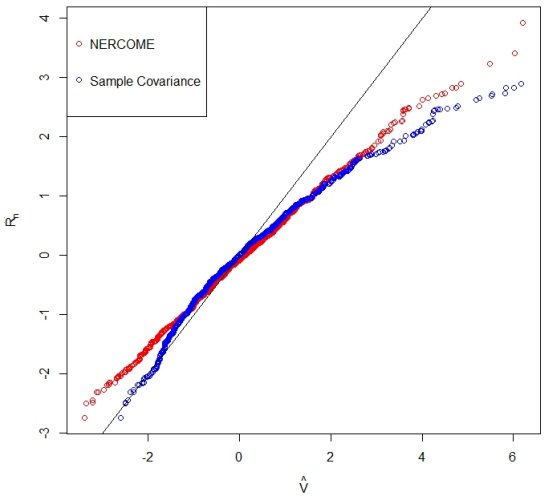
(a) V vs $\widehat{V}$: Normal



(b) $\widehat{R}_n$ vs $\widehat{V}$: Normal

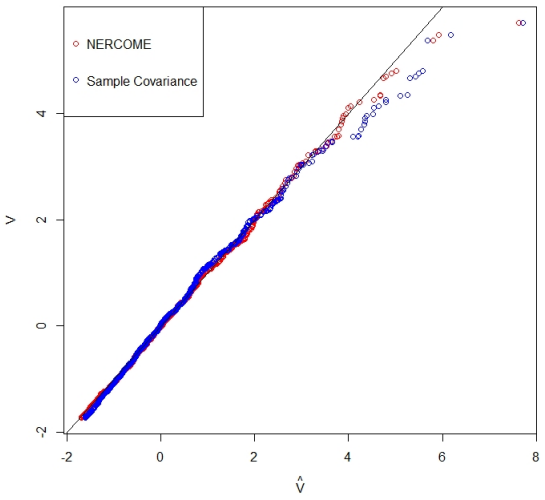


(c) V vs $\widehat{V}$: t_5

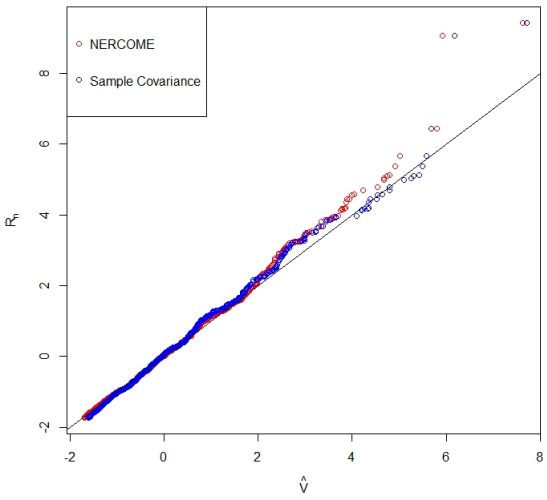


(d) $\widehat{R}_n$ vs $\widehat{V}$: t_5

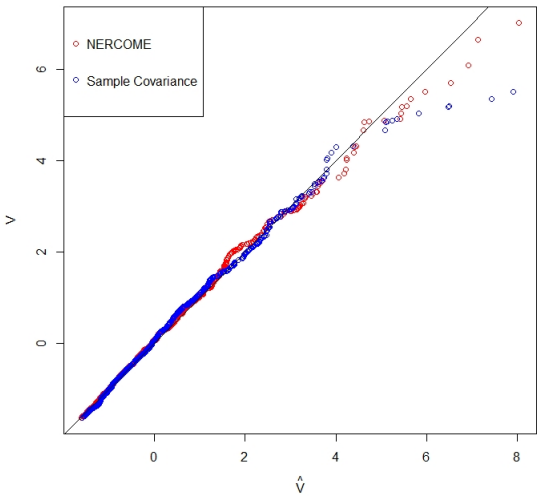
Fig. 6.1 Profile (i). (a) QQ-plot of V v.s. $\widehat{V}$, normally distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, normally distributed data; (c) QQ-plot of V v.s. $\widehat{V}$, t_5 distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, t_5 distributed data. Red: NERCOME; blue: sample covariance matrix.



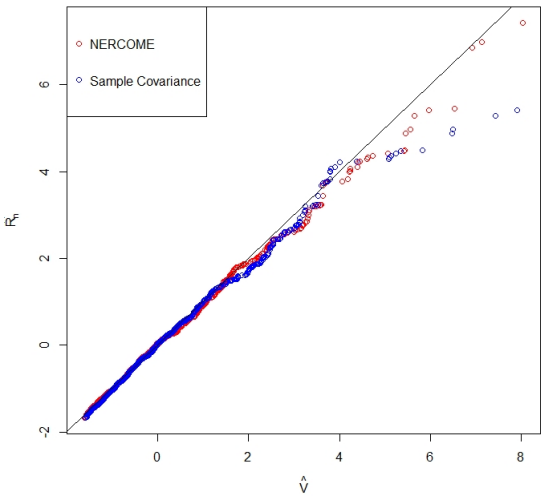
(a) V vs $\widehat{V}$: Normal



(b) $\widehat{R}_n$ vs $\widehat{V}$: Normal

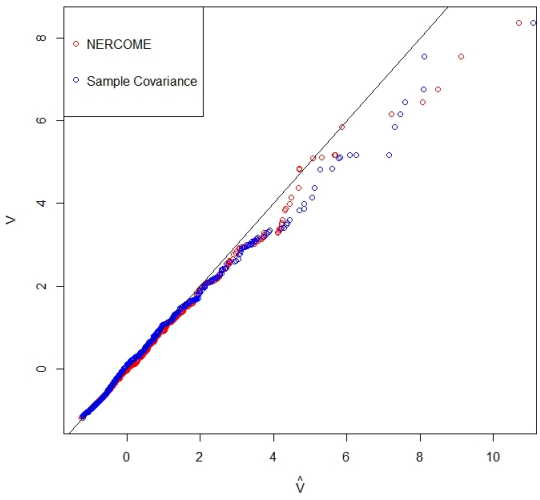


(c) V vs $\widehat{V}$: t_5

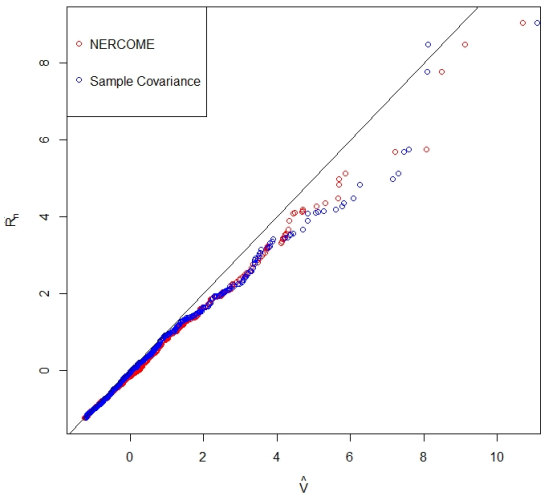


(d) $\widehat{R}_n$ vs $\widehat{V}$: t_5

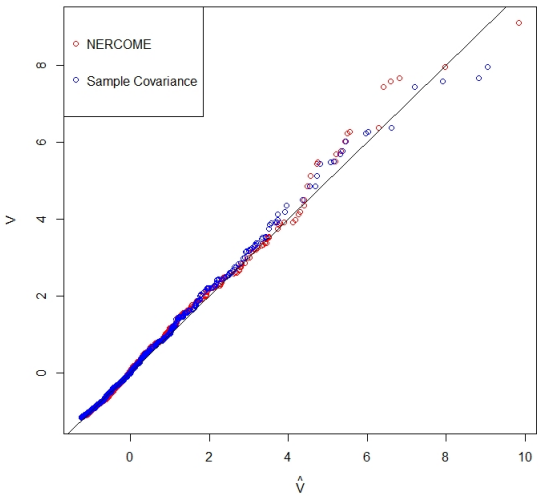
Fig. 6.2 Profile (ii). (a) QQ-plot of V v.s. $\widehat{V}$, normally distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, normally distributed data; (c) QQ-plot of V v.s. $\widehat{V}$, t_5 distributed data; (d) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, t_5 distributed data. Red: NERCOME; blue: sample covariance matrix.



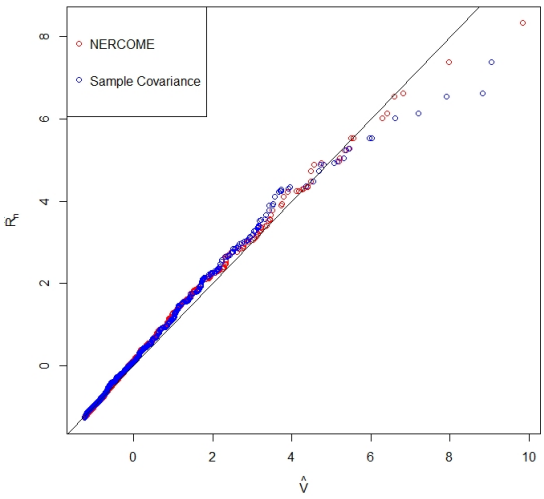
(a) V vs $\widehat{V}$: Normal



(b) $\widehat{R}_n$ vs $\widehat{V}$: Normal

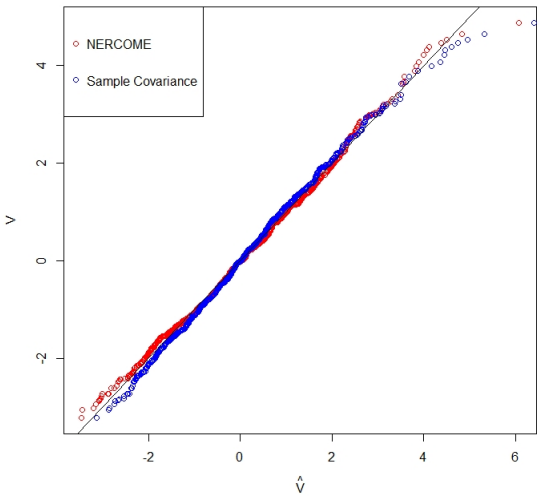


(c) V vs $\widehat{V}$: t_5

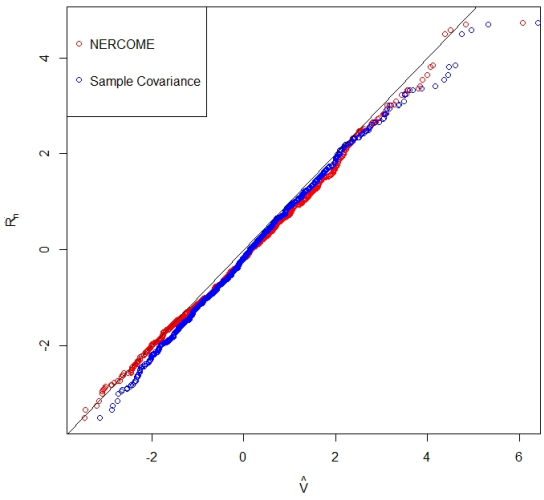


(d) $\widehat{R}_n$ vs $\widehat{V}$: t_5

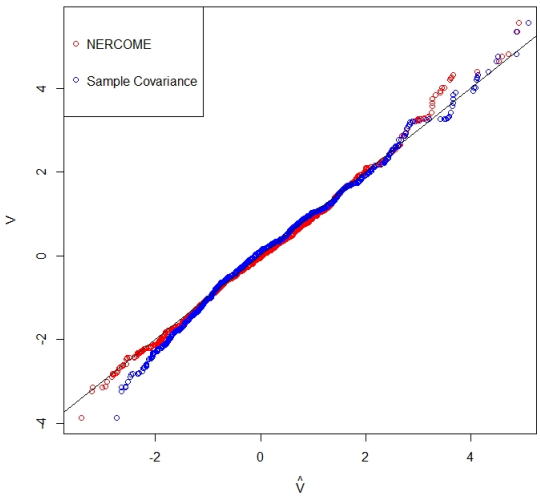
Fig. 6.3 Profile (iii). (a) QQ-plot of V v.s. $\widehat{V}$, normally distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, normally distributed data; (c) QQ-plot of V v.s. $\widehat{V}$, t_5 distributed data; (d) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, t_5 distributed data. Red: NERCOME; blue: sample covariance matrix.



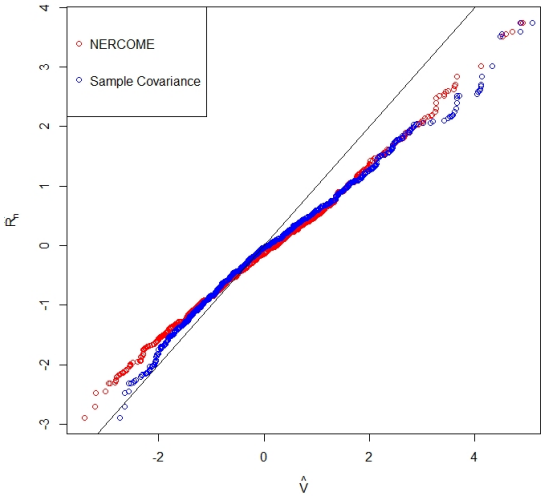
(a) V vs $\widehat{V}$: Normal



(b) $\widehat{R}_n$ vs $\widehat{V}$: Normal

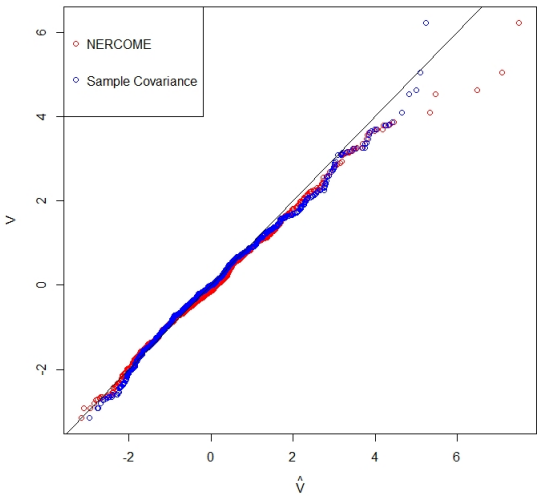


(c) V vs $\widehat{V}$: t_5

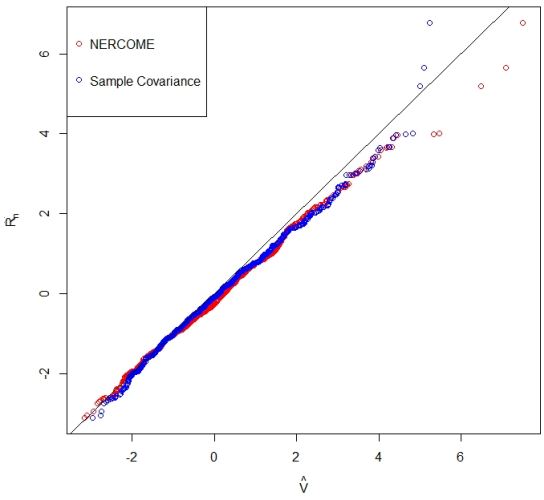


(d) $\widehat{R}_n$ vs $\widehat{V}$: t_5

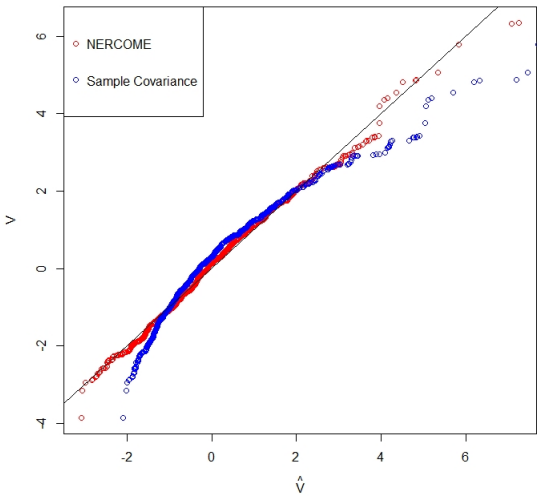
Fig. 6.4 Profile (iv). (a) QQ-plot of V v.s. $\widehat{V}$, normally distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, normally distributed data; (c) QQ-plot of V v.s. $\widehat{V}$, t_5 distributed data; (d) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, t_5 distributed data. Red: NERCOME; blue: sample covariance matrix.



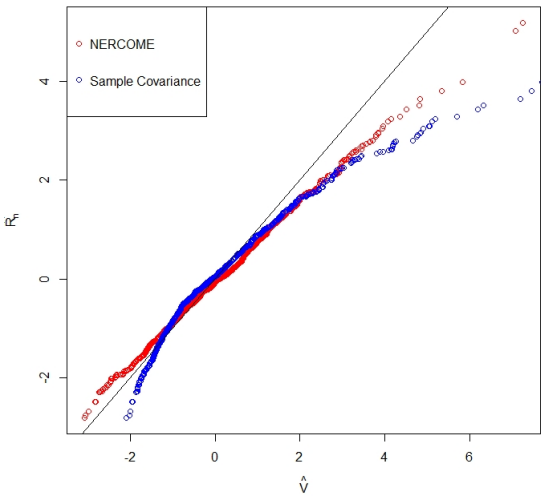
(a) V vs $\widehat{V}$: Normal



(b) $\widehat{R}_n$ vs $\widehat{V}$: Normal



(c) V vs $\widehat{V}$: t_5



(d) $\widehat{R}_n$ vs $\widehat{V}$: t_5

Fig. 6.5 Profile (v). (a) QQ-plot of V v.s. $\widehat{V}$, normally distributed data; (b) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, normally distributed data; (c) QQ-plot of V v.s. $\widehat{V}$, t_5 distributed data; (d) QQ-plot of $\widehat{R}_n$ v.s. $\widehat{V}$, t_5 distributed data. Red: NERCOME; blue: sample covariance matrix.

cases, whereas the sample covariance matrix may not exhibit normalized consistency (Xu et al., 2015). Thus, the empirical studies align with our theoretical results. In profiles (ii) and (iii), where $\left\|\mathbf{\Sigma}_p\right\| \asymp f$, both NERCOME and the sample covariance matrix yield good estimators of V, and $\widehat{R}_n$ converges to $\widehat{V}$ quickly. Although we have not proven the normalized consistency of NERCOME in profile (ii) and (iii), simulation studies show that NERCOME also performs well in this case.

The empirical studies demonstrate that, in general, using NERCOME can provide better approximations of V compared to using the sample covariance matrix (Xu et al., 2015), especially when $\left\|\mathbf{\Sigma}_p\right\| = o(f)$ and the data are heavy-tailed distributed. This suggests that the hypothesis testing procedure proposed by Xu et al. (2015) should be more accurate if NERCOME is used to approximate the critical values.

References

Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165 – 180.

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

Amengual, D. and Watson, M. W. (2007). Consistent estimation of the number of dynamic factors in a large n and t panel. *Journal of Businessand Economic Statistics*, 25(1):91–96.

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

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1).

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, 191(1):1–18.

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. (2007). Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics*, 25(1):52–60.

Bai, J. and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, 235(2):1893–1916.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, 6(2):311–329.

Bai, Z. and Silverstein, J. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics, New York, 2 edition.

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

Bailey, N., Kapetanios, G., and Pesaran, M. H. (2021). Measurement of factor strength: Theory and practice. *Journal of Applied Econometrics*, 36(5):587–613.

Barigozzi, M. and Hallin, M. (2024). Dynamic factor models: a genealogy. *arXiv:2310.17278*.

Barigozzi, M., He, Y., Li, L., and Trapani, L. (2023a). Robust tensor factor analysis. *arXiv:2303.18163*.

Barigozzi, M., He, Y., Li, L., and Trapani, L. (2023b). Statistical inference for large-dimensional tensor factor models by iterative projections. *arXiv:2206.09800*.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis*. Wiley.

Bernanke, B. S. (1986). Alternative explanations of the money-income correlation. *Carnegie-Rochester Conference Series on Public Policy*, 25:49–99.

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

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144.

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

Chang, J., He, J., Yang, L., and Yao, Q. (2023). Modelling matrix time series via a tensor cp-decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):127–148.

Chen, E. Y. and Chen, R. (2022). Modeling dynamic transport network with matrix factor models: An application to international trade flow. *Journal of Data Science*, page 490–507.

Chen, E. Y. and Fan, J. (2021). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055.

Chen, E. Y., Xia, D., Cai, C., and Fan, J. (2020). Semiparametric tensor factor analysis by iteratively projected SVD. *arXiv:2007.02404*.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960.

Chen, R., Xiao, H., and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1):539–560.

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Cheng, C., Wei, Y., and Chen, Y. (2021). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(11):7380–7419.

Connor, G. and Korajczyk, R. A. (2022). Semi-strong factors in asset returns. *Journal of Financial Econometrics*, 22(1):70–93.

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., Guo, J., and Zheng, S. (2022). Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 117(538):852–861.

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

Fan, J., Wang, W., and Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5 – 22. Special Issue on Financial Engineering and Risk Management.

Forni, M. and Gambetti, L. (2010). The dynamic effects of monetary policy: A structural factor model approach. *Journal of Monetary Economics*, 57(2):203–216.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, 25(5):1319–1347.

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.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, 50(6):1243–1255.

Forni, M. and Lippi, M. (2001). The generalized dynamic factor model: Representation theory. *Econometric Theory*, 17(6):1113–1141.

Freyaldenhoven, S. (2022). Factor models with local factors — determining the number of relevant factors. *Journal of Econometrics*, 229(1):80–102.

Freyaldenhoven, S. (2023). Identification through sparsity in factor models: The l_1-rotation criterion. *Federal Reserve Bank of Philadelphia*.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins University Press, 3rd edition.

Guo, S., Box, J. L., and Zhang, W. (2017). A dynamic structure for high-dimensional covariance matrices and its application in portfolio allocation. *Journal of the American Statistical Association*, 112(517):235–253.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Hallin, M. and Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics*, 163(1):29–41.

Han, Y., Chen, R., Yang, D., and Zhang, C.-H. (2020). Tensor factor model estimation by iterative projection. *arXiv: 2006.02611*.

Han, Y., Zhang, C. H., and Chen, R. (2022). Rank determination in tensor factor model. *Electronic Journal of Statistic*, 16:1726–1803.

Hartigan, J. A. (2014). Bounding the maximum of dependent random variables. *Electronic Journal of Statistics*, 8(2):3126 – 3140.

He, Y., Kong, X., Yu, L., Zhang, X., and Zhao, C. (2023a). Matrix factor analysis: From least squares to iterative projection. *Journal of Business and Economic Statistics*, 0(0):1 – 13.

He, Y., Kong, X.-B., Liu, D., and Zhao, R. (2023b). Robust statistical inference for large-dimensional matrix-valued time series via iterative huber regression. *arXiv:2306.03317*.

He, Y., Wang, Y., Yu, L., Zhou, W., and Zhou, W.-X. (2022). Matrix kendall's tau in high-dimensions: A robust statistic for matrix factor model. *arXiv:2207.09633*.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif. University of California Press.

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. PMID: 20617121.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4).

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business and Economic Statistics*, 28(3):397–409.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Kong, X.-B. (2017). On the number of common factors with high-frequency data. *Biometrika*, 104(2):397–410.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist.*, 44(3):928–953.

Lam, C. (2021). Rank determination for time series tensor factor model using correlation thresholding. *Manuscript*.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278.

Lam, C. and Feng, P. (2018). A nonparametric eigenvalue-regularized integrated covariance matrix estimator for asset return data. *Journal of Econometrics*, 206(1):226 – 257.

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.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Latala, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282.

Latała, R. (2004). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, 12(3):209.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2013). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. ECON - Working Papers 122, Department of Economics - University of Zurich.

Lettau, M. (2022). High dimensional factor models with an application to mutual fund characteristics. *National Bureau of Economic Research*.

Li, H., Li, Q., and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *Journal of Econometrics*, 197(1):76–86.

Li, Z., Lam, C., Yao, J., and Yao, Q. (2019). On testing for high-dimensional white noise. *Ann. Statist.*, 47(6):3382–3412.

Liu, T., Yuan, M., and Zhao, H. (2022). Characterizing spatiotemporal transcriptome of the human brain via low-rank tensor decomposition. *Statistics in Biosciences*.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887.

Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electron. J. Statist.*, 6:2125–2149.

Moon, H. R. and Weidner, M. (2015). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016.

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

Pan, J. and Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90(1):239–244.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika*, 94(4):1006–1013.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1 -penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–980.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341 – 360.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics*, 34(1):1373–1403.

Stock, J. and Watson, M. (2005). Implications of dynamic factor models for var analysis. *NBER Working Papers*. No. 11467.

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

Tao, M., Su, J., and Wang, L. (2019). Land cover classification of PolSAR image using tensor representation and learning. *Journal of Applied Remote Sensing*, 13(1):016516.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

Trzcinka, C. (1986). On the number of factors in the arbitrage pricing model. *The Journal of Finance*, 41(2):347–368.

Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. *Problems in Measuring Change*, page 122–137.

Tucker, L. R. (1964). The extension of factor analysis to three-dimensional matrices. *Contributions to Mathematical Psychology*, page 110–127.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Uematsu, Y. and Yamagata, T. (2022). Estimation of sparsity-induced weak factor models. *Journal of Business Economic Statistics*, 41(1):213–227.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248. Special Issue on Financial Engineering and Risk Management.

Wang, L. and Paul, D. (2014). Limiting spectral distribution of renormalized separable sample covariance matrices when $p/n \to 0$. *Journal of Multivariate Analysis*, 126:25–52.

Williams, B. (2019). Identification of the linear factor model. *Econometric Reviews*, 39(1):92–109.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):427–450.

Xu, M., Zhang, D., and Wu, W. B. (2015). l^2 asymptotics for high-dimensional data. *arXiv:1405.7244*.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yokota, T., Lee, N., and Cichocki, A. (2017). Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria. *IEEE Transactions on Signal Processing*, 65(5):1196–1206.

Yu, L., He, Y., Kong, X., and Zhang, X. (2022). Projected estimation for large-dimensional matrix factor models. *Journal of Econometrics*, 229(1):201–217.

Zhang, A. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2).

Zhang, A. R. and Xia, D. (2018). Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64:7311–7338.