Various Statistical Inferences for High-dimensional Time Series: Bootstrap, Homogeneity Pursuit and Autocovariance Test

Daning Bi

A thesis submitted for the degree of Doctor of Philosophy in Statistics

The Australian National University

May 2021

Declaration

I hereby declare that the work in this thesis is my own except otherwise stated.

Daning Bi

21 May 2021

To my beloved family.

Acknowledgments

This thesis is based on my PhD program, founded by the Research School of Finance, Actuarial Studies and Statistics (RSFAS) at the Australian National University (ANU). I would like to thank RSFAS and ANU for providing me an opportunity and scholarships for doing research and meeting prestigious researchers. During the period of PhD studies, I have been helped by many individuals. I would like to express my sincere gratitude to those who made this thesis possible.

I would like to express my sincere appreciation to all of my supervisors, Dr. Yanrong Yang, Dr. Gen Nowak, and Professor Alan Welsh, for their continuous support and invaluable advice throughout this period. I am eternally grateful to Dr. Yanrong Yang for her invaluable guidance during my PhD studies. She has always been very generous and patient in supervising me to conduct the research projects in my thesis and always encourage me to pursue the best. I sincerely thank Dr. Gen Nowak and Professor Alan Welsh for providing their expertise in fields of statistics. Their wide knowledge and expert thinking have inspired me throughout my PhD studies.

RSFAS has been providing excellent opportunities for me to work with my colleagues and fellow candidates. In particular, I would like to thank Dr. Le Chang and Mr. Adam Nie with whom I worked closely for Chapter 2 and Chapter 3 of this thesis. I appreciate the help of Associate Professor Timothy Higgins for his considerable help and guidance during the whole period of my study at ANU. I would also like to thank Dr. Tao Zou for the generous advices on conducting research. Moreover, thanks to Mr. Adam Nie, Mr. Chen Tang, Dr. Yuan Gao, Dr. Jiali Wang, and Dr. Yang Yang for being great friends and making my candidature experience truly enjoyable.

Last but not the least, I wish to acknowledge the great support and uncondi-

tional love of my family, my wife, Lingyu; my mother, Wei; and my father Jifei. They kept me going on and this work would not have been possible without their companion.

Abstract

This thesis aims to study various statistical inferences for high-dimensional data, especially high-dimensional time series, including sieve bootstrap, homogeneity pursuit, and an equivalence test for spiked eigenvalues of autocovariance matrix. The primary techniques used in this thesis are novel dimension-reduction methods developed from factor models and principal component analysis (PCA).

Chapter 2 proposes a novel sieve bootstrap method for high-dimensional time series and applies it to sparse functional time series where the actual observations are not dense, and pre-smoothing is misleading. Chapter 3 introduces an iterative complement-clustering principal component analysis (CPCA) to study high-dimensional data with group structures, where both homogeneity and sub-homogeneity (group-specific information) can be identified and estimated. Lastly, Chapter 4 proposes a novel test statistic named the autocovariance test to compare the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series. In all chapters, dimension-reduction methods are applied for novel statistical inferences. In particular, Chapters 2 and 4 focus on the spiked eigenstructure of autocovariance matrix and use factors to capture the temporal dependence of the high-dimensional time series. Meanwhile, Chapter 3 aims to simultaneously estimate homogeneity and sub-homogeneity, which form a more complicated spiked eigenstructure of the covariance matrix, despite that the group-specific information is relatively weak compared with the homogeneity and traditional PCA fails to capture it.

The theoretical and asymptotic results of all three statistical inferences are provided in each chapter, respectively, where the numerical evidence on the finite-sample performance for each method is also discussed. Finally, these three statistical inferences are applied on particulate matter concentration data, stock return data, and age-specific mortality data for multiple countries, respectively, to provide valid statistical inferences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the developments in computer sciences and information technologies, an increasing number of data are now being collected for statistical analysis. High-dimensional data, especially high-dimensional time series data, are now widely seen in many disciplines such as economics, finance, meteorology, and biology. Despite the developments in computing powers, novel statistical methods for statistical inferences such as estimation, prediction, and hypothesis testing are in urgent demand for many scientific questions.

Unlike univariate or multivariate data with finite dimensions, when the data dimension and the number of unknown parameters grow with the sample size, the sample estimates of unknown parameter matrices such as covariance and autocovariance matrices for high-dimensional data are generally not consistent. This phenomenon is usually referred to as the 'curse of dimensionality', which has received innumerable attention in the past few years. Among numerous attempts to deal with the 'curse of dimensionality', most of them grow from either regularisation or dimension-reduction methods. In this thesis, novel statistical inferences based on variations of dimension-reduction methods, including factor models (Bai and Ng, 2002a; Bai, 2003a) and principal component analysis (PCA) (Jolliffe, 2002), are developed to study high-dimensional data, especially high-dimensional time series.

Factor models (Bai and Ng, 2002a; Bai, 2003a) are introduced for studying large dimensional data-set where both the data dimension N and sample size T tend to infinity. As an efficient dimension-reduction method, factor models transfer the study on high-dimensional data to low-dimensional factors. Following the

original work for independent data, Lam et al. (2011); Lam and Yao (2012) apply factor models on high-dimensional time series, where the temporal dependence of the high-dimensional data are assumed to be captured by low-dimensional factors. On the other hand, the idea of principal component analysis (PCA) can date back to the early twentieth century in Pearson (1901), and Hotelling (1933) for multivariate data with fixed dimensions. With the developments in recent years, PCA has become a well-known statistical method to represent high-dimensional data onto a low-dimensional space. Besides, (Fan et al., 2013a) summarises the relationship between PCA and factor models, where PCA can be utilised as a statistical technique to estimate factor models since PCA on the sample covariance matrix consistently estimates the eigenspace spanned by factor loadings. Moreover, both methods have received a vast number of developments in the past few years, and various variations have been introduced to solve different statistical problems. A complete survey on PCA and factor models' recent developments can be found in Johnstone and Paul (2018).

This thesis mainly contributes to three statistical inferences for high-dimensional data, especially high-dimensional time series. They are sieve bootstrap, homogeneity and sub-homogeneity pursuit, and the equivalence test for spiked eigenvalues of autocovariance matrix (the autocovariance test). Typically, the sieve bootstrap in Chapter 2 and the autocovariance test in Chapter 4 focus on the spiked eigenstructure of the autocovariance matrix and use factors to capture the temporal dependence of the original high-dimensional time series. On the other hand, Chapter 3 aims to develop a novel method to estimate homogeneity and sub-homogeneity (group-specific information), where a more complicated spiked eigenstructure exists in the covariance matrix and the group-specific information is relatively weak compared with the homogeneity hence hard to be estimated by traditional PCA methods.

First of all, Chapter 2 introduces a novel sieve bootstrap method for high-dimensional time series and applies it on sparse functional time series where traditional functional bootstrap methods fail to work as the observations are not dense enough, and pre-smoothing on sparse observations is misleading.

Bootstrap (Efron, 1979), as a pervasive and powerful tool, can be utilised to create confidence intervals and perform hypothesis testings when theoretical results are not available or hard to be applied. For time series data, the traditional non-parametric bootstrap method (Efron, 1979) fails to work since the temporal dependence of time series data cannot be correctly mimicked. Kreiss and Lahiri (2012) summarise a few variations of the traditional bootstrap method, including the block bootstrap (Kunsch, 1989), the autoregressive (AR) sieve bootstrap (Kreiss, 1988), and the bootstrap for Markov chains (Kulperger and Rao, 1989), for studying statistical inferences of time series data. Nonetheless, as discussed in El Karoui and Purdom (2018), traditional non-parametric bootstrap methods for univariate and multivariate data are not directly applicable for high-dimensional data, including time series, due to the 'curse of dimensionality'. In particular, the validity of the AR sieve bootstrap (Kreiss, 1988) relies on a so-called Boundedness Condition (Wiener and Masani, 1958) as presented in Chapter 2. Nevertheless, this condition cannot be fulfilled when $N \to \infty$ with the sample size T, hence the AR sieve bootstrap is not valid for high-dimensional time series. Consequently, to overcome the issues associated with the 'curse of dimensionality', Chapter 2 assumes high-dimensional time series follow factor models in (Lam et al., 2011) and apply the AR sieve bootstrap on low-dimensional factor time series. Finally, the bootstrapped data is transferred back to the original high-dimensional space by factor models to provide statistical inferences.

Secondly, a novel 'iterative complement-clustering principal component analysis' (CPCA) is introduced in Chapter 3, which aims at studying high-dimensional data with group structures. With recent improvements in computation powers, an increasing number of high-dimensional data, including time series, are now being collected. Moreover, it is natural to consider combining similar data for an aggregated analysis, which may benefit statistical inferences such as estimation and prediction. However, as discussed in Boivin and Ng (2006), grouping up more data does not always benefit statistical analysis since data from different populations may exhibit quite different patterns that increase the complexity and heterogeneity of the whole data-set. More specifically, for a particular group of

data, we consider it contains two styles of information. One is the information shared with all the groups, which forms the homogeneity for the whole data-set. However, another style of information, named sub-homogeneity, is group-specific and belongs to a particular group of data exclusively. Furthermore, since the sub-homogeneity is relatively weak compared with the homogeneity, when directly applying traditional dimension-reduction methods, such as PCA, on the whole data-set, sub-homogeneity in each group are not necessarily captured. To alleviate this issue, the whole data-set can be clustered into groups, where sub-homogeneity can be identified and estimated from each cluster. To achieve that, we propose a novel method named CPCA to identify both homogeneity and sub-homogeneity and handle the interaction between them iteratively. To be more specific, since each group's data is still high-dimensional, the CPCA method applies traditional PCA in each cluster first and then aggregates the principal component scores for a further PCA. Meanwhile, a leave-one-out principal component regression (PCR) clustering method inspired by Chiou and Li (2007) is performed for each iteration to improve the result of clustering actively. Consequently, this CPCA is essential for estimations and predictions of both homogeneity and sub-homogeneity and can correctly identify the clustering of the original high-dimensional data-set.

Thirdly, Following the idea in Chapter 3 on combining similar high-dimensional data for aggregated analysis, Chapter 4 extends the study to high-dimensional time series. In contrast to the new method CPCA proposed in Chapter 3, which effectively estimates both homogeneity and sub-homogeneity for combined data, Chapter 4 studies whether two high-dimensional time series data have the same spiked eigenstructure of the autocovariance matrices. In particular, a novel equivalence test named autocovariance test on spiked eigenvalues of autocovariance matrices is proposed in Chapter 4 for comparing two high-dimensional time series. Due to the 'curse of dimensionality', traditional statistical testing methods based on sample estimates of covariance and auto-covariance matrices generally fail to work when the data dimension N grows to infinity. To overcome the issues associated with the 'curse of dimensionality', the high-dimensional time

series data is assumed to follow factor models as in Chapter 2. Moreover, the test statistic is built based on a central limit theorem (CLT) for spiked eigenvalues of the sample autocovariance matrices for high-dimensional time series, which is in my joint work with others (Bi et al., 2020). Meanwhile, to implement the test procedure, the sieve bootstrap proposed in Chapter 2 is also employed for estimating specific unknown parameters. Consequently, this autocovariance test facilitates combined analysis for high-dimensional time series from multiple populations. For example, the autocovariance test can be applied to the age-specific mortality data for multiple countries to test whether human mortality rates have the same spiked eigenvalues in autocovariance matrices for countries worldwide. This work is also inspiring since the aggregated analysis may improve the estimation accuracy and provide more consistent forecastings for human mortality rates worldwide.

The rest of this thesis is organised as follows. Three aforementioned statistical inferences are proposed in Chapter 2 to 4 in order, where statistical models, implementations, asymptotic results, simulation studies, and real data applications are all addressed with discussions. Besides, the conclusions and future works are presented in Chapter 5. Furthermore, technical proofs of theorems in Chapter 2 to 4 and some auxiliary lemmas are included in appendices. It is also worth noting that each chapter uses its own notations.

Chapter 2

Sieve Bootstrap for High-dimensional Time Series: A Factor Model Approach

2.1 Introduction

In this chapter, we propose a novel sieve bootstrap method for high-dimensional time series. Studying statistical inferences for mean, variance and many other statistics is a major problem for modern statistics. For high-dimensional data especially time series, deriving the theoretical properties including the central limit theorem for certain statistics such as leading eigenvalues of covariance matrix and autocovariance matrices can be rather involved especially when the data dimension goes to infinity. Bootstrap (Efron, 1979), as an alternative, has become more attractive when pursuing statistical inferences for both independent and dependent data. However, as discussed in El Karoui and Purdom (2018), conventional nonparametric bootstrap does not work in general for high-dimensional data including time series due to the 'curse of dimensionality', where sample estimates of certain statistics are no longer statistically consistent to their population counterparts. Consequently, novel bootstrap methods that can be applied on high-dimensional time series is in urgent demand.

The first contribution of this work is that we develop a sieve bootstrap method for high-dimensional time series, where a factor model is introduced. Since being developed, the conventional nonparametric bootstrap (Efron, 1979) has become

very popular in studying statistical inferences as it is not only accurate but also easy to apply on real data. However, unlike for independent data, the conventional nonparametric bootstrap (Efron, 1979) is not readily applicable for time series, since the temporal dependence of the observations cannot be correctly mimicked by simple resampling. Consequently, a few variations and modifications have been made on the conventional bootstrap method for the purpose of applying bootstrap for dependent data. And among all bootstrap methods for time series, block-wise bootstrap methods (Kunsch, 1989) and autoregressive (AR) sieve bootstrap methods (Kreiss, 1988; Bühlmann, 1997) have received the most discussions and developments in the past few years. The AR sieve bootstrap was introduced by Kreiss (1988) and has been well studied from the case of linear time series (Bühlmann, 1997) to strictly stationary time series fulfilling a general moving average (MA) (∞) representation (Kreiss et al., 2011). Prior to this work, the theoretical requirement and validity of a general AR sieve bootstrap method for certain type of statistics have been discussed for univariate (Kreiss et al., 2011), multivariate (Meyer and Kreiss, 2015) and functional time series (Paparoditis, 2018), respectively. However, for high-dimensional time series where the data dimension N increases with the sample size T, the AR sieve bootstrap method is not readily applicable. This is because the dimension of the spectral density matrix of underlying multivariate time series diverges with data dimension N, and an infinite order vector AR or MA representation do not exist (see Wold theorem in Anderson (1971), and boundedness condition in Wiener and Masani (1958)). Besides, for real data applications where the dimension N of observed time series is not required to go to infinity, estimating high-dimensional coefficient matrices in an one-side AR or MA representation of the original time series is still very complicated and time-consuming. Recently, Krampe et al. (2019) consider sieve bootstrap for VAR model of linear time series where the VAR coefficients are assumed to be sparse, while we consider a different set-up in this work, where the dimension N of time series is allowed grow but the observations are assumed to fulfil a (strong) factor model with finite number of factors. Factor models (Bai and Ng, 2002b; Bai, 2003b), are originally

introduced and developed to study large dimensional data with $N, T \to \infty$ through dimension reduction techniques. Later on, Fan et al. (2011) utilises factor models to estimate a large covariance matrix where the covariance of error terms is assumed to be approximately sparse. To study the temporal dependence of a factor model, Lam et al. (2011) propose a factor model which can be estimated based on the accumulated squared autocovariance matrices. Consequently, a finite-dimensional factor process is then developed to explore the dependence structure of the original high-dimensional time series. In our work, we consider the strong factors' case in Lam et al. (2011), where the spiked eigenvalues of accumulated squared (first k_0) autocovariance matrices of high-dimensional time series $\{\boldsymbol{y}_t\}$ are of order N. In summary, we propose an AR sieve bootstrap for high-dimensional time series using a factor model approach. The proposed AR sieve bootstrap using factor model is not only an efficient statistical method for studying inference of high-dimensional time series but also an indispensable building block of AR sieve bootstrap methods under high-dimensional set-up.

The second contribution of this work is that we compare the proposed novel sieve bootstrap for high-dimensional time series with the sieve bootstrap method for functional time series (Paparoditis, 2018) in terms of their applications on sparse and unsmoothed functional observations. And we suggest that the sparse and unsmoothed observations need to be treated as high-dimensional time series and the sieve bootstrap proposed in this work needs to be applied. In the literature of functional time series studies, a very fundamental assumption is that the actual observations come from a smoothed functional curve and statistical inferences for functional data usually require the observations to be dense. In a classic functional set-up, dense and discrete points are observed on a sample of T curves. Denoted by N_t the number of observations for the curve t, the discussions on the density of observations in functional data literature are generally through assumptions made on N_t. Typically, when N_t is much larger than the sample size T, the data can be considered dense functional data where each curve can be well smoothed before analysis. Discussions on the density of functional observations and smoothing methods can be found in Ramsay

and Silverman (2002), Hall et al. (2006) and Wang and Fan (2017). However, in the case where N_t is small compared with sample size T for all t, the discrete observations should be considered as sparse along the population functional curve (Wang and Fan, 2017). The fundamental problem of sparse functional data is that the local patterns of population functional curve are generally not captured by those sparse observations.

To illustrate the potential problems of pre-smoothing sparse observations for functional time series analysis, we consider a toy example. For a square-integrable functional process $\{\mathcal{X}(u),\ u \in \mathcal{I}\}$, let $y_{i,t}$ be the i-th observation of $\{\mathcal{X}_t(\cdot)\}$, observed at a random time t with the measurement errors defined as $\epsilon_{i,t}$ for $t = 1, 2, ..., T$ and $i = 1, 2, ..., N$. Consider now a model of functional observations

$$y_{i,t} = \mathcal{X}_t(u_i) + \epsilon_{i,t}, \quad u_i \in \mathcal{I}, \tag{2.1}$$

where $\epsilon_{i,t}$ is independent and identically distributed (i.i.d.) with $\mathbb{E}(\epsilon_{i,t}) = 0$, $\mathbb{V}(\epsilon_{i,t}) = \sigma^2$ and $\mathcal{I}$ is a functional support. In this model, the observations of $\{\mathcal{X}_t(\cdot)\}$ are assumed to be equally spaced, and the number of measurements N assesses the density and design of the actual observations. In the functional data analysis, $\mathcal{X}_t(u_i)$ can be estimated or recovered by some smoothing methods such as a linear smoother as follows,

$$\widehat{\mathcal{X}_t}(u_i) = \sum_{j=1}^{N} w_i(u_j) y_{t,i},$$

where $w_i(u_j)$ is the weight of j-th point on the i-th point with $\sum_{j=1}^{N} w_i(u_j) = 1$ for $t = 1, 2, ..., T$ and $i = 1, 2, ..., N$. Various smoothing methods have been developed for functional data, and Ramsay and Silverman (2002) study choosing smoothing basis for different types of functional data. Besides, Yao et al. (2005) compare the functional data with longitude data and discuss the impact of pre-smoothing on a functional model, and Zhang and Wang (2016) extend this discussion to investigate the asymptotic properties of local linear smoothers on various types

of sampling designs. However, the accuracy of the smoothing curve is highly related to the density of observations and measurement errors. If observations along the curve are equally spaced, the change of density can affect the quality of smoothness and its recovering power to the population curve. For a relatively sparse curve, smoothing can fail to work under certain situations; for example, when there are local patterns that observations are too sparse to capture. To visually depict this phenomenon, we provide a toy example by simulations in Section 2.5. We consider a contaminated functional time series model generated from three Fourier bases with different frequencies reflecting local patterns. The details of the simulation setting can be found in Section 2.5.1. The curves in Figure 2.1 are plotted based on 401 grid points defined on a functional support $[0, 1]$, whereas the actual number of observations N along each curve are chosen as 51, 21 and 5 to address different observation densities. As shown in Figure

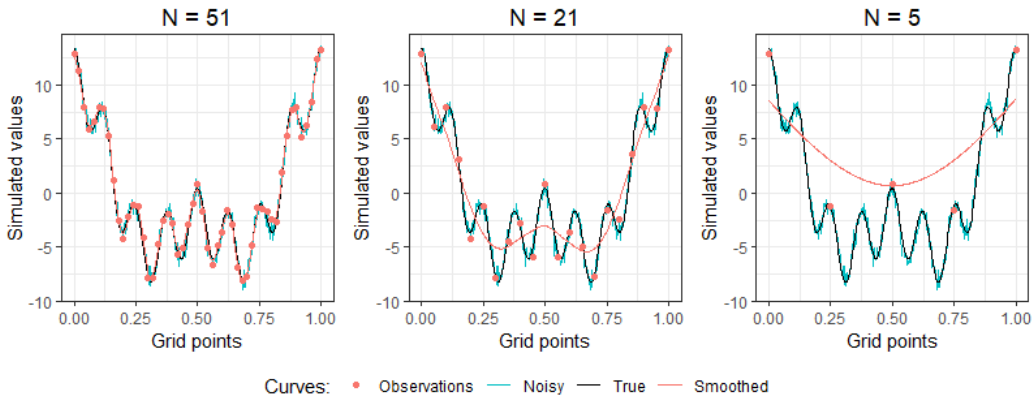


Figure 2.1: Example of smoothing error of sparse functional time series observations

2.1, when the observations (red points) become sparse (but still equally spaced), the (red) smoothing curve can lead to an obvious misleading result with local patterns not accurately captured by the smoothing curve. The errors associated with pre-smoothing on those sparse observations are generally large. In this situation, the assumption of dense functional data suffers from insufficient observations along each curve. As a result, we cannot adopt the pre-smoothing results based on functional set-up but instead treat the data as multivariate time series with growing dimensions. In other words, when N grows with sample size

T but at a relatively slower rate, the real data may adapt to a high-dimensional set-up rather than a functional set-up, which makes statistical inferences and applications rather different. This phenomenon is associated with an area where functional data analysis and high-dimensional data analysis may overlap yet follow different assumptions and produce quite different asymptotic results.

In contrast to functional data analysis, where the increase of observations along a curve can practically improve pre-smoothing and recovering the functional curve, the growing of dimensions is associated with the increase of complexity for high-dimensional data analysis. This key difference makes it vital to choose between functional time series and high-dimensional time series methods. In this work, we consider the situation where N is growing but not fast enough. The curve smoothed from the sparse observations is inaccurate, especially to local patterns of a functional curve. We propose an AR sieve bootstrap method for studying the inferences of this type of high-dimensional time series. The rest of this chapter is organised as follows. Section 2.2 introduces factor models for high-dimensional time series and discusses the AR representation of the factor time series, a building block of general AR sieve bootstrap. In Section 2.3, the estimation procedure for factor models and sieve bootstrap procedure for factor time series is introduced with regularity conditions on factor models. The additional assumptions and asymptotic validity of our novel sieve bootstrap method are discussed in Section 2.4. An overall mean statistics of factor time series and spiked eigenvalues of squared autocovariance matrices are introduced. In Section 2.5, we first explore the impact of density of observations on pre-smoothing results and then verify the validity of our novel sieve bootstrap methods on the overall mean statistics and the spiked eigenvalues of squared autocovariance matrices. Section 2.6 provides an example of applying our novel sieve bootstrap method to PM_{10} data. Conclusions are presented in Section 2.7. Technical proofs and auxiliary lemmas are presented in Appendixes 2.A and 2.B in additional supplementary documents.

2.2 Model

In this chapter, we study a situation where the actual observations from functional time series are sparse such that smoothing methods generally fail to work, especially for local patterns of the raw functional curve. Hence, the sparse observations we considered are high-dimensional time series instead of functional time series. In general, bootstrap methods fail to work when the dimension N of the time series grows to infinity. To effectively overcome this problem and reduce the dimension for bootstrap, we consider a strictly stationary N-dimensional time series $\{\boldsymbol{y}_t \in \mathbb{R}^N,\ t \in \mathbb{Z}\}$ following a general unobservable factor model

$$\boldsymbol{y}_t = \boldsymbol{Q}\boldsymbol{f}_t + \boldsymbol{u}_t, \tag{2.2}$$

where $\{\boldsymbol{f}_t \in \mathbb{R}^r,\ t \in \mathbb{Z}\}$ are $r \times 1$ unobserved finite-dimensional factor time series and $\{\boldsymbol{u}_t \in \mathbb{R}^r,\ t \in \mathbb{Z}\}$ are $N \times 1$ white noises with mean zero and covariance matrix $\boldsymbol{\Sigma}_u$. Unobservable (approximate) factor models have received numerous discussions and there are various identification condition and assumptions on $\boldsymbol{Q}$, $\boldsymbol{f}_t$ and $\boldsymbol{u}_t$ depending on objects. In our work, we adapt the idea in Lam et al. (2011) to consider a factor model where temporal dependence of $\{\boldsymbol{y}_t\}$ can be fully captured by the factors $\{\boldsymbol{f}_t\}$ with a constant factor loading matrix $\boldsymbol{Q}$. In other words, we do not allow for a direct dynamic system on $\{\boldsymbol{f}_t\}$, therefore we still maintain a static relationship between $\{\boldsymbol{y}_t\}$ and $\{\boldsymbol{f}_t\}$. In addition, to establish a sieve bootstrap method which mimic the temporal dependence of the original data, we also adapt the assumptions used in literature of factor models, such as Bai and Ng (2002b) and Bai (2003b) where the high-dimensional noise components $\{\boldsymbol{u}_t\}$ are independent of the finite-dimensional factors $\{\boldsymbol{f}_t\}$. The exact assumptions and conditions for estimation of this factor model is introduced and discussed in Section 2.3.

For the $r \times 1$ unobserved factors $\{\boldsymbol{f}_t\}$, under certain assumptions that will be specified in Section 2.4.1, we can consider it to admit a general (vector) AR representations. That is, there exists an infinite sequence of $r \times r$ matrices

$\{\boldsymbol{A}_l \in \mathbb{R}^{r\times r},\ l \in \mathbb{N}\}$ such that factors $\{\boldsymbol{f}_t\}$ can be expressed as

$$\boldsymbol{f}_t = \sum_{l=1}^{\infty} \boldsymbol{A}_l \boldsymbol{f}_{t-l} + \boldsymbol{e}_t,\ t \in \mathbb{Z}, \tag{2.3}$$

where $\{\boldsymbol{e}_t \in \mathbb{R}^r,\ t \in \mathbb{Z}\}$ are full rank uncorrelated white noises innovation processes with $\mathbb{E}(\boldsymbol{e}_t) = 0$ and $\mathbb{E}(\boldsymbol{e}_t \boldsymbol{e}_s^\top) = \mathbf{1}_{t=s}\boldsymbol{\Sigma}_e$, with $\boldsymbol{\Sigma}_e$ a full rank $r \times r$ covariance matrix. The (vector) AR representation is also the AR analogue of the Wold representation of $\boldsymbol{f}_t$, and it is represented by a MA representation based on a function of the same innovation processes $\{\boldsymbol{e}_t \in \mathbb{R}^r,\ t \in \mathbb{Z}\}$ as in (2.3):

$$\boldsymbol{f}_t = \sum_{l=1}^{\infty} \boldsymbol{\Psi}_l \boldsymbol{e}_{t-l} + \boldsymbol{e}_t,\ t \in \mathbb{Z}, \tag{2.4}$$

where $\{\boldsymbol{\Psi}_l \in \mathbb{R}^{r\times r},\ l \in \mathbb{N}\}$ are the coefficients matrices of the power series $\left(\boldsymbol{I}_r - \sum_{l=1}^{\infty} \boldsymbol{A}_l z^l\right)^{-1}$, for $|z| \leq 1$ (Brockwell and Davis, 1991). The (vector) AR representation in (2.3) is more attractive for statistical applications and has received more attentions since it relates $\boldsymbol{f}_t$ to its past values. Sieve bootstrap, on the other hand, utilises the Wold representation in (2.4) to generate bootstrap factors by resampling from the innovations $\boldsymbol{e}_t$. In practice, since neither the factors $\{\boldsymbol{f}_t\}$ or their loadings $\boldsymbol{Q}$ are observable, sieve bootstrap for sparse time series is performed on estimates of $\{\boldsymbol{f}_t\}$ rather than true factors. Hence, we need to introduce the estimation and bootstrap procedure first.

2.3 Estimation and bootstrap procedure

2.3.1 Estimation of factor models

Since $\{\boldsymbol{f}_t\}$ in model (2.3) are assumed to contain all the temporal dependence of $\{\boldsymbol{y}_t\}$, we can utilise and modify the idea in Lam et al. (2011) to estimate $\{\boldsymbol{f}_t\}$. Define the accumulated squared autocovariance of $\{\boldsymbol{y}_t\}$ up to a prescribed lag $k_0 > 0$ as

$$L = \sum_{k=1}^{k_0} \boldsymbol{\Gamma}_y(k) \boldsymbol{\Gamma}_y(k)^\top, \tag{2.5}$$

where $\boldsymbol{\Gamma}_y(k) = \text{Cov}(\boldsymbol{y}_t, \boldsymbol{y}_{t+k})$ is the autocovariance of $\{\boldsymbol{y}_t\}$ at lag k, for $k = 1, 2, ..., k_0$. L then collects the temporal dependence of $\{\boldsymbol{y}_t\}$ by pooling up the information contained in first k_0-lags of autocovariance with the square form facilitating the spectral decomposition on L.

Remark 2.1. The reason of not to consider the covariance matrix $\boldsymbol{\Sigma}_y$ into L is undemanding. As discussed in Lam et al. (2011), for the factor model (2.2), $\boldsymbol{\Sigma}_y = \boldsymbol{\Gamma}_y(0) = \boldsymbol{Q}\boldsymbol{\Gamma}_f(0)\boldsymbol{Q}^\top + \boldsymbol{\Sigma}_u$, where $\boldsymbol{\Gamma}_f(0)$ is the covariance matrix of $\{\boldsymbol{y}_t\}$ and $\boldsymbol{\Sigma}_u$ is the covariance matrix of $\{\boldsymbol{u}_t\}$. Hence to exclude $\boldsymbol{\Sigma}_y$ from L can filter out the impact of covariance on $\{\boldsymbol{u}_t\}$, especially for $N \to \infty$.

It is then straightforward to use spectral (eigenvalue) decomposition on L to estimate the factor loading matrix $\boldsymbol{Q}$, and the factors $\{\boldsymbol{f}_t\}$ from L. Before discussing the estimation procedure details, we summarise the assumptions and identification conditions for the factor model defined in (2.2) first. Recall that none of the terms in (2.2) are observable except $\{\boldsymbol{y}_t\}$, we need the following conditions to identify and estimate factors $\{\boldsymbol{f}_t\}$ and their corresponding loading matrix $\boldsymbol{Q}$.

Assumptions 2.1 (Factor models)**.** *For factor models (2.2), we impose the following assumptions,*

(i) $\{\boldsymbol{f}_t\}$ *are strictly stationary with* $\mathbb{E}\boldsymbol{f}_t = \boldsymbol{0}$ *and* $\mathbb{E}\left\|\boldsymbol{f}_t\right\|^2 < \infty$; $\{\boldsymbol{u}_t\} \sim WN(0, \boldsymbol{\Sigma}_u)$ *are uncorrelated white noise with covariance matrix* $\boldsymbol{\Sigma}_u$*, and all eigenvalues of* $\boldsymbol{\Sigma}_u$ *are uniformly bounded as* $N \to \infty$*;* $\boldsymbol{f}_t$ *are independent of* $\boldsymbol{u}_s$ *for any* $t, s \in \mathbb{Z}$*.*

(ii) (Identification for $\{\boldsymbol{Q}\boldsymbol{f}_t\}$ *and* $\{\boldsymbol{f}_t\}$*)* $\frac{1}{N}\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r$ *and for a prescribed integer* $k_0 > 0$*, the* $r \times r$ *matrices* $\boldsymbol{\Gamma}_f(k) = Cov(\boldsymbol{f}_t, \boldsymbol{f}_{t+k})$ *are full rank for all* $k = 0, 1, ..., k_0$ *with the eigenvalues* $\{\lambda_i(\boldsymbol{f}),\ i = 1, 2, ..., r\}$ *of* $\sum_{k=1}^{k_0} \boldsymbol{\Gamma}_f(k)\boldsymbol{\Gamma}_f(k)^\top$ *fulfilling* $\infty > \lambda_1(\boldsymbol{f}) \geq \lambda_2(\boldsymbol{f}) \geq \cdots \geq \lambda_r(\boldsymbol{f}) > 0$ *as* $N \to \infty$*.*

(iii) $\{\boldsymbol{y}_t\}$*, therefore* $\{\boldsymbol{f}_t\}$*, is* ψ*-mixing with the mixing coefficients* $\psi(\cdot)$ *satisfying the condition that* $\sum_{t\geq 1} \psi(t)^{1/2} < \infty$*, and* $\mathbb{E}|y_{j,t}|^4 < \infty$ *element-wisely.*

Assumption 2.1 (i) states the strict stationarity on $\{\boldsymbol{f}_t\}$, which has been used in literature of factor models, such as Fan et al. (2013b) and is commonly seen

in AR sieve bootstrap literature, such as Kreiss et al. (2011) and Meyer and Kreiss (2015). Apart from the stationarity, Assumption 2.1 (i) also states that factor time series $\{\boldsymbol{f}_t\}$ and error terms $\{\boldsymbol{u}_t\}$ are independent at any time lags, which is stronger than the assumption in Lam et al. (2011), but is required for us to apply bootstrap methods by resampling from the innovations $\{\boldsymbol{e}_t\}$ in Wold representation of $\{\boldsymbol{f}_t\}$ as in (2.4), since sieve bootstrap does not work for high-dimensional noises $\{\boldsymbol{u}_t\}$.

We impose Assumption 2.1 (ii) to identify the factor components $\{\boldsymbol{Q}\boldsymbol{f}_t\}$ from the original high-dimensional data. The conditions that $\frac{1}{N}\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r$ and eigenvalues $\{\lambda_i(\boldsymbol{f}),\ i = 1, 2, ..., r\}$ of $\sum_{k=1}^{k_0} \boldsymbol{\Gamma}_f(k)\boldsymbol{\Gamma}_f(k)^\top$ fulfil $\infty > \lambda_1(\boldsymbol{f}) \geq \lambda_2(\boldsymbol{f}) \geq \cdots \geq \lambda_r(\boldsymbol{f}) > 0$ as $N \to \infty$ are sufficient for $\{\boldsymbol{Q}\boldsymbol{f}_t\}$ to be identifiable from $\{\boldsymbol{u}_t\}$ when $N \to \infty$, since the $N \times N$ matrix L can be represented as

$$L = \sum_{k=1}^{k_0} \boldsymbol{\Gamma}_y(k)\boldsymbol{\Gamma}_y(k)^\top = N\boldsymbol{Q}\left\{\sum_{k=1}^{k_0} \boldsymbol{\Gamma}_f(k)\boldsymbol{\Gamma}_f(k)^\top\right\}\boldsymbol{Q}^\top, \tag{2.6}$$

with the first r eigenvalues of $\frac{1}{N^2}L$ non-vanishing. In other words, the columns of $\boldsymbol{Q}$ can be considered as the eigenvectors of L corresponding to r nonzero eigenvalues scaled by $\sqrt{N}$. As a consequence, Assumption 2.1 (ii) implies the pervasiveness of r factors $\{\boldsymbol{f}_t\}$ when N goes to infinity, which is equivalent to the strong factors' case according to the definition in Lam et al. (2011).

The ψ-mixing in Assumption 2.1 (iii) is introduced to specify the week dependence structure of $\{\boldsymbol{f}_t\}$, which is considered in Lam et al. (2011) to simplify the technical proof of consistency on loading matrix $\boldsymbol{Q}$. However, it is not the weakest possible. In the meantime, Assumption 2.1 (ii) together with the mixing condition in (iii) is also sufficient for the absolute summability condition on $\{\boldsymbol{f}_t\}$ when $N \to \infty$, which is preliminary for AR sieve bootstrap to be applicable on $\{\boldsymbol{f}_t\}$, since otherwise the Wold representation is not guaranteed to exist (Cheng and Pourahmadi, 1993).

To further explain the use of Assumption 2.1 with the estimation procedure, first notice that $\{\boldsymbol{f}_t\}$ are strong factors and no linear combinations of the components of $\{\boldsymbol{f}_t\}$ are white noises (WN) as implied by Assumption 2.1 (ii). Recall that

L is non-negative definite and can be represented as in (2.6) with $\frac{1}{N}\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r$. Since the middle part of (2.6) is symmetric, we can apply spectral decomposition on it and recognize L as $N\boldsymbol{Q}\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{Q}^\top$ where $\boldsymbol{U}$ is an $r \times r$ orthonormal matrix and $\boldsymbol{D}$ is an $r \times r$ diagonal matrix. Therefore, by Assumption 2.1 (ii), we can treat $\boldsymbol{Q}\boldsymbol{U}$ as $\boldsymbol{Q}$ for inferences' purpose and estimate $\boldsymbol{Q}$ and $\boldsymbol{f}_t$ based on the spectral decomposition of L.

Remark 2.2. The factor model discussed in Lam et al. (2011) assumes the temporal dependence of a high-dimensional time series $\{\boldsymbol{y}_t\}$ can be captured by the factor time series $\{\boldsymbol{f}_t\}$. In general, let $\boldsymbol{P}$ be a $N \times (N-r)$ matrix in which the columns are orthogonal to those of $\boldsymbol{Q}$, then $\boldsymbol{\Gamma}_y(k)^\top\boldsymbol{P} = 0$. Hence as long as $\boldsymbol{\Gamma}_f(k)$ is full rank, $\mathcal{M}(\boldsymbol{Q})$ is the orthogonal complement to the linear space spanned by the eigenvectors of $\boldsymbol{\Gamma}_y(k)^\top$ that are associated with those zero eigenvalues (i.e. the eigenvectors spanning $\mathcal{M}(\boldsymbol{P})$ (Lam et al., 2011)). This also justifies the use of autocovariance but not covariance when estimating this factor time series, since $\boldsymbol{\Sigma}_y = \boldsymbol{\Gamma}_y(0) = \boldsymbol{Q}\boldsymbol{\Gamma}_f(0)\boldsymbol{Q}^\top + \boldsymbol{\Sigma}_u$ and $\boldsymbol{\Gamma}_y(0)^\top\boldsymbol{P} \neq 0$. The fact that $\boldsymbol{\Gamma}_y(k)\boldsymbol{\Gamma}_y(k)^\top$ is non-negative definite for all $k = 1, 2, ..., k_0$ guarantees that the columns of $\boldsymbol{P}$ are those orthogonal eigenvectors of L corresponding to zero eigenvalues and the sum in (2.5) pools up the temporal dependence of $\{\boldsymbol{y}_t\}$ from different time lags.

With such regularity conditions in Assumptions 2.1, we can estimate the factors and their loadings, and construct a pseudo-time series with AR sieve bootstrap. To facilitate the estimation process, we define $\boldsymbol{Q}^o = \frac{1}{\sqrt{N}}\boldsymbol{Q}$ as the (unscaled) orthonormal factor loading matrix such that $\boldsymbol{Q}^{o\top}\boldsymbol{Q}^o = \boldsymbol{I}_r$ and $\boldsymbol{f}_t^o$ as the scaled factors such that $\boldsymbol{y}_t = \boldsymbol{Q}^o\boldsymbol{f}_t^o + \boldsymbol{u}_t$ is equivalent to model (2.2) with different scaling on $\boldsymbol{Q}$ and $\{\boldsymbol{f}_t\}$. The detailed estimation and bootstrap procedure of our proposed method is illustrated as follows.

2.3.2 Bootstrap procedures

Step 1: Estimation of $\boldsymbol{Q}$:

To utilise the idea in Lam et al. (2011) to estimate $\boldsymbol{Q}$ and $\{\boldsymbol{f}_t\}$ using L, the accumulated squared autocovariance matrices of $\{\boldsymbol{y}_t\}$ up to a prescribed lag

$k_0 > 0$, we first define the sample accumulation of squared autocovariance up to lag k_0 as

$$\widetilde{L} = \sum_{k=1}^{k_0} \widetilde{\mathbf{\Gamma}}_y(k)\widetilde{\mathbf{\Gamma}}_y(k)^\top, \tag{2.7}$$

with $\widetilde{\mathbf{\Gamma}}_y(k)$ the sample autocovariance at lag k defined as

$$\widetilde{\mathbf{\Gamma}}_y(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(\boldsymbol{y}_{t+k} - \overline{\boldsymbol{y}})(\boldsymbol{y}_t - \overline{\boldsymbol{y}})^\top.$$

By applying spectral (eigenvalue) decomposition on $\widetilde{L}$, we can obtain $\widehat{\boldsymbol{Q}^o} = (\widehat{\boldsymbol{q}_1^o}, \widehat{\boldsymbol{q}_2^o}, ..., \widehat{\boldsymbol{q}_r^o})$ with $\widehat{\boldsymbol{q}_i^o}$ the eigenvector of $\widetilde{L}$ corresponding to the i-th largest eigenvalue of $\widetilde{L}$. $\widehat{\boldsymbol{Q}^o}$ is then a natural estimator of the unscaled loading matrix $\boldsymbol{Q}^o$. And by scaling up $\widehat{\boldsymbol{Q}^o}$ with $\sqrt{N}$, the square root of dimension, we ended up with $\widehat{\boldsymbol{Q}} = \sqrt{N}\widehat{\boldsymbol{Q}^o}$ as the estimator of $\boldsymbol{Q}$.

As discussed in Lam et al. (2011), the estimation results are not sensitive to the choice of k_0, and the numeral results associated with $k_0 = 1$ to $k_0 = 5$ are similar. In general, when dimension N is large compared with T, a relatively larger k_0 may be considered for better accuracy of sample estimates, while $k_0 = 1$ is computational more efficient when the sample size T is large compared with dimension N. Besides, for finite samples, some of the non-spiked eigenvalues of $\widetilde{L}$ may not be exactly zero, therefore we can use the ratio-based estimator as discussed in Lam et al. (2011) to estimate the number of factor r. As defined in Lam et al. (2011), the ratio-based estimator for r is

$$\widehat{r} = \underset{1\le j\le R}{\operatorname{argmin}}\, \widehat{\lambda}_{j+1}/\widehat{\lambda}_j,$$

with $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \cdots \ge \widehat{\lambda}_N$ the eigenvalues of $\widetilde{L}$ and R an integer satisfying $r \le R < N$. And practically, R can be taken as $N/2$ or $N/3$ for computation efficiency (Lam et al., 2011).

Step 2: Estimation of $\{\boldsymbol{f}_t\}$:

With $\widehat{\boldsymbol{Q}}$ the estimator of $\boldsymbol{Q}$, it is then straightforward to estimate $\{\boldsymbol{f}_t\}$ by

$$\widehat{\boldsymbol{f}}_t = \widehat{\boldsymbol{Q}}^\top \boldsymbol{y}_t.$$

Step 3: Sieve bootstrap on $\{\widehat{\boldsymbol{f}}_t\}$:

To apply sieve bootstrap on $\{\widehat{\boldsymbol{f}}_t\}$, we can, first of all, fit a pth order VAR model on the r-dimensional time series $\{\widehat{\boldsymbol{f}}_t\}$ as

$$\widehat{\boldsymbol{f}}_t = \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p}(r)\widehat{\boldsymbol{f}}_{t-l} + \widehat{\boldsymbol{e}}_{t,p},\ t = p+1, p+2, ..., T,$$

and denote the residuals by $\widehat{\boldsymbol{e}}_{t,p}$,

$$\widehat{\boldsymbol{e}}_{t,p} = \widehat{\boldsymbol{f}}_t - \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p}(r)\widehat{\boldsymbol{f}}_{t-l},\ t = p+1, p+2, ..., T,$$

where $\{\widehat{\boldsymbol{A}}_{l,p},\ l = 1, 2, ..., p;\ t = p+1, p+2, ..., T\}$ are Yule-Walker estimators of the AR coefficient matrices. We can then generate $\{\boldsymbol{e}_t^*\}$ the bootstrap sample of residuals by resampling from the empirical distribution of the centered residual vectors. Consequently, based on the idea of sieve bootstrap (see, e.g. Kreiss, 1992; Meyer and Kreiss, 2015; Paparoditis, 2018), we can generate the r-dimensional pseudo time series $\{\boldsymbol{f}_t^*,\ t = 1, 2, ..., T\}$ by simulating the VAR model with bootstrap residuals $\{\boldsymbol{e}_t^*\}$. Therefore, a sieve bootstrap sample of $\{\boldsymbol{f}_t^*\}$ is generated by

$$\boldsymbol{f}_t^* = \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p}(r)\boldsymbol{f}_{t-l}^* + \boldsymbol{e}_t^*,$$

where $\{\boldsymbol{e}_t^*\}$ are i.i.d. random vectors following the empirical distribution of the centered residual vectors $\{\widetilde{\boldsymbol{e}}_t\}$, where $\widetilde{\boldsymbol{e}}_{t,p} = \widehat{\boldsymbol{e}}_{t,p} - \overline{\widehat{\boldsymbol{e}}}_{T,p}$ and $\overline{\widehat{\boldsymbol{e}}}_{T,p} = 1/(T-p)\sum_{t=p+1}^{T} \widehat{\boldsymbol{e}}_{t,p}$.

Step 4: Generating bootstrap data $\{\boldsymbol{y}_t^*\}$:

Lastly, the bootstrap time series $\{\boldsymbol{y}_t^*\}$ can be constructed as

$$\boldsymbol{y}_t^* = \sum_{j=1}^{r} \boldsymbol{f}_{j,t}^* \widehat{\boldsymbol{q}}_j,$$

where $\widehat{\boldsymbol{q}}_j = \sqrt{N}\widehat{\boldsymbol{q}}_j^0$ is the scaled eigenvector of $\widehat{\boldsymbol{L}}$ corresponding to the jth largest eigenvalue. Following this sieve bootstrap procedure, the pseudo-time series $\{\boldsymbol{y}_t^*\}$ can mimic the temporal dependence of the original data $\{\boldsymbol{y}_t\}$ via a factor model. Notice that, when N is fixed and relatively small, another bootstrap procedure such as wild bootstrap can be applied on $\widehat{\boldsymbol{u}}_t = \boldsymbol{y}_t - \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{f}}_t$ when generating $\{\boldsymbol{y}_t^*\}$. However, when N goes to infinity, traditional bootstrap methods generally fail to work for the high-dimensional noises $\{\boldsymbol{u}_t\}$. Therefore, it is generally not valid to apply bootstrap on $\{\widehat{\boldsymbol{u}}_t\}$ when generating $\{\boldsymbol{y}_t^*\}$. As a consequence, $\{\boldsymbol{y}_t^*\}$ correctly mimic the temporal dependence of $\{\boldsymbol{y}_t\}$ through the factors $\{\boldsymbol{f}_t\}$ but not the noises $\{\boldsymbol{u}_t\}$ in the factor model (2.2).

Remark 2.3. Since traditional bootstrap procedures are not valid for high-dimensional noises $\{\boldsymbol{u}_t\}$, our sieve bootstrap time series $\{\boldsymbol{y}_t^*\}$ do not contain bootstrap noises. As a result, $\{\boldsymbol{y}_t^*\}$ can only provide valid inferences for statistics that are temporal dependent or not depending on $\{\boldsymbol{u}_t\}$ since $\{\boldsymbol{u}_t\}$ are independent of $\{\boldsymbol{f}_t\}$ and $\boldsymbol{Q}$. For statistics relying on $\{\boldsymbol{u}_t\}$, such as $\boldsymbol{\Gamma}_y(0)$, the covariance matrix of $\{\boldsymbol{y}_t\}$, our sieve bootstrap method is not valid since $\boldsymbol{\Sigma}_u$ cannot be mimicked.

2.4 Asymptotic theory

2.4.1 Regularity assumptions

Before introducing the additional regularity assumptions, we fix some notations first. We use $\|\cdot\|_2$ to denote the L_2 norm (also known as spectral norm or operator norm) of a matrix or vector, and $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. And we use $a \asymp b$ to denote the case that $a = O_P(b)$ and $b = O_P(a)$.

In addition to Assumptions 2.1 made on the factor model (2.2), to apply sieve bootstrap on $\{\widehat{\boldsymbol{f}}_t\}$, the estimates of factors $\{\boldsymbol{f}_t\}$, we also need some regularity conditions on $\{\boldsymbol{f}_t\}$ for sieve bootstrap to be consistent and valid. Denoted by $W(\cdot)$, the spectral density matrix of a vector process for all frequencies $\omega \in (0, 2\pi]$, then the spectral density matrix of $\{\boldsymbol{f}_t\}$ can be defined as

$$W_f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_f(k) e^{-i\omega k}, \omega \in (0, 2\pi].$$

Assumptions 2.2. *In model (2.2), we strengthen Assumption 2.1 such that* $\{\boldsymbol{f}_t\}$ *are strictly stationary and purely nondeterministic stochastic processes of full rank with* $\mathbb{E}\boldsymbol{f}_t = \mathbf{0}$ *and* $\mathbb{E}\left\|\boldsymbol{f}_t\right\|^2 < \infty$. $\boldsymbol{\Gamma}_f(k)$*, the autocovariance matrix of* $\boldsymbol{f}_t$ *at lag* k *fulfils the matrix norm summability condition* $\sum_{k=-\infty}^{\infty}(1+|k|)^{\gamma}\left\|\boldsymbol{\Gamma}_f(k)\right\|_F < \infty$ *for some* $\gamma \geq 0$ *that will be specified later on.*

Assumption 2.2 is introduced to fulfil the requirement for the existence of a general VAR representation (2.3). This type of conditions are commonly seen in sieve bootstrap literature, such as Kreiss et al. (2011) and Meyer and Kreiss (2015), and it is worth noting that $\{\boldsymbol{e}_t\}$, the innovation processes in the VAR representation of factor processes $\{\boldsymbol{f}_t\}$, are linear independent as $\{\boldsymbol{f}_t\}$ are full rank (Wiener and Masani, 1958). In addition, following the heredity of mixing properties in Assumption 2.1, $\{\boldsymbol{f}_t\}$ are strict stationary and also $\psi-$mixing, which in turn implies the decaying of $\boldsymbol{\Gamma}_f(k)$ as $k \to \infty$. The matrix norm summability condition on $\boldsymbol{\Gamma}_f(k)$, as in Assumption 2.2, then specifies the rate of decaying that is required for a vector AR representation to be valid as stated in the next Lemma. Besides, since we assume the dimension r of $\{\boldsymbol{f}_t\}$ is finite, $\mathbb{E}\boldsymbol{f}_t = 0$ in Assumption 2.1 is made on a multivariate factor process, therefore, will not affect the results we derived on the consistency and validity of the proposed sieve bootstrap method. The assumption $\mathbb{E}\boldsymbol{f}_t = 0$ can be relaxed to $\mathbb{E}\boldsymbol{f}_t = \mu_f$ with the cost of a more lengthy proof of theorems in this work.

Lemma 2.1. *Under Assumption 2.1 and 2.2 with* $\gamma = 0$*, the spectral density matrix* $W(\cdot)$ *of* $\{\boldsymbol{f}_t\}$ *fulfils the following so-called Boundedness Condition (Wiener and Masani,*

1958):

$$c\boldsymbol{I}_r \leq W_f(\omega) \leq d\boldsymbol{I}_r, \text{ for all } \omega \in (0, 2\pi],$$

where $\boldsymbol{I}_r$ *is the unity matrix with dimension* $r \times r$ *and* $0 < c \leq d < \infty$.

Proof of Lemma 2.1. The upper bound $d\boldsymbol{I}_r$ for all $\omega \in (0, 2\pi]$ follows directly from the norm summability condition stated in Assumption 2.2. The assumption of strong factors in Assumption 2.1 implies the positivity on eigenvalues of the spectral density matrix $W(\cdot)$. Therefore, denoted by $\sigma_i(\omega)$, the minimum eigenvalue of $W_f(\omega)$ for $i = 1, 2, ..., r$, then $\sigma_i(\omega)$ is continuous in $(0, 2\pi]$ and strictly positive. Denoted by $\sigma_{\min} = \min_{\omega \in (0,2\pi]}(\sigma_i(\omega))$, the minimum eigenvalue of the spectral density matrix of $\{\boldsymbol{f}_t\}$, then there exists a constant $c > 0$ so that $\sigma_{\min} \geq c$ for all frequencies $\omega \in (0, 2\pi]$. □

The continuity and Boundedness properties in Lemma 2.1 then entail the existence of a vector AR representation for any vector process satisfying Assumption 2.2 (see, e.g. Meyer and Kreiss, 2015; Cheng and Pourahmadi, 1993; Wiener and Masani, 1958). That is, the AR representation (2.3) and Wold representation (2.4) are valid under Assumption 2.2.

The validity of sieve bootstrap on a class of strictly stationary vector series fulfilling Assumption 2.2 has been discussed in Meyer and Kreiss (2015), where some additional conditions on the convergence of Yule-Walker estimators of the finite predictor coefficients on $\{\boldsymbol{f}_t\}$ are also introduced. We summarise these conditions in Assumption 2.3 and leave the results of Meyer and Kreiss (2015) to Lemma 2.6 in Appendix B, as they are preliminary for showing the bootstrap consistency and validity.

Assumptions 2.3. *The Yule-Walker estimators* $\{\widetilde{\boldsymbol{A}}_{l,p}, l = 1, 2, ..., p\}$ *of* $\{\boldsymbol{A}_{l,p}, l = 1, 2, ..., p\}$, *the finite predictor coefficients matrices on the VAR model of* $\{\boldsymbol{f}_t\}$, *fulfils that* $p^2 \sum_{l=1}^{p} \|\widetilde{\boldsymbol{A}}_{l,p} - \boldsymbol{A}_{l,p}\|_F = O_P(1)$, *as* $T \to \infty$ *and* $p \to \infty$.

Assumption 2.3 requires $p \to \infty$ at a relatively slower rate of sample size T, which is required for the convergence of the Yule-Walker estimator of $\boldsymbol{A}_p =$

$(\boldsymbol{A}_{1,p}, ..., \boldsymbol{A}_{p,p})$. In other words, the order p of the AR terms in sieve bootstrap depends on the sample size T and has to be chosen properly. For $\{\boldsymbol{f}_t\}$ fulfilling Assumption 2.2, Assumption 2.3 is also satisfied if we choose $p = O\left((T/\ln T)^{1/6}\right)$, for example (Meyer and Kreiss, 2015). Assumptions 2.2 and 2.3 are widely discussed in literature of sieve bootstrap, for example, in Kreiss et al. (2011) and Meyer and Kreiss (2015). In summary, Assumption 2.2 ensures the existence of VAR representation in (2.3) and specifies the rate of decaying for the coefficient matrices and Assumption 2.3 relates to the convergence of Yule-walker estimators $\{\widetilde{\boldsymbol{A}}_{l,p}\}$ to the finite predictor coefficient matrices $\{\boldsymbol{A}_{l,p}\}$.

Assumptions 2.4. *The dimension N and AR(p) satisfy $N \to \infty$, $p \to \infty$ when $T \to \infty$ such that $p^{11/2}(N^{-1/2} + T^{-1/2}) \to 0$.*

In addition to Assumption 2.3, Assumption 2.4 is introduced as the bootstrap procedure is performed on the estimated factors $\{\widehat{\boldsymbol{f}}_t\}$ rather than true unobservable factors $\{\boldsymbol{f}_t\}$, where the error comes from both the estimation of factors and finite order approximation of sieve AR representations. In other words, we need to control the error imposed by the bootstrap procedure by restricting the speed that the AR order p goes to infinity. On the other hand, the order on dimension N in Assumption 2.4 also indicates 'blessing of dimensionality', since $\{\boldsymbol{f}_t\}$ are assumed to be strong factors according to definitions in Lam et al. (2011).

2.4.2 Bootstrap validity for generalised mean statistics

One of the most fundamental problems in functional data analysis is to estimate the mean function from observations with noises. Some methods and applications can be found, for example, in Ramsay and Silverman (2002) and Cai and Yuan (2011). Under the setting where observations are generally sparse, statistical inferences for (general) mean statistics of high-dimensional data is also fundamental.

The validity of general AR and VAR sieve bootstrap has been well discussed in Kreiss et al. (2011) and Meyer and Kreiss (2015). The key idea is that the general AR and VAR sieve bootstrap doesn't mimic the behaviour of the underlying

processes in (2.3) or (2.4), but the behaviour of a so-called companion processes $\{\check{\boldsymbol{f}}_t\}$ defined in the same form as $\{\boldsymbol{f}_t\}$ but with i.i.d. white noises $\{\check{\boldsymbol{e}}_t\}$ rather than $\{\boldsymbol{e}_t\}$, where $\mathcal{L}(\check{\boldsymbol{e}}_t) = \mathcal{L}(\boldsymbol{e}_t)$ though $\{\boldsymbol{e}_t\}$ are only uncorrelated for $t \neq s$. In other words, except for the Gaussian case, the general AR and VAR sieve bootstrap works for statistics that only depend on up-to-second-order quantities of $\{\boldsymbol{f}_t\}$. Without additional assumptions on the distribution of $\{\boldsymbol{e}_t\}$, the higher-order properties of $\{\check{\boldsymbol{f}}_t\}$ and $\{\boldsymbol{f}_t\}$ are not necessarily the same. With this result, the bootstrap consistency for a class of general mean statistics stated below have been investigated by Kreiss et al. (2011) and Meyer and Kreiss (2015). Based on this result, we can study statistical inferences for such class of statistics based on the unobservable factor terms $\boldsymbol{Q}$ and $\{\boldsymbol{f}_t\}$ by bootstrapping $\{\widehat{\boldsymbol{f}}_t\}$. The following general mean statistics is introduced by Kunsch (1989), and has been widely discussed in sieve bootstrap literature such as Bühlmann (1997), Kreiss et al. (2011) and Meyer and Kreiss (2015). Consider for a class of general mean statistics

$$M_T = \eta\left(\frac{1}{T-m+1}\sum_{t=1}^{T-m+1} g(\boldsymbol{f}_t, ..., \boldsymbol{f}_{t+m-1})\right), \tag{2.8}$$

for functions $g : \mathbb{R}^{mr} \to \mathbb{R}^d$ and $\eta : \mathbb{R}^d \to \mathbb{R}$ for some $d \geq 1$ and $1 \leq m \leq T$, fulfilling the following smoothness condition: $\eta(\boldsymbol{z})$ has continuous partial derivatives for all $\boldsymbol{z}$ in a neighbourhood of $\theta = \mathbb{E}g(\boldsymbol{f}_t, ..., \boldsymbol{f}_{t+m-1})$ and the differentials $\sum_{i=1}^{d} \partial\eta/\partial z_i|_{z=\theta} z_i$ do not vanish. The function g has continuous partial derivatives of order h $(h \geq 1)$ that satisfy a Lipschitz condition for all $i = 1, 2, ..., d$. We summarise our first result on bootstrap consistency of $\boldsymbol{Q}\overline{\boldsymbol{f}_T}$, the mean statistics of the unobservable factor component $\{\boldsymbol{Q}\boldsymbol{f}_t\}$, in the following Theorem.

Theorem 2.1. *Suppose that Assumptions 2.1, 2.2 ($\gamma = 1$), 2.3 and 2.4 are satisfied for fixed and known number of factors r. In addition, if we further assume that*

(a) $\mathbb{E}\left(e_{j,t}^{2(h+2)}\right) < \infty$ *for each element* $e_{j,t}$ *in* $\{\boldsymbol{e}_t\}$, *(see (2.8) for the definition of h).*

(b) The empirical distribution of $\{\boldsymbol{e}_t\}$ *converges weakly to the distribution function of* $\mathcal{L}(\boldsymbol{e}_t)$.

(c) $\lim_{T\to\infty} \mathbb{V}(\sqrt{T}\overline{\boldsymbol{f}_T}) = \sum_{k\in\mathbf{Z}} \boldsymbol{\Gamma}_f(k) > 0$.

Then, for any vector $\boldsymbol{c} \in \mathbb{R}^N$ *such that* $\|\boldsymbol{c}^\top \boldsymbol{Q}\|_{\ell_1} < \infty$ *and* $0 < \sum_{k\in\mathbf{Z}} \boldsymbol{c}^\top \boldsymbol{Q}\boldsymbol{\Gamma}_f(k)\boldsymbol{Q}^\top \boldsymbol{c} < \infty$ *as* $N \to \infty$, *we can conclude that*

$$d_K\left(\mathcal{L}^*\left(\sqrt{T}\boldsymbol{c}^\top \widehat{\boldsymbol{Q}}\left(\overline{\boldsymbol{f}_T^*} - \mathbb{E}^*\overline{\boldsymbol{f}_T^*}\right)\right), \mathcal{L}\left(\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\left(\overline{\boldsymbol{f}_T} - \mathbb{E}\overline{\boldsymbol{f}_T}\right)\right)\right) \xrightarrow{p} 0,$$

when $N \to \infty$ *and* $T \to \infty$, *where* d_K *denotes the Kolmogorov distance.*

Theorem 2.1 states the validity of proposed sieve bootstrap methods on estimated factors $\{\widehat{\boldsymbol{f}}_t\}$. In general, statistical inferences such as bootstrap standard errors or bootstrap confidence intervals can be computed for a linear combination of factor components $\{\boldsymbol{Q}\boldsymbol{f}_t\}$, which makes inferences on the mean level of originally high-dimensional time series possible. On the other hand, the bootstrap inferences can be considered an alternative statistical tool for practical use compared with the asymptotic results, which can be rather difficult to derive. The factor model in (2.2) filtered out the time-invariant noises $\{\boldsymbol{u}_t\}$ and reduced the dimension of $\{\boldsymbol{y}_t\}$. In turn, the factor model facilitates the development of the sieve bootstrap procedure.

Remark 2.4. As discussed in Kreiss et al. (2011) and Meyer and Kreiss (2015), AR sieve bootstrap in fact mimics the behavior of a companion process $\check{\boldsymbol{f}}_t$ which shares the same first and second-order quantities as $\{\boldsymbol{f}_t\}$. Hence for the mean statistics, AR sieve bootstrap works without any additional assumptions made on the higher-order moments of $\{\boldsymbol{f}_t\}$. Also, for sieve bootstrap to be asymptotically valid on $\{\boldsymbol{f}_t\}$, the dimension r needs not to go to infinity. Therefore we imposed the assumption on strong factors. To study the impact of factor strength on the validity of AR sieve bootstrap, we also consider weak factors in simulation studies in Section 2.5.

2.4.3 Bootstrap consistency for autocovariance matrices

For high-dimensional i.i.d. data, the covariance matrix plays an important role in dimension reduction techniques, such as factor models and principal component analysis. However, for high-dimensional dependent data, the autocovariance

matrices are vital or even more crucial than the covariance matrix. Lam et al. (2011) provides a discussion on the use of autocovariance in dimension-reduction techniques. Therefore, it is critical to find the bootstrap consistency for the autocovariance matrices under the proposed sieve bootstrap method. In the next theorem, we show that the proposed sieve bootstrap method can guarantee the asymptotic consistency on the autocovariance matrices, which in turn implies the validity of using bootstrap data $\{\boldsymbol{y}_t^*\}$ to approximate the original data $\{\boldsymbol{y}_t\}$.

Recall that $\boldsymbol{\Gamma}_f(k) = \mathrm{Cov}(\boldsymbol{f}_t, \boldsymbol{f}_{t+k})$ is the autocovariance of unobservable factor $\{\boldsymbol{f}_t\}$ at lag k, for $k > 0$. Without the loss of generality, we again assume the means of factors are 0 to simplify the notations used in the next theorem. Define $\boldsymbol{\Gamma}_f^*(k) = \mathrm{Cov}(\boldsymbol{f}_t^*, \boldsymbol{f}_{t+k}^*)$ as the autocovariance of bootstrap factor $\{\boldsymbol{f}_t^*\}$ at lag k, then we have the following theorem on the asymptotic consistency of $\boldsymbol{\Gamma}_f^*(k)$.

Theorem 2.2. *Suppose that Assumptions 2.1, 2.2 ($\gamma = 1$) and 2.3 are satisfied for fixed and known number of factors r. In addition, if we further assume that*

(a) $\mathbb{E}\left(e_{j,t}^{2(h+2)}\right) < \infty$ *for each element $e_{j,t}$ in $\{\boldsymbol{e}_t\}$, (see (2.8) for the definition of h).*

(b) The empirical distribution of $\{\boldsymbol{e}_t\}$ converges weakly to the distribution function of $\mathcal{L}(\boldsymbol{e}_t)$.

Then for $k \in \mathbb{N}$, we have

$$\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_f(k)\right\|_2 \xrightarrow{p} 0,$$

when $N \to \infty$ and $T \to \infty$.

Let $\{\delta_i(k)\}_{i=1}^r$ be the ordered spiked eigenvalues of $\frac{1}{N^2}\boldsymbol{\Gamma}_y(k)\boldsymbol{\Gamma}_y(k)^\top$, the squared autocovariance matrices of $\{\boldsymbol{y}_t\}$ at lag $k > 0$. And define $\{\delta_i^*(k)\}_{i=1}^r$ to be the first r largest eigenvalues of $\frac{1}{N^2}\boldsymbol{\Gamma}_y^*(k)\boldsymbol{\Gamma}_y^*(k)^\top$, the bootstrap squared autocovariance matrices of $\{\boldsymbol{y}_t^*\}$ at lag $k > 0$, where $\boldsymbol{\Gamma}_y^*(k) = \mathrm{Cov}^*(\boldsymbol{y}_t^*, \boldsymbol{y}_{t+k}^*)$. As a consequence of Theorem 2.2, we immediately have the following Proposition on the convergence of spiked eigenvalues of the bootstrap squared autocovariance matrices to their population counterparts.

Proposition 2.1. *Under the same Assumptions of Theorem 2.2, for* $i = 1, 2, ..., r$ *and* $k \in \mathbb{N}$, *we have*

$$\left\| \mathbf{\Gamma}_y^*(k) - \mathbf{\Gamma}_y(k) \right\|_2 \xrightarrow{p} 0, \tag{2.9}$$

and

$$|\delta_i^*(k) - \delta_i(k)| \xrightarrow{p} 0, \tag{2.10}$$

when $N \to \infty$ *and* $T \to \infty$.

The study of spiked eigenvalues of squared autocovariance matrices of high-dimensional time series is necessary but significant in many applications. However, there are very few inference tools available in the literature due to the difficulties and complexities of studying dependent data when $N \to \infty$. Proposition 2.1 verifies the bootstrap consistency on spiked eigenvalues of squared autocovariance matrices and provides statistical tools to study the properties of spiked eigenvalues based on sieve bootstrap.

Remark 2.5. The results of Proposition 2.1 are on the whole probability space, which allows for the use of autocovariances and their spiked eigenvalues computed from a bootstrap sample $\{\boldsymbol{y}_t^*\}$ to approximate the autocovariances and corresponding spiked eigenvalues of the original data $\{\boldsymbol{y}_t\}$.

2.5 Simulation studies

In this section, we first compare the proposed sieve bootstrap method's performances under functional time series assumptions and multivariate (high-dimensional) time series assumptions. We then study the sieve bootstrap confidence intervals for the general mean statistics and eigenvalues of the squared autocovariance matrix by evaluating the empirical coverage probability. Lastly, we also examine the proposed sieve bootstrap method's performance when the factors $\{\boldsymbol{f}_t\}$ are assumed to be weak, and the dimension N goes to infinity. This

is particularly important for statistical inferences of high-dimensional factor modelling.

2.5.1 Smoothing on sparse discrete functional time series

To study the impact of smoothing on the sparse functional time series observations, we can compare bootstrap samples' empirical distributions under various densities of observations. To start, we first assume the data are originated from functional curves, which are temporal dependent. Recall model (2.1) that

$$y_{t,i} = \mathcal{X}_t(u_i) + \epsilon_{t,i}, \quad u_i \in \mathcal{I},$$

where $\epsilon_{t,i}$ is i.i.d. with $\mathbb{E}(\epsilon_{t,i}) = 0$ and $\mathbb{V}(\epsilon_{t,i}) = \sigma^2$, for $t = 1, 2, ..., T$ and $i = 1, 2, ..., N$. In this model, the number of measurements N reflect the density of the actual observations. To study the impact of density, we assume the observations are equally spaced and generated from a three factors' model

$$\boldsymbol{y}_t = \boldsymbol{Q}\boldsymbol{f}_t + \boldsymbol{u}_t,$$

where $u_{t,i}$, the element in $\{\boldsymbol{u}_t\}$, is independent $\mathcal{N}(0,1)$ random noise, $\boldsymbol{Q}$ is a $N \times 3$ matrix with each column a Fourier basis and $\cos(2\pi i/N)$, $\cos(4\pi i/N)$, $0.5\cos(16\pi i/N)$ as i-th element, respectively. The factors $\{\boldsymbol{f}_t\}$ follows a VAR(1) model with a coefficient matrix

$$\begin{bmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{bmatrix}$$

and errors independent simulated from $\mathcal{N}(0,1)$. The Fourier basis is selected to produce a smoothed population curve, with the third basis reflecting local patterns. Hence, we can generate discrete observations from a functional curve with local patterns. In Section 2.1, we have presented plots of $\{\boldsymbol{y}_t\}$ at a particular time t with three different densities of observations to illustrate a smoothing's

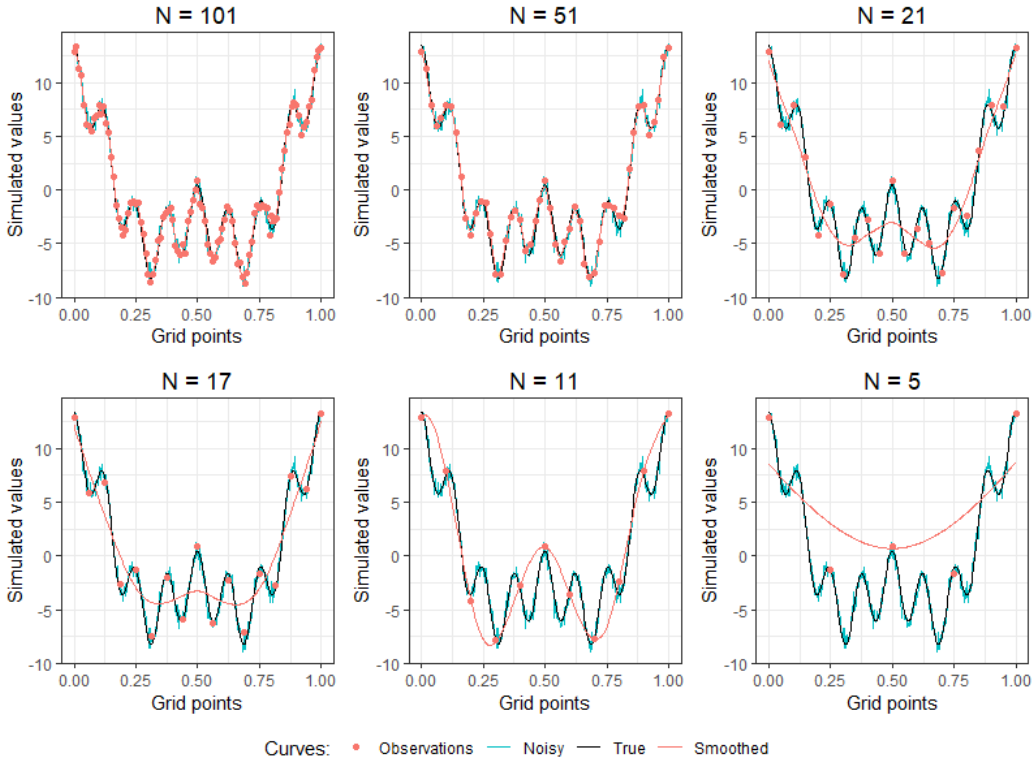


Figure 2.2: Example of smoothing errors on sparse functional observations

potential issue. This section takes it one step further and considers a wider choice of densities so that the actual dimensions of observations along each curve are $N = 101, 51, 21, 17, 11$ and 5.

For the same choice of time t as in Section 2.1, we have generated 6 plots under various densities in Figure 2.2 to compare the smoothing results with the population true curve and noisy curve with small measurement errors. The smoothing results are obtained using B-splines with the number of basis functions set to N, the actual number of observations in each case, and the roughness penalties selected based on generalised cross-validation (GCV). As depicted in Figure 2.2, when the actual number of observations N is relatively small, for example, $N < 21$, some local patterns of the population curve are generally not captured. In addition, the smoothing curve sometimes also averaged out the actual observations to achieve relatively flat results, for example, when $N = 21, 17$ and 5 as in Figure 2.2. As a result, the observations after smoothing are generally less spread than the original observations, which produces very different

bootstrap samples and inferences' results. To see that, we generate $B = 499$ sieve bootstrap samples and computed two summary statistics to compare the bootstrap distribution based on original observations with smoothed observations. We use sieve bootstrap to obtain estimates of a so-called (standardised) overall mean statistic, computed as $\overline{\overline{y^*}} = \frac{\sqrt{T}}{\sqrt{N}} \mathbf{1}^\top \widehat{\boldsymbol{Q}} \overline{\boldsymbol{f}^*}$ according to Theorem 2.1, and δ_1^*, the estimate of (standardised) largest eigenvalue of squared lag-1 sample autocovariance matrix as defined in Proposition 2.1, to compare bootstrap samples from original observations with bootstrap samples from pre-smoothed observations.

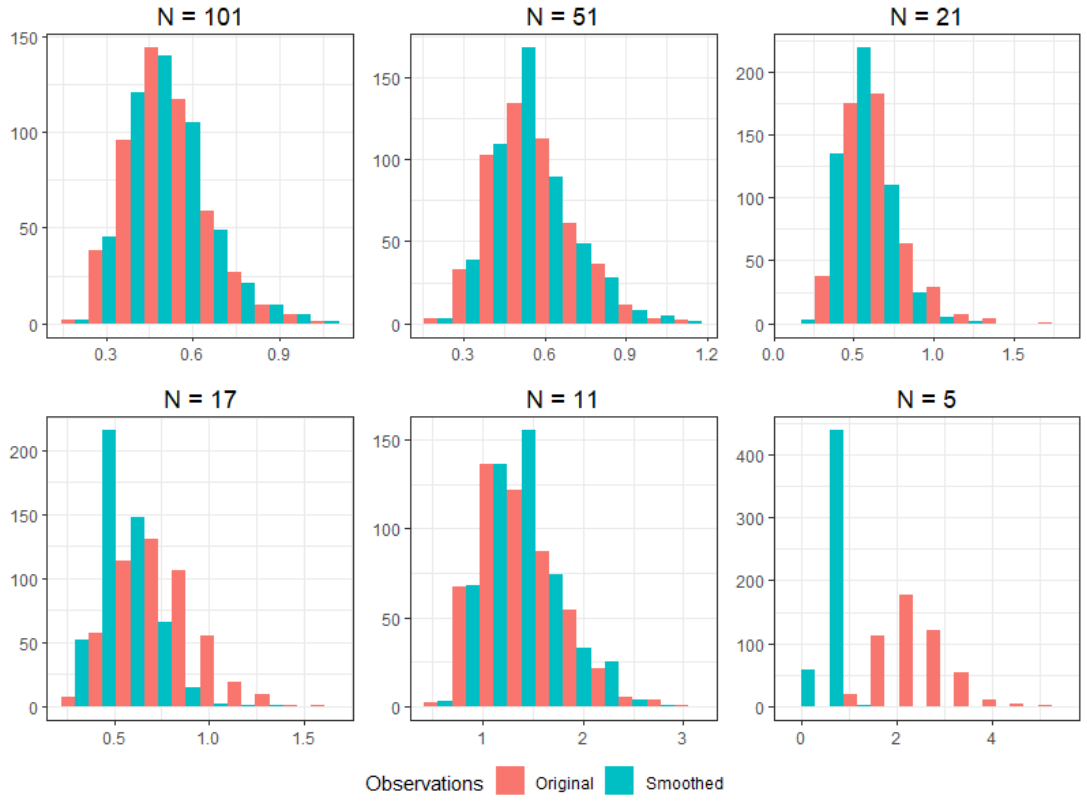


Figure 2.3: Histograms of δ_1^*, the sieve bootstrap estimates of the largest eigenvalue of squared lag-1 sample autocovariance matrix

Figures 2.3 and 2.4 compare the histograms and boxplots of δ_1^*, the sieve bootstrap estimates of largest eigenvalue of squared lag-1 autocovariance matrix, while Figures 2.5 and 2.6 compare the histograms and boxplots of $\overline{\overline{y^*}}$, the sieve bootstrap estimates of overall mean statistic. As seen in Figure 2.2, when $N = 21, 17$ and 5, the pre-smoothed observations are averaged out compared with the original observations. As a result, the bootstrap estimates of the two

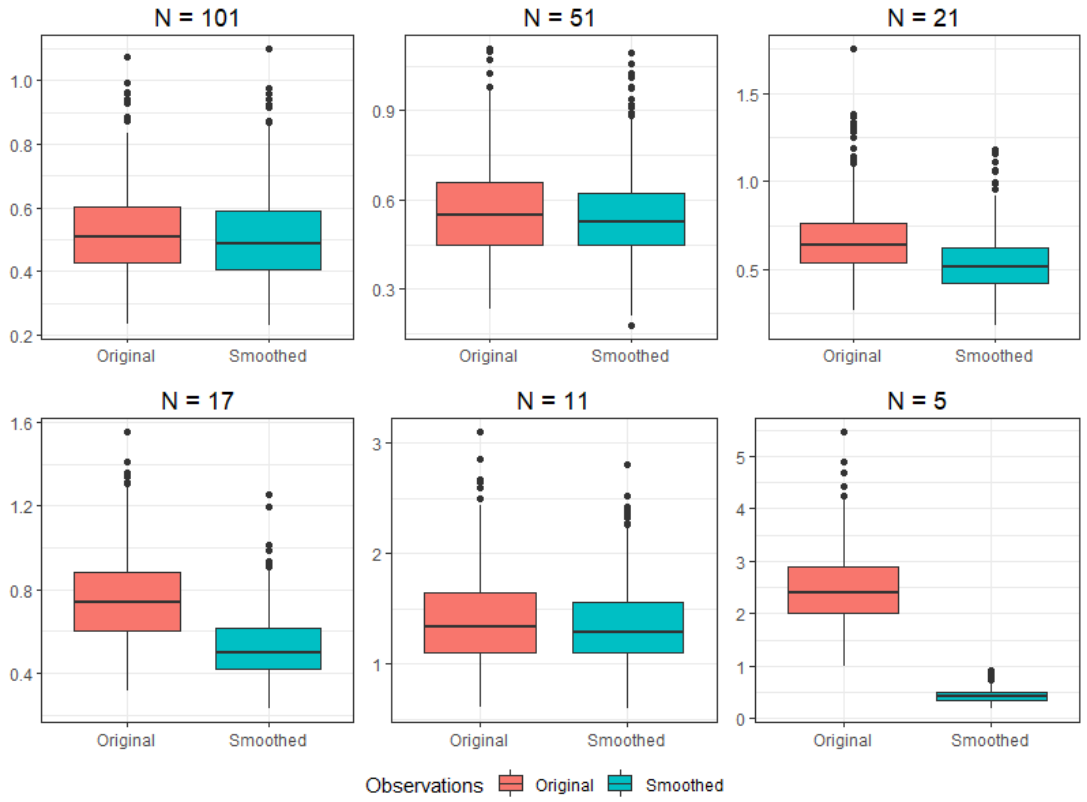


Figure 2.4: Boxplots of δ_1^*, the sieve bootstrap estimates of the largest eigenvalue of squared lag-1 sample autocovariance matrix

statistics perform differently before and after smoothing, when $N = 21, 17$ and 5. Figures 2.3 and 2.5 use boxplots to present the difference of empirical distributions of $\overline{\overline{y^*}}$ and δ_1^* for $N = 21, 17$ and 5, whereas Figures 2.4 and 2.6 illustrate the impact of smoothing by comparing the histograms of $\overline{\overline{y^*}}$ and δ_1^*.

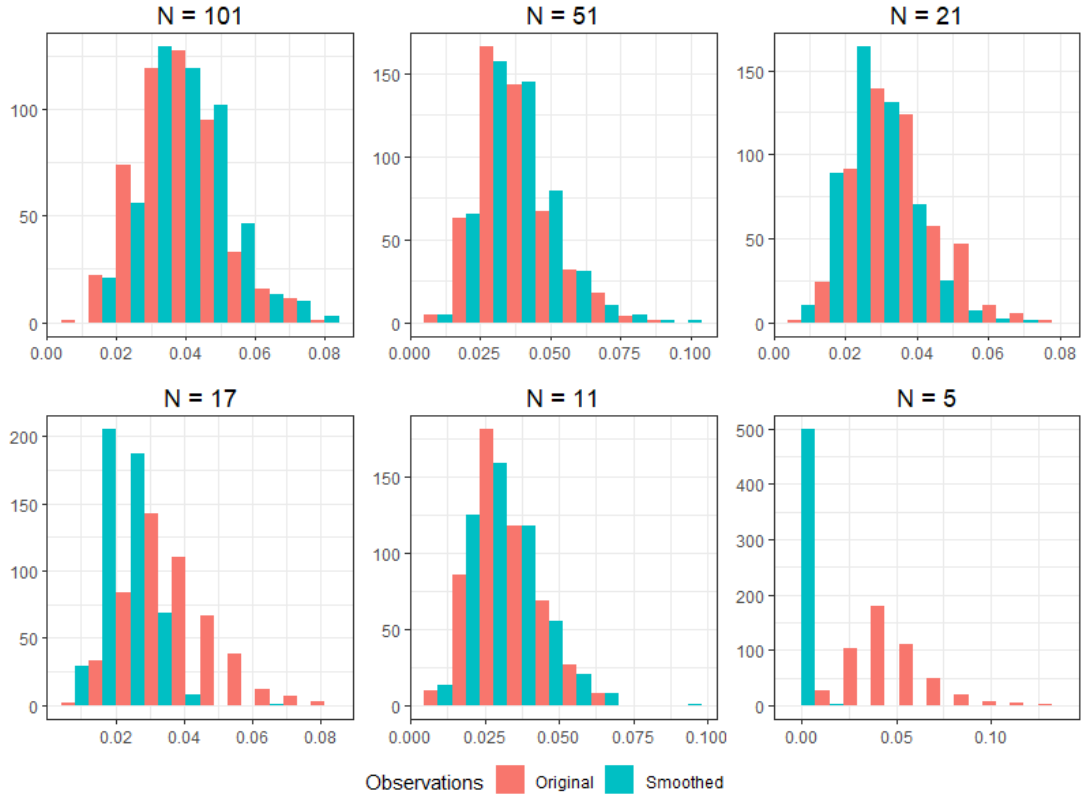


Figure 2.5: Histograms of $\overline{\overline{y^*}}$, the sieve bootstrap estimates of overall mean statistic

The last example we presented in Figure 2.7 illustrates results of sieve bootstrap estimates (bootstrap average) of the functional mean curve when we pre-smooth the observations under various densities of data. As shown in Figure 2.7, when the actual observations are relatively dense, for example, $N \geq 51$, sieve bootstrap estimates of the mean functional curve are close to the pre-smoothed curve and the population curve. However, when the observations are sparse, for example, $N \leq 21$, sieve bootstrap estimates of the mean functional curve do not correctly capture the local patterns of the population curve, which is due to the unacceptable smoothing results. This result is also typical evidence of the impact of pre-smoothing on sieve bootstrap for functional time series. Hence, when

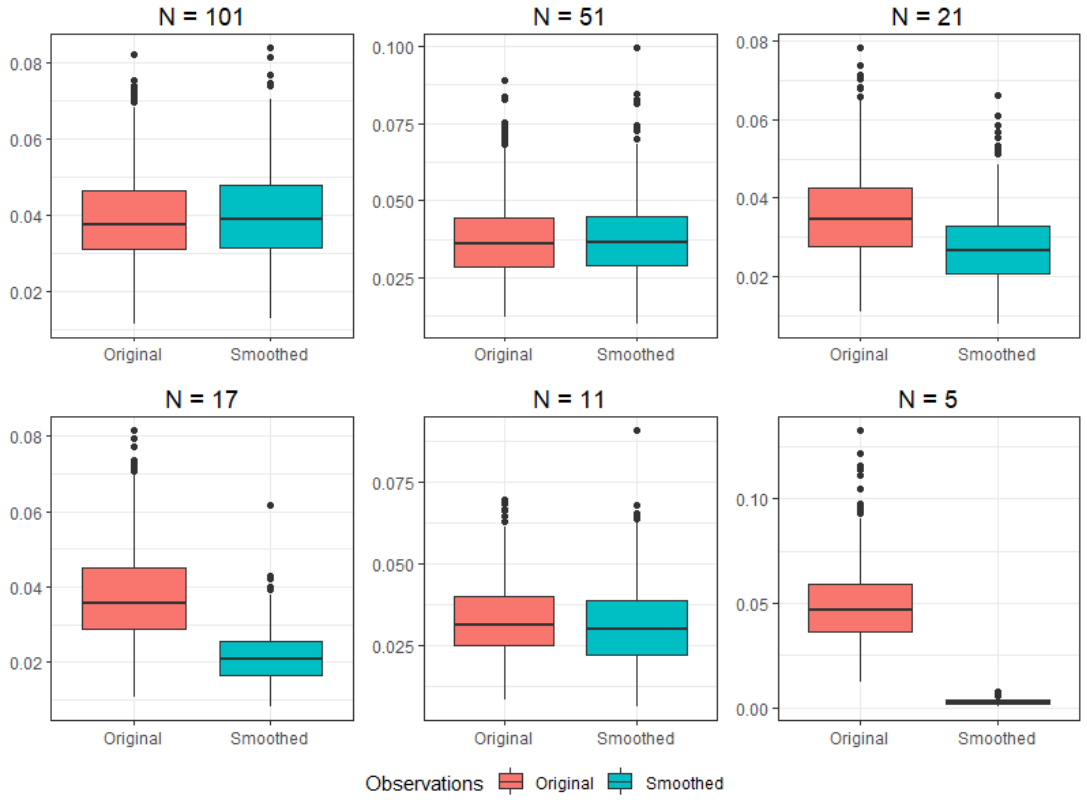


Figure 2.6: Boxplots of $\overline{\overline{y^*}}$, the sieve bootstrap estimates of overall mean statistic

the actual functional time series observations are sparse, pre-smoothing may significantly impact statistical inferences, including bootstrap. In fact, for many real-world time series data, the rule on considering a data set as dense functional time series is generally not clear and often varies across researchers and problems. Practically speaking, the impact of observations' density is only about whether to pre-smooth the functional time series before performing bootstrap or other statistical analysis.

Nonetheless, the theoretical assumptions behind functional time series and high-dimensional time series vary, leading to very different theoretical results on statistical inferences, including sieve bootstrap. On the other hand, this difference in data structure assumptions demonstrates the importance of developing statistical methods on sparse functional time series observations. It verifies our contributions on the building blocks of sieve bootstrap for high-dimensional time series.

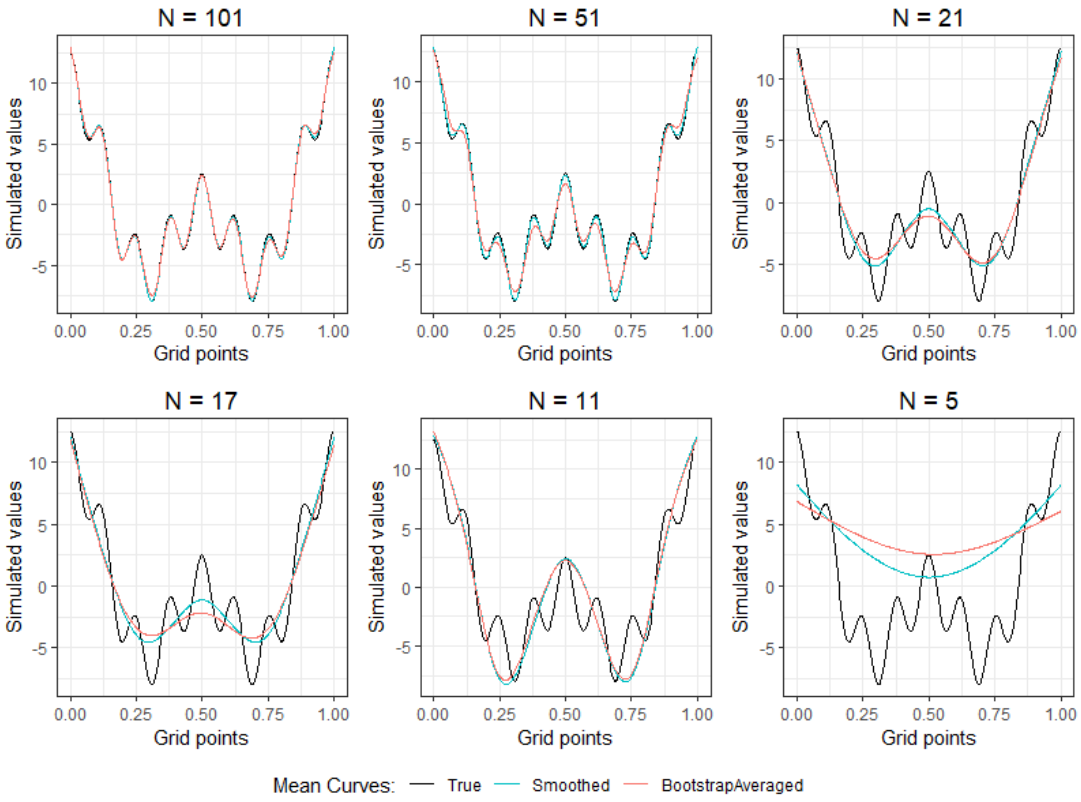


Figure 2.7: Example of errors of sieve bootstrap mean curve for sparse functional observations

2.5.2 Sieve bootstrap for mean statistic

We study the validity and consistency of our proposed sieve bootstrap method for high-dimensional factor time series models. To achieve this, we use simulation to evaluate the empirical coverage and average width of bootstrap confidence intervals for the overall mean statistics defined in Theorem 2.1 first. Recall model (2.2) that $\boldsymbol{y}_t = \boldsymbol{Q}\boldsymbol{f}_t + \boldsymbol{u}_t$ and its equivalent form $\boldsymbol{y}_t = \boldsymbol{Q}^o\boldsymbol{f}_t^o + \boldsymbol{u}_t$ with different scales on $\boldsymbol{Q}$ and $\boldsymbol{f}_t$. To address the problem under a general high-dimensional factor time series model, we no longer assume the original data are generated from a functional curve as in Section 2.5.1. Instead, we generate the factor loading matrix $\boldsymbol{Q}^o$ by an arbitrary QR decomposition on standard multivariate normal random variables, where $\boldsymbol{Q}^o$ fulfils $\boldsymbol{Q}^{o\top}\boldsymbol{Q}^o = \boldsymbol{I}_r$ with r the number of factors but is not necessarily a smoothing basis. We then assume the observations $\{\boldsymbol{y}_t\}$ are from a two factors model $\{\boldsymbol{y}_t\} = \boldsymbol{Q}^o\boldsymbol{f}_t^o + \boldsymbol{u}_t$, where $\{u_{i,t}\}$ are independent $\mathcal{N}(0,1)$ random noises, $\boldsymbol{Q}$ is a $N \times 2$ matrix with each column an orthogonal basis, and both factors of $\boldsymbol{f}_t^o$ are strong factors following an AR(1) model with mean 0 and the AR coefficient 0.5. In other words, both factors are generated from $f_{i,t} = 0.5f_{i,t-1} + e_{i,t}$, with $i = 1, 2$.

To study the impact of factor strength or dimensionality and signal to noise ratio, we simulate data from two cases with different factor strengths. In the first case, we assume the error term $e_{i,t}$ in AR(1) model of both factors are independent $\mathcal{N}(0, N)$ and $\mathcal{N}(0, 0.5N)$, respectively, where N is the dimension of original data. The use of order N in the variance of error terms in the AR(1) model of both factors reflects both factors' pervasiveness, or equivalently, that the strength of factors is set to be the strongest. The use of different scales 1 and 0.5 in the variance of $e_{i,t}$ for $i = 1, 2$ is to ensure that the first two largest eigenvalues of accumulated squared autocovariance matrices that are associated with the two factors are spiked and unequal.

In the second case, we consider a weak factor model where the error terms $e_{i,t}$ in AR(1) model of both factors $\{\boldsymbol{f}_t\}$ are independent $\mathcal{N}(0, N^{0.2})$ and $\mathcal{N}(0, 0.5N^{0.2})$, respectively. In this case, the factors are relatively weak since the variance of

$e_{i,t}$ in AR(1) model are of order $N^{0.2}$, which means the spikiness of the first two largest eigenvalues of accumulated squared autocovariance matrices weakens when N increases. The use of 0.5 as the AR coefficient in both cases reflects a moderate temporal dependence within each factor. Generally speaking, a larger AR coefficient or strong temporal dependence within each factor also demands a relatively large sample size T for better sieve bootstrap results. In comparison, a smaller AR coefficient or weaker temporal dependence within each factor can lead to the overestimating problem on the number of factors, which is already considered in the second case.

In both cases, we repeat the simulation by 500 times and each time we generate $B = 499$ bootstrap samples to create a confidence interval for the (standardised) overall mean statistics defined as $\theta_y := \frac{\sqrt{T}}{\sqrt{N}}\mathbf{1}^\top \boldsymbol{Q}\boldsymbol{\mu}_f$, with $\boldsymbol{\mu}_f$ the population mean of $\{\boldsymbol{f}_t\}$ for strong factors' case and $\theta_y := \frac{\sqrt{T}}{\sqrt{N^{0.2}}}\mathbf{1}^\top \boldsymbol{Q}\boldsymbol{\mu}_f$ for weak factors' case, where the two statistics are standardised by the factor strength for comparison of the length of confidence intervals as below.

Specifically, we first compute $B = 499$ sieve bootstrap estimates of (standardised) overall mean statistics as $\overline{\overline{y^*}} = \frac{\sqrt{T}}{\sqrt{N}}\mathbf{1}^\top \widehat{\boldsymbol{Q}}\overline{\boldsymbol{f}^*}$, and then create bootstrap intervals based on it. In this example, we investigate the performance of our proposed sieve bootstrap method based on two types of bootstrap intervals, the nonparametric bootstrap interval using quantiles and the parametric bootstrap interval based on normality. Both bootstrap intervals are practically popular, computationally efficient and easy to implement. For an arbitrary statistic θ and its sample estimate $\widehat{\theta}$, the nonparametric bootstrap interval using quantiles are calculated as $\left(2\widehat{\theta} - \theta^*_{(1-\alpha/2)},\ 2\widehat{\theta} + \theta^*_{(\alpha/2)}\right)$, where $\theta^*_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the bootstrap estimates θ^*. The nonparametric bootstrap interval using quantiles are sometimes referred as reverse percentile interval as the order of upper and lower quantiles are reversed in the formula. The idea of nonparametric bootstrap interval using quantiles is to use the bootstrap distribution of $(\theta^* - \widehat{\theta})$ to approximate the distribution of $(\widehat{\theta} - \theta)$. On the other hand, the parametric bootstrap interval based on normality can be computed as $\left(\widehat{\theta} - b^* - \sqrt{v^*}z_{(1-\alpha/2)}, \widehat{\theta} - b^* - \sqrt{v^*}z_{(1-\alpha/2)}\right)$, where b^* and v^* are the bootstrap

estimates of bias and variance of $\widehat{\theta}$, and $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of standard normal distribution. Similar to the nonparametric bootstrap interval using quantiles, the parametric bootstrap interval based on normality also assumes the bootstrap distribution of $(\theta^* - \widehat{\theta})$ correctly approximates the distribution of $(\widehat{\theta} - \theta)$, but are constructed in a parametric way. To achieve improved empirical coverage and average width of intervals, more sophisticated intervals with additional corrections on bias and variance may also be constructed, such as double bootstrap, with a higher cost of computations. Since this example's main purpose is to inspect the validity and consistency of our proposed sieve bootstrap method under various cases, we only use these two ways of bootstrap intervals as they are simple and computationally efficient. Finally, to get a comprehensive comparison on the performance of two types of intervals, we compute empirical coverage, average width, and interval score (Gneiting and Raftery, 2007) of bootstrap intervals under various combinations of N and T. The interval score of a bootstrap interval (l, u) is computed as

$$S_\alpha = (u-l) + \frac{2}{\alpha}(l-\theta)\mathbb{1}\{\theta < l\} + \frac{2}{\alpha}(\theta-u)\mathbb{1}\{\theta > u\},$$

with the idea of rewarding narrower intervals but putting penalties on intervals missing true statistics θ. Therefore, when the empirical coverage and average width of two bootstrap intervals are close, the average interval score can be used for overall comparison.

In Tables 2.1 and 2.2, we present the empirical coverage, average width and interval score of nonparametric bootstrap intervals using quantiles and parametric bootstrap intervals based on normality for θ_y in strong factors' case. The nominate coverages we investigated are 95%, 90%, and 80% with different combinations of N and T for comparison. As shown in both tables, when the sample size T is large enough and the factors are strong, or the signal to noise ratio is not affected by N, the empirical coverage is reasonably close to the nominated coverage and are not largely affected by the ratio of N/T. Besides, bootstrap intervals' average width is also similar for various combinations of N and T.

This result is often referred to as the 'Blessing of dimensionality' in the literature of high-dimensional statistics. The performance of bootstrap confidence intervals generally benefits from the increase of both N and T. Between nonparametric bootstrap intervals using quantiles and parametric bootstrap intervals based on normality, the average interval scores are very close for almost all combinations of N and T. Hence, we conclude that both intervals perform well in strong factors' case.

Table 2.1: Empirical coverage, average width and interval score of nonparametric bootstrap intervals using quantiles for θ_y of a strong factor model

Nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
	50	0.936	8.333	12.545	0.894	6.997	11.379	0.804	5.454	10.069
	100	0.950	8.367	10.517	0.906	7.033	9.799	0.784	5.498	9.012
200	200	0.946	8.201	11.404	0.892	6.909	10.336	0.794	5.407	9.378
	500	0.944	8.415	12.056	0.884	7.076	10.972	0.766	5.522	9.895
	1000	0.934	8.021	13.144	0.882	6.752	11.586	0.808	5.278	9.789
	50	0.942	8.507	12.616	0.890	7.168	11.454	0.790	5.591	10.207
	100	0.930	8.275	12.210	0.864	6.959	11.417	0.800	5.449	10.062
500	200	0.940	8.525	12.891	0.902	7.177	11.310	0.812	5.609	9.639
	500	0.940	8.608	13.715	0.880	7.240	12.188	0.800	5.646	10.629
	1000	0.948	8.572	13.150	0.904	7.229	11.363	0.800	5.642	9.912
	50	0.944	8.452	11.809	0.896	7.104	10.932	0.784	5.553	9.696
	100	0.946	8.415	12.187	0.890	7.093	11.050	0.790	5.530	9.938
1000	200	0.936	8.114	12.200	0.880	6.827	11.039	0.772	5.324	10.177
	500	0.952	8.347	11.236	0.904	7.022	10.194	0.828	5.476	8.993
	1000	0.952	8.355	12.103	0.884	7.029	10.906	0.782	5.476	9.982

Table 2.2: Empirical coverage, average width and interval score of parametric bootstrap intervals based on normality for θ_y of a strong factor model

Parametric bootstrap intervals based on normality										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
	50	0.936	8.388	12.346	0.886	7.039	11.504	0.794	5.484	10.124
	100	0.948	8.445	10.461	0.906	7.087	9.748	0.796	5.522	9.066
200	200	0.948	8.282	11.344	0.890	6.951	10.227	0.798	5.416	9.270
	500	0.948	8.482	11.642	0.888	7.118	10.732	0.766	5.546	9.869
	1000	0.928	8.094	13.232	0.890	6.793	11.551	0.812	5.292	9.805
	50	0.940	8.582	12.444	0.902	7.202	11.314	0.784	5.611	10.161
	100	0.928	8.352	11.975	0.874	7.009	11.323	0.796	5.461	10.044
500	200	0.942	8.597	13.018	0.910	7.215	11.255	0.810	5.622	9.636
	500	0.944	8.669	13.605	0.886	7.275	12.147	0.796	5.668	10.585
	1000	0.950	8.649	13.049	0.904	7.258	11.335	0.804	5.655	9.956
	50	0.940	8.526	11.860	0.898	7.156	10.847	0.780	5.575	9.706
	100	0.948	8.492	12.313	0.892	7.127	10.936	0.790	5.553	9.882
1000	200	0.942	8.185	12.060	0.888	6.869	11.136	0.778	5.352	10.194
	500	0.954	8.411	11.134	0.904	7.059	10.216	0.828	5.500	8.990
	1000	0.952	8.423	11.870	0.894	7.069	10.882	0.784	5.508	9.911

However, as shown in Tables 2.3 and 2.4, when the factors are weak, with the

factor strength set to $N^{0.2}$, the empirical coverage tends to increase with N/T, and the bootstrap intervals become wider and wider. This suggests that bootstrap overestimates the standard error of (standardised) overall mean statistics when N increases. When the factors are weak, the spikiness of the first two largest eigenvalues of accumulated squared autocovariance matrices decreases. The number of factors is overestimated, which brings the noises into bootstrap samples. As a result, neither of the two types of bootstrap intervals performs well when factors are weak, and N/T is large. The bootstrap distribution of the (standardised) overall mean statistics suffers from comparably fatter tails.

Table 2.3: Empirical coverage, average width and interval score of nonparametric bootstrap intervals using quantiles for θ_y of a weak factor model

Nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.972	8.700	10.109	0.938	7.319	9.162	0.840	5.706	8.324
	100	0.990	8.854	9.248	0.960	7.429	8.112	0.868	5.809	7.138
	200	0.988	9.141	9.757	0.960	7.705	8.658	0.894	6.011	7.382
	500	0.996	10.263	10.321	0.984	8.643	8.957	0.934	6.742	7.637
	1000	0.962	10.979	19.094	0.958	9.232	14.167	0.932	7.214	10.428
500	50	0.936	8.524	13.455	0.866	7.182	11.862	0.780	5.617	10.336
	100	0.958	8.603	11.340	0.906	7.235	10.602	0.816	5.673	9.332
	200	0.964	8.872	11.071	0.926	7.480	9.893	0.856	5.852	8.767
	500	0.982	9.471	11.053	0.960	7.967	9.635	0.862	6.220	8.220
	1000	0.996	10.251	10.728	0.978	8.610	9.306	0.932	6.723	7.827
1000	50	0.930	8.506	12.453	0.876	7.165	11.454	0.774	5.591	10.392
	100	0.940	8.462	11.969	0.882	7.102	11.029	0.774	5.547	10.057
	200	0.948	8.315	11.400	0.908	6.984	10.121	0.810	5.438	9.288
	500	0.968	8.959	10.526	0.934	7.528	9.288	0.860	5.870	8.158
	1000	0.980	9.183	9.962	0.948	7.730	9.178	0.878	6.047	8.273

2.5.3 Sieve bootstrap for spiked eigenvalues of squared autocovariance matrix

The study on spiked eigenvalues of high-dimensional covariance matrix has received massive attention in the past decades. For time series data, researchers are generally interested in the spiked eigenvalues of the squared autocovariance matrix. However, the theoretical results of these spiked eigenvalues of squared autocovariance matrix for high-dimensional time series are much more involved and hard to be applied for practical analysis. As an alternative, the bootstrap can be considered for real data applications when the theoretical results do not

Table 2.4: Empirical coverage, average width and interval score of parametric bootstrap intervals based on normality for θ_y of a weak factor model

Parametric bootstrap intervals based on normality										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.974	8.771	9.914	0.944	7.361	9.071	0.834	5.735	8.339
	100	0.986	8.925	9.350	0.962	7.490	8.187	0.870	5.836	7.175
	200	0.988	9.222	9.851	0.960	7.739	8.606	0.894	6.030	7.370
	500	0.996	10.355	10.467	0.986	8.690	9.018	0.934	6.771	7.615
	1000	0.964	11.079	19.357	0.958	9.298	14.231	0.932	7.244	10.363
500	50	0.940	8.605	13.329	0.878	7.221	11.776	0.770	5.626	10.327
	100	0.962	8.684	11.013	0.906	7.288	10.354	0.808	5.678	9.284
	200	0.972	8.967	10.826	0.926	7.526	9.758	0.862	5.863	8.721
	500	0.984	9.551	11.057	0.962	8.016	9.594	0.866	6.245	8.244
	1000	0.994	10.327	10.965	0.978	8.667	9.373	0.940	6.752	7.799
1000	50	0.930	8.585	12.507	0.886	7.205	11.569	0.780	5.614	10.395
	100	0.942	8.524	12.153	0.878	7.153	11.116	0.774	5.573	10.001
	200	0.954	8.358	11.114	0.908	7.014	10.150	0.802	5.465	9.213
	500	0.972	9.013	10.452	0.944	7.564	9.320	0.866	5.894	8.175
	1000	0.982	9.272	9.892	0.944	7.781	9.171	0.878	6.062	8.197

exist or hard to be implemented. As discussed in Proposition 2.1, the bootstrap estimates $\delta_i^*(k)$ are generally consistent to $\delta_i(k)$. However, without a general central limit theorem (CLT) on $\widehat{\delta}_i(k)$, the spiked eigenvalues of squared sample autocovariance matrix, the validity of sieve bootstrap estimate is generally hard to derive theoretically. Therefore, we use simulation to study our sieve bootstrap method's performance on estimating $\delta_i(k)$. To be more specific, the data we generated are based on the strong factors' case model in Section 2.5.2. We continue the study on validity and consistency of our sieve bootstrap method by accessing the empirical coverage of bootstrap intervals on the first two largest eigenvalues δ_1 and δ_2 of squared lag-1 autocovariance matrix. In order to get a comprehensive comparison based on average width and interval score of bootstrap intervals for various combination of N and T, instead of δ_1 and δ_2, the bootstrap intervals can be created based on standardised eigenvalues $\delta_1^0 = \frac{\sqrt{T}}{N^2}\delta_1$ and $\delta_2^0 = \frac{\sqrt{T}}{N^2}\delta_2$.

First of all, we compute the empirical coverage, average width, and interval score for nonparametric bootstrap intervals using quantiles and parametric bootstrap intervals based on normality for strong factors. As shown in Tables 2.5 to 2.8, neither of two types bootstrap intervals can provide an desired result as the empirical coverage probabilities are consistently lower than the nominate probabilities for each interval, especially when T is small. While the 'blessing of

dimensionality' may improve the empirical coverage of both intervals on δ_1 and δ_2 for large N, the results are not as good for the overall mean statistic. They consistently underestimated empirical coverage probabilities are due to the skewness of sampling distribution of $\widehat{\delta}_i(k)$, the eigenvalues of sample lag-k autocovariance matrices, especially for a relatively small T. In general, nonparametric bootstrap intervals using quantiles and parametric bootstrap intervals based on normality perform well when the sampling distributions are symmetric. However, the parametric bootstrap interval based on normality, which is symmetric, and the nonparametric bootstrap interval using quantiles, which is reversely skewed, do not perform well when the sample statistic follows a skewed distribution. To consider for this skewness, an unreversed nonparametric bootstrap interval using quantiles, computed as $\left(\theta^*_{(\alpha/2)},\ \theta^*_{(1-\alpha/2)}\right)$, can also be computed and compared since the skewness of sample statistic is retained by the bootstrap estimates.

Table 2.5: Empirical coverage, average width and interval score of nonparametric bootstrap intervals using quantiles for δ_1^0 of a strong factor model

Nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.842	11.682	28.717	0.810	9.653	19.626	0.768	7.385	13.919
	100	0.842	11.775	27.056	0.832	9.719	18.711	0.786	7.455	13.268
	200	0.854	11.625	26.011	0.830	9.592	18.102	0.782	7.342	12.964
	500	0.842	11.625	27.406	0.820	9.621	18.971	0.774	7.380	13.436
	1000	0.858	11.484	24.363	0.836	9.450	17.167	0.786	7.251	12.377
500	50	0.876	11.127	22.388	0.848	9.264	16.777	0.760	7.185	13.468
	100	0.902	11.310	22.295	0.876	9.447	16.504	0.786	7.307	13.178
	200	0.900	11.610	22.324	0.868	9.679	16.786	0.768	7.508	13.986
	500	0.888	11.352	23.279	0.850	9.481	17.251	0.750	7.330	13.685
	1000	0.900	11.342	21.580	0.872	9.452	16.159	0.800	7.327	12.785
1000	50	0.930	11.381	19.688	0.902	9.544	15.411	0.820	7.440	12.960
	100	0.904	10.920	22.254	0.862	9.183	17.069	0.756	7.135	13.601
	200	0.916	11.244	19.419	0.888	9.426	15.552	0.788	7.327	12.762
	500	0.938	11.277	18.019	0.896	9.472	14.308	0.798	7.363	12.137
	1000	0.932	11.303	18.043	0.892	9.466	14.796	0.790	7.357	12.636

As shown in Tables 2.9 and 2.10, unreversed nonparametric bootstrap intervals using quantiles outperform the other two competitors for δ_1 with almost all combinations of N and T and for δ_2 with small T. The failure of nonparametric bootstrap intervals using quantiles and parametric bootstrap intervals based on normality, on the other hand, verifies the skewness on the distribution of $\widehat{\delta}_i(k)$. Although some bias-corrected intervals may also be constructed, for example, by double bootstrap, to improve the empirical coverage probabilities further, those

Table 2.6: Empirical coverage, average width and interval score of nonparametric bootstrap intervals using quantiles for δ_2^0 of a strong factor model

Nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.824	2.274	6.551	0.744	1.884	5.088	0.612	1.450	4.406
	100	0.820	2.207	7.632	0.766	1.828	5.568	0.666	1.414	4.387
	200	0.836	2.200	6.303	0.770	1.821	4.852	0.654	1.397	4.118
	500	0.866	2.259	5.204	0.794	1.875	4.344	0.684	1.443	3.886
	1000	0.814	2.175	7.343	0.756	1.810	5.241	0.642	1.389	4.213
500	50	0.896	2.665	5.580	0.826	2.230	4.495	0.722	1.727	3.979
	100	0.894	2.519	5.194	0.842	2.102	4.021	0.746	1.630	3.344
	200	0.896	2.592	5.265	0.842	2.172	4.153	0.756	1.683	3.575
	500	0.912	2.565	4.980	0.868	2.146	3.849	0.778	1.665	3.261
	1000	0.896	2.579	4.799	0.846	2.160	3.964	0.744	1.675	3.562
1000	50	0.924	2.728	4.662	0.878	2.290	3.952	0.792	1.785	3.343
	100	0.918	2.689	4.374	0.874	2.252	3.613	0.778	1.753	3.223
	200	0.904	2.670	5.078	0.856	2.241	4.149	0.756	1.744	3.518
	500	0.938	2.695	4.118	0.872	2.259	3.589	0.780	1.759	3.201
	1000	0.908	2.635	4.957	0.868	2.213	3.926	0.758	1.724	3.384

Table 2.7: Empirical coverage, average width and interval score of parametric bootstrap intervals based on normality for δ_1^0 of a strong factor model

Parametric bootstrap intervals based on normality										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.900	12.001	21.334	0.868	10.072	16.830	0.790	7.847	13.373
	100	0.898	12.119	19.448	0.868	10.171	15.888	0.812	7.924	12.638
	200	0.910	11.947	19.026	0.878	10.026	15.385	0.802	7.812	12.419
	500	0.918	12.004	20.782	0.880	10.074	16.199	0.796	7.849	12.975
	1000	0.928	11.819	18.451	0.902	9.919	14.744	0.814	7.728	11.834
500	50	0.930	11.320	17.496	0.874	9.500	14.953	0.788	7.402	12.868
	100	0.934	11.489	18.826	0.892	9.642	15.345	0.814	7.512	12.901
	200	0.936	11.806	19.009	0.884	9.908	15.920	0.784	7.720	13.795
	500	0.934	11.521	18.831	0.882	9.669	15.571	0.766	7.533	13.356
	1000	0.928	11.487	17.414	0.892	9.640	14.745	0.802	7.511	12.433
1000	50	0.944	11.528	17.524	0.896	9.674	14.991	0.808	7.538	13.001
	100	0.928	11.071	19.106	0.880	9.291	15.728	0.776	7.239	13.158
	200	0.928	11.392	16.591	0.902	9.561	14.526	0.786	7.449	12.613
	500	0.950	11.420	15.860	0.914	9.584	13.501	0.790	7.467	12.002
	1000	0.936	11.420	15.883	0.910	9.584	14.020	0.804	7.467	12.494

Table 2.8: Empirical coverage, average width and interval score of parametric bootstrap intervals based on normality for δ_2^0 of a strong factor model

Parametric bootstrap intervals based on normality										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.844	2.329	6.049	0.776	1.955	5.166	0.646	1.523	4.390
	100	0.840	2.269	7.042	0.788	1.904	5.520	0.672	1.484	4.392
	200	0.860	2.251	5.834	0.790	1.890	4.850	0.696	1.472	4.015
	500	0.868	2.318	5.169	0.814	1.945	4.525	0.706	1.516	3.865
	1000	0.858	2.234	6.576	0.778	1.875	5.119	0.664	1.461	4.171
500	50	0.894	2.707	5.001	0.834	2.272	4.553	0.740	1.770	3.969
	100	0.908	2.553	4.421	0.864	2.142	3.824	0.772	1.669	3.299
	200	0.914	2.631	4.639	0.848	2.208	4.033	0.760	1.721	3.561
	500	0.934	2.608	4.476	0.878	2.189	3.803	0.792	1.705	3.263
	1000	0.910	2.620	4.101	0.844	2.199	3.877	0.754	1.713	3.541
1000	50	0.942	2.765	4.445	0.896	2.321	3.813	0.798	1.808	3.324
	100	0.942	2.719	3.923	0.880	2.282	3.500	0.778	1.778	3.232
	200	0.920	2.705	4.526	0.860	2.270	4.048	0.756	1.769	3.504
	500	0.946	2.730	3.799	0.880	2.291	3.476	0.788	1.785	3.190
	1000	0.930	2.671	4.284	0.874	2.242	3.743	0.770	1.746	3.338

methods on reducing the error of bootstrap intervals generally have significant requirements on computations and are beyond the scope of this work.

Table 2.9: Empirical coverage, average width and interval score of unreversed nonparametric bootstrap intervals using quantiles for δ_1^0 of a strong factor model

Unreversed nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.942	11.682	16.165	0.900	9.653	14.147	0.810	7.385	12.086
	100	0.952	11.775	14.332	0.900	9.719	12.771	0.818	7.455	11.336
	200	0.960	11.625	13.247	0.910	9.592	12.188	0.824	7.342	10.910
	500	0.948	11.625	14.970	0.918	9.621	13.025	0.818	7.380	11.344
	1000	0.970	11.484	13.951	0.926	9.450	12.077	0.830	7.251	10.399
500	50	0.952	11.127	14.622	0.904	9.264	13.088	0.798	7.185	11.784
	100	0.944	11.310	16.024	0.906	9.447	14.138	0.818	7.307	12.024
	200	0.940	11.610	17.097	0.878	9.679	15.135	0.778	7.508	13.444
	500	0.956	11.352	15.419	0.900	9.481	13.710	0.768	7.330	12.577
	1000	0.966	11.342	14.164	0.914	9.452	12.760	0.798	7.327	11.576
1000	50	0.944	11.381	16.588	0.898	9.544	14.825	0.798	7.440	13.120
	100	0.936	10.920	16.243	0.900	9.183	14.309	0.786	7.135	12.541
	200	0.944	11.244	14.869	0.898	9.426	13.542	0.784	7.327	12.327
	500	0.964	11.277	14.435	0.924	9.472	12.799	0.796	7.363	11.696
	1000	0.956	11.303	14.641	0.896	9.466	13.749	0.800	7.357	12.565

2.6 Particulate matter concentration

We apply the proposed sieve bootstrap methods on a real data set of high-dimensional time series. The raw data are observations of PM_{10} particles in the air, collected on a half-hour basis in Graz, Austria from 1 Oct. 2010 to 31 Mar.

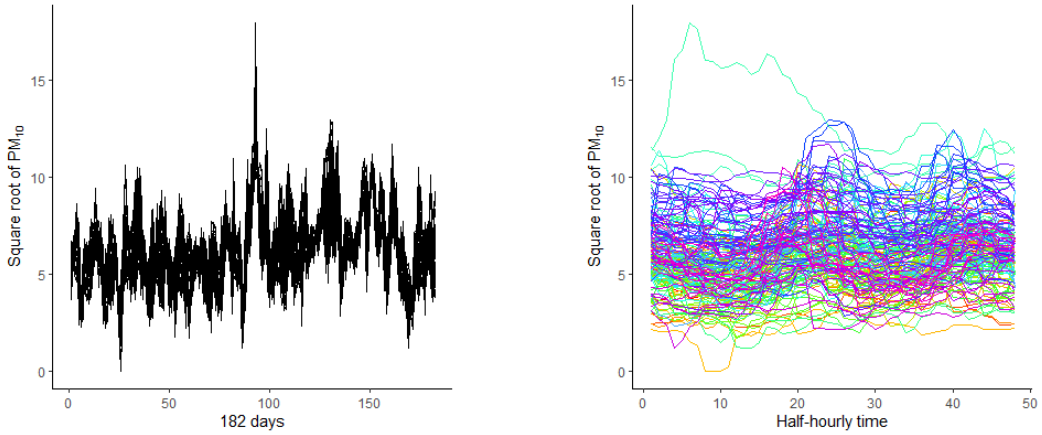
Table 2.10: Empirical coverage, average width and interval score of unreversed nonparametric bootstrap intervals using quantiles for δ_2^0 of a strong factor model

Unreversed nonparametric bootstrap intervals using quantiles										
		95%			90%			80%		
T	N	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score	Empirical coverage	Average width	Average interval score
200	50	0.852	2.274	5.738	0.776	1.884	4.721	0.664	1.450	3.888
	100	0.860	2.207	5.817	0.798	1.828	4.725	0.682	1.414	3.777
	200	0.876	2.200	5.351	0.790	1.821	4.410	0.664	1.397	3.728
	500	0.898	2.259	4.414	0.832	1.875	3.779	0.698	1.443	3.260
	1000	0.850	2.175	6.195	0.768	1.810	4.989	0.660	1.389	4.038
500	50	0.906	2.665	4.854	0.862	2.230	4.204	0.758	1.727	3.523
	100	0.904	2.519	4.682	0.852	2.102	4.029	0.720	1.630	3.478
	200	0.914	2.592	4.626	0.864	2.172	3.984	0.748	1.683	3.404
	500	0.930	2.565	4.373	0.884	2.146	3.683	0.770	1.665	3.149
	1000	0.910	2.579	3.995	0.868	2.160	3.718	0.760	1.675	3.267
1000	50	0.932	2.728	4.244	0.874	2.290	3.785	0.788	1.785	3.319
	100	0.932	2.689	4.154	0.864	2.252	3.752	0.778	1.753	3.303
	200	0.912	2.670	4.703	0.860	2.241	4.137	0.764	1.744	3.548
	500	0.946	2.695	3.969	0.892	2.259	3.526	0.806	1.759	3.148
	1000	0.924	2.635	4.453	0.860	2.213	3.972	0.758	1.724	3.399

2011. PM_{10} particles represent a common type of air pollutant that can be found in smoke and dust with an aerodynamic diameter of less than 0.01mm.

This data set has been studied in Hörmann et al. (2015) for topics of dynamic functional principal component analysis (FPCA) and in Shang (2018) for comparisons of bootstrap methods for stationary functional time series. The original data is preprocessed by a square-root transformation to stabilize the variance and avoid heavy-tailed observations as directed in Aue et al. (2015) and Hörmann et al. (2015). The square-root of PM_{10} levels contained in a 48×182 matrix are then plotted in Figure 2.8a as high-dimensional time series over 182 days with dimension of 48 and in Figure 2.8b as 182 repeats of 48 half-hourly observations within each day. In general, the PM_{10} concentration levels are relatively high in winters when the temperatures are low and the pollutants related to daily life such as traffics and heating lack space to disperse in the atmosphere. Therefore, the day-to-day PM_{10} levels in winter are highly temporally dependent, while the half-hourly observations in each day experience similar local patterns which are mainly related to people's day-to-day life and temperature.

In Hörmann et al. (2015) and Shang (2018), observations of half-hourly PM_{10} levels as in Figure 2.8b are assumed to come from a functional curve. In general, for a functional time series, the original observations are smoothed before further studies such as FPCA and functional bootstrap. Hence, according to Hörmann



(a) Univariate time-series plot (b) Functional time-series plot

Figure 2.8: Observed time series of (square-root) PM_{10} levels

et al. (2015) and Shang (2018), there are 182 temporal dependent functional curves each smoothed from 48 observations. However, as illustrated in this work, the pre-smoothing results rely heavily on the smoothness condition of the functional curve. When the observations are not dense enough, pre-smoothing may cause a loss of information, especially on local patterns. To maintain the original features of time series observations to the greatest extent, we instead treat the data as a multivariate or high-dimensional time series. We then perform the proposed sieve bootstrap methods with a factor model on this 48 by 182 matrix of time series. This creates a bootstrap confidence interval for the mean levels of (square root) PM_{10} which are temporal dependent at each half-hourly time point, and to create a bootstrap confidence surface for the lag-1 autocovariance matrix of (square root) PM_{10} levels.

In Figure 2.9, a 90% nonparametric bootstrap interval using quantiles is created on the mean levels of (square root) PM_{10}, defined as $\theta_y := \boldsymbol{Q}\boldsymbol{\mu}_f$ with $\boldsymbol{\mu}_f$ the population mean of temporal dependent factors $\{\boldsymbol{f}_t\}$. From this plot of sample estimate and confidence interval of θ_y, it is clear that local patterns, for example, between 4th and 10-th half-hourly time points, are preserved flawlessly by our proposed sieve bootstrap methods based on high-dimensional time series. Similarly, a sample estimate and a 90% unreversed nonparametric bootstrap interval using quantiles for lag-1 autocovariance matrix $\mathrm{Cov}(\boldsymbol{y}_t, \boldsymbol{y}_{t+1})$

of temporal dependent (square root) PM_{10} levels at 48 half-hourly time points are also computed and presented in Figure 2.10. This unreversed nonparametric bootstrap interval using quantiles provides interval estimates on autocovariance of (square root) PM_{10} levels between two consecutive days, where, as shown in Figure 2.10, the local patterns are again completely preserved by our proposed sieve bootstrap methods.

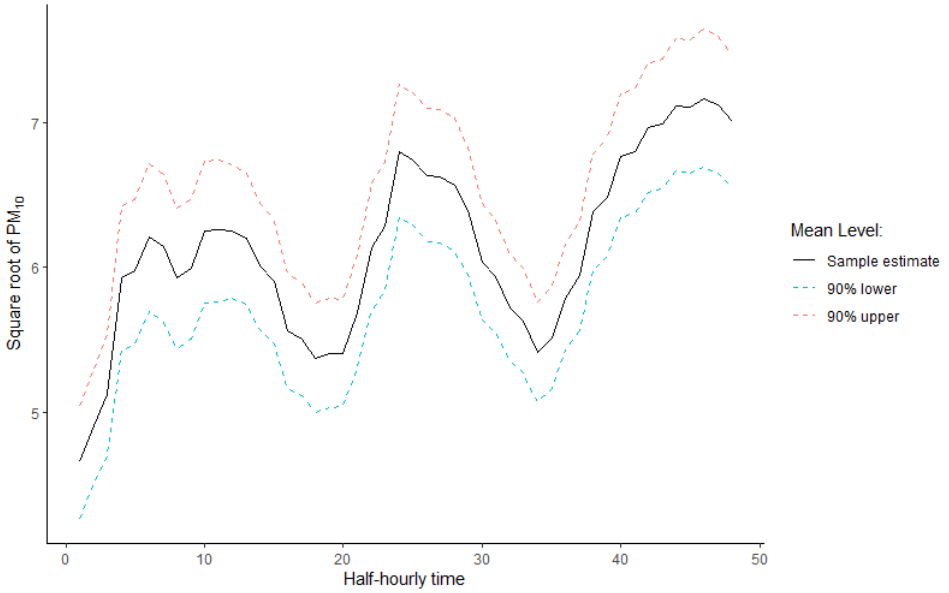


Figure 2.9: 90% Sieve bootstrap confidence interval for the mean of temporal dependent (square root) PM_{10} levels at 48 half-hourly time

2.7 Conclusions and discussions

We first introduce pre-smoothing failure on sparse functional time series observations, especially when there are local patterns in the population curve. We also address statistical inferences, such as bootstrap associated with pre-smoothing of sparse observations under functional set-up. We then suggest alternatively treating the sparse observations as multivariate high-dimensional time series. We adapt dimension-reduction methods, such as factor models, to pursue statistical inferences, such as bootstrap. Specifically, we suggest using autocovariance to estimate the factor model and perform a sieve bootstrap on the estimated factors to provide ultimate inferences on the original time series. Our proposed sieve

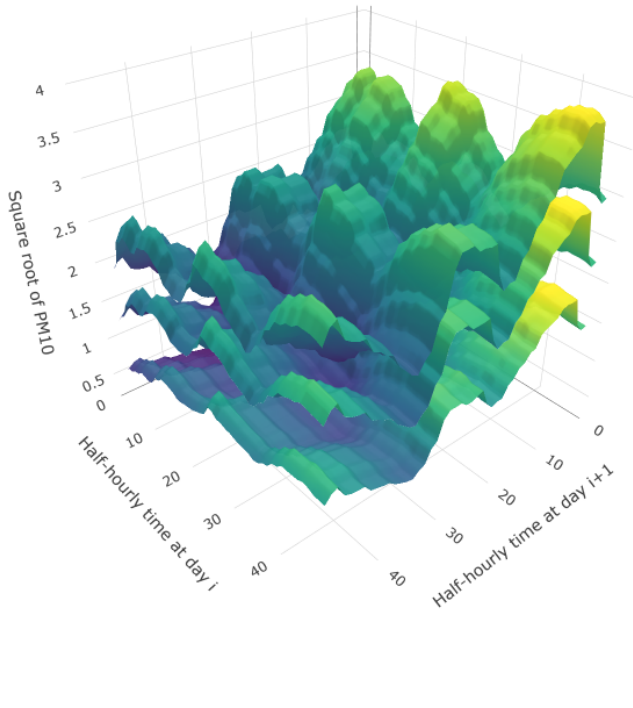


Figure 2.10: 90% Sieve bootstrap confidence surface for lag-1 autocovariance of temporal dependent (square root) PM_{10} levels at 48 half-hourly time point

bootstrap methods using factor models provide valid statistical inferences on a general overall mean statistic and maintain consistency on bootstrap estimates of spiked eigenvalues of autocovariance matrices. Simulation studies provide numerical evidence on the finite-sample performance of the sieve bootstrap methods on high-dimensional time series following strong factor models. At last, we apply our methods to PM_{10} data for constructing bootstrap confidence intervals for mean vector and autocovariance matrix, respectively.

Our work is crucial as a building block of sieve bootstrap methods under high-dimensional set-up and inspirational for studying the differences and connections between studies on functional and high-dimensional time series. Our future work includes further exploration and justification on density or sparsity of functional time series observations and the impact of pre-smoothing, which are fundamental for data analysis.

2.A Appendix A: Technical proofs of theorems

Proof of Theorem 2.1. Let $\boldsymbol{f}_t^b = \sum_{l=1}^{p} \widetilde{\boldsymbol{A}}_{l,p} \boldsymbol{f}_{t-l}^b + \boldsymbol{e}_{t,p}^b$, where $\{\widetilde{\boldsymbol{A}}_{l,p},\ l = 1, 2, ..., p\}$ are the estimators of AR coefficient matrices based on true factors $\{\boldsymbol{f}_t\}$, and $\{\boldsymbol{e}_{t,p}^b, t = p+1, p+2, ..., T\}$ are generated by i.i.d. resampling from the centered residuals $(\widetilde{\boldsymbol{e}}_{t,p} - \overline{\widetilde{\boldsymbol{e}}}_{T,p})$ with $\widetilde{\boldsymbol{e}}_{t,p} = \boldsymbol{f}_t - \sum_{l=1}^{p} \widetilde{\boldsymbol{A}} \boldsymbol{f}_{t-l}$ and $\overline{\widetilde{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p} \sum_{t=p+1}^{T} \widetilde{\boldsymbol{e}}_{t,p}$. Therefore, $\{\boldsymbol{f}_t^b\}$ are bootstrap pseudo-variables generated based on the true factors $\{\boldsymbol{f}_t\}$ rather than $\{\widehat{\boldsymbol{f}}_t\}$. Recall that $\{\boldsymbol{f}_t^*\}$ are bootstrapped based on the centered residuals $(\widehat{\boldsymbol{e}}_{t,p} - \overline{\widehat{\boldsymbol{e}}}_{T,p})$ with $\widehat{\boldsymbol{e}}_{t,p} = \widehat{\boldsymbol{f}}_t - \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p} \widehat{\boldsymbol{f}}_{t-l}$ and $\overline{\widehat{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p} \sum_{t=p+1}^{T} \widehat{\boldsymbol{e}}_{t,p}$, and we define $\mathbb{E}^*$ and Cov^* as the expectation and covariance with respect to the measure assigning probability $1/(T-p)$ to each observation, respectively. Therefore, $\mathbb{E}^* \overline{\boldsymbol{f}_T^*} = \overline{\widehat{\boldsymbol{f}}_T}$ by definition and we can write

$$
\begin{aligned}
\sqrt{T} \boldsymbol{c}^\top \widehat{\boldsymbol{Q}} \left(\overline{\boldsymbol{f}_T^*} - \mathbb{E}^* \overline{\boldsymbol{f}_T^*} \right) &=: \mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3 \\
&= \sqrt{T} \boldsymbol{c}^\top \boldsymbol{Q} \left(\overline{\boldsymbol{f}_T^b} - \mathbb{E}^* \overline{\boldsymbol{f}_T^b} \right) + \sqrt{T} \boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q} \right) \left(\overline{\boldsymbol{f}_T^*} - \mathbb{E}^* \overline{\boldsymbol{f}_T^*} \right) \\
&\quad + \sqrt{T} \boldsymbol{c}^\top \boldsymbol{Q} \left[\left(\overline{\boldsymbol{f}_T^*} - \mathbb{E}^* \overline{\boldsymbol{f}_T^*} \right) - \left(\overline{\boldsymbol{f}_T^b} - \mathbb{E}^* \overline{\boldsymbol{f}_T^b} \right) \right],
\end{aligned}
$$

with obvious definitions of $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$.

For the term $\mathcal{M}_1$, under Assumptions 2.1 (iii), 2.2 and the additional assumption in Theorem 2.1 that $\lim_{T\to\infty} \mathbb{V}(\sqrt{T}\overline{f_T}) = \sum_{k\in\mathbf{Z}} \mathbf{\Gamma}_f(k) < \infty$, using Theorem 2.1 in Politis et al. (1997), we have the following CLT for $\sqrt{T}\,\overline{f_T}$

$$\sqrt{T}\left(\overline{f_T} - \mathbb{E}\overline{f_T}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \sum_{k\in\mathbf{Z}} \mathbf{\Gamma}_f(k)\right).$$

Moreover, under the additional assumptions in Theorem 2.1, $\boldsymbol{c}^\top \boldsymbol{Q}$ is an r-dimensional vector such that $\|\boldsymbol{c}^\top \boldsymbol{Q}\|_{\ell_1} < \infty$ for a fixed r, Therefore, under Assumptions 2.1 (ii) and 2.2, we can use Cramer-Wold Theorem (Cramér and Wold, 1936) to conclude for the scalar $\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\overline{\boldsymbol{f}_T}$ that

$$\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\left(\overline{f_T} - \mathbb{E}\overline{f_T}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \boldsymbol{c}^\top \boldsymbol{Q}\left(\sum_{k\in\mathbf{Z}} \mathbf{\Gamma}_f(k)\right)\boldsymbol{Q}^\top \boldsymbol{c}\right),$$

when $T, N \to \infty$.

Besides, under the strong mixing condition on true factors $\{\boldsymbol{f}_t\}$, the empirical moments of $\{\boldsymbol{e}_t\}$ converge to its population counterpart. Therefore, under all the assumptions of 2.1, we fulfil all the conditions of Theorem 4.1 in Meyer and Kreiss (2015). Consequently, we can use Theorem 4.1 in Meyer and Kreiss (2015) to conclude that the general VAR sieve bootstrap is valid for $\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\overline{\boldsymbol{f}_T}$ since $\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\overline{\boldsymbol{f}_T}$ shares the same CLT with its counterpart generated from the companion process as discussed in Meyer and Kreiss (2015). Hence

$$d_K\left(\mathcal{L}^*\left(\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\left(\overline{f_T^b} - \mathbb{E}^*\overline{f_T^b}\right)\right), \mathcal{L}\left(\sqrt{T}\boldsymbol{c}^\top \boldsymbol{Q}\left(\overline{f_T} - \mathbb{E}\overline{f_T}\right)\right)\right) = o_P(1)$$

as $T, N \to \infty$.

Therefore, to see the assertion in Theorem 2.1, we need to show that when $T, N \to \infty$, both $\mathcal{M}_2$ and $\mathcal{M}_3$ tend to 0 in probability, then apply Slutsky's theorem.

To show $\mathcal{M}_2 \to 0$ in probability for $T, N \to \infty$, we first of all notice that

$$\sqrt{T}\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\left(\overline{f_T^*} - \mathbb{E}^*\overline{f_T^*}\right) = \frac{1}{\sqrt{T}}\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\sum_{t=1}^{T}\left(f_t^* - \overline{\widehat{f}_T}\right).$$

Therefore, we can show that

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{T}\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\left(\overline{f_T^*} - \mathbb{E}\overline{f_T^*}\right)\right]^2 \\
=&\mathbb{E}\left[\frac{1}{T}\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\sum_{t=1}^{T}\left(f_t^* - \overline{\widehat{f}_T}\right)\right]\left[\sum_{s=1}^{T}\left(f_s^* - \overline{\widehat{f}_T}\right)^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)^\top \boldsymbol{c}\right] \\
=&\left[\frac{1}{T}\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(f_t^* - \overline{\widehat{f}_T}\right)\left(f_s^* - \overline{\widehat{f}_T}\right)^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)^\top \boldsymbol{c}\right] \\
\leq&\frac{1}{T}\left\|\boldsymbol{c}^\top \left(\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right)\right\|^2 \left\|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(f_t^* - \overline{\widehat{f}_T}\right)\left(f_s^* - \overline{\widehat{f}_T}\right)^\top\right\|_F \\
=&O_P\left(\frac{1}{T^2}\left\|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(f_t^* - \overline{\widehat{f}_T}\right)\left(f_s^* - \overline{\widehat{f}_T}\right)^\top\right\|_F\right),
\end{aligned}$$

where the last line follows from the fact that $\|\boldsymbol{c}^\top \boldsymbol{Q}\|_{\ell_1} < \infty$ for $N, T \to \infty$ under the additional assumptions in Theorem 2.1, $\|\boldsymbol{Q}\|_2 \asymp \sqrt{N}$, and $\left\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_2 = O_P\left(N^{1/2}T^{-1/2}\right)$ by Lemma 2.3.

Define $\Sigma_{e,p}^* := \mathbb{E}^*\left(\boldsymbol{e}_t^*\boldsymbol{e}_t^{*\top}\right)$, then

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left(f_t^* - \overline{\widehat{f}_T}\right)\left(f_s^* - \overline{\widehat{f}_T}\right)^\top &= \sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left(\left(\sum_{l_1=0}^{\infty}\widehat{\boldsymbol{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1}^*\right)\left(\sum_{l_2=0}^{\infty}\widehat{\boldsymbol{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2}^*\right)^\top\right) \\
&= \sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\sum_{l_1=0}^{\infty}\sum_{l_2=0}^{\infty}\left(\widehat{\boldsymbol{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1}^*\boldsymbol{e}_{s-l_2}^{*\top}\widehat{\boldsymbol{\Psi}}_{l_2,p}^\top\right) \\
&= \sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{l=0}^{\infty}\widehat{\boldsymbol{\Psi}}_{l,p}\mathbb{E}^*\left(\boldsymbol{e}_{t-l}^*\boldsymbol{e}_{t-l}^{*\top}\right)\widehat{\boldsymbol{\Psi}}_{s-t+l,p}^\top
\end{aligned}$$

where $\boldsymbol{e}_{t-l_1}^*$ and $\boldsymbol{e}_{t-l_2}^*$ are i.i.d. bootstrapped therefore $\mathbb{E}^*\left(\boldsymbol{e}_{t-l_1}^*\boldsymbol{e}_{t-l_2}^{*\top}\right) = \boldsymbol{0}$ for $l_1 \neq l_2$.

Hence we can show that

$$\begin{aligned}\frac{1}{T^2}\left\|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left(f_t^*-\overline{\widehat{f}_T}\right)\left(f_s^*-\overline{\widehat{f}_T}\right)^\top\right\|_F \leq& \frac{1}{T^2}\left\|\Sigma_{e,p}^*\right\|_F\sum_{l=0}^{\infty}\left\|\widehat{\mathbf{\Psi}}_{l,p}\right\|_F\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{\Psi}}_{s-t+l,p}\right\|_F\\ =&O_P\left(\frac{1}{T}\right),\end{aligned}$$

where we note that Lemmas 2.6 and 2.8 imply the summability of $\left\|\widehat{\mathbf{\Psi}}_{l,p}\right\|_F$, hence $\sum_{s=1}^{T}\left\|\widehat{\mathbf{\Psi}}_{s-t+l,p}\right\|_F$ is bounded for $T\to\infty$. Therefore, $\frac{1}{T}\sum_{l=0}^{\infty}\left\|\widehat{\mathbf{\Psi}}_{l,p}\right\|_F\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{\Psi}}_{s-t+l,p}\right\|_F$ is bounded for $T\to\infty$, and we can conclude that $\mathbb{E}^*\left[\sqrt{T}\boldsymbol{c}^\top\left(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\right)\left(\overline{f_T^*}-\mathbb{E}^*\overline{f_T^*}\right)\right]^2\to 0$ in probability, which suffices for $\mathcal{M}_2\to 0$ in probability conditional on the sample.

For $\mathcal{M}_3$, we first write

$$\begin{aligned}&\mathbb{E}^*\left[\sqrt{T}\boldsymbol{c}^\top\boldsymbol{Q}\left\{\left(\overline{f_T^*}-\mathbb{E}^*\overline{f_T^*}\right)-\left(\overline{f_T^b}-\mathbb{E}^*\overline{f_T^b}\right)\right\}\right]^2\\ =&\mathbb{E}^*\left\|\sqrt{T}\boldsymbol{c}^\top\boldsymbol{Q}\left\{\left(\overline{f_T^*}-\overline{\widehat{f}_T}\right)-\left(\overline{f_T^b}-\overline{\widetilde{f}_T}\right)\right\}\right\|^2\\ \leq&\|\boldsymbol{c}^\top\boldsymbol{Q}\|^2\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\|\left\{\left(f_t^*-\overline{\widehat{f}_T}\right)-\left(f_t^b-\overline{\widetilde{f}_T}\right)\right\}\left\{\left(f_s^*-\overline{\widehat{f}_T}\right)-\left(f_s^b-\overline{\widetilde{f}_T}\right)\right\}^\top\right\|_F\\ =&O_P\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\|\left\{\left(f_t^*-\overline{\widehat{f}_T}\right)-\left(f_t^b-\overline{\widetilde{f}_T}\right)\right\}\left\{\left(f_s^*-\overline{\widehat{f}_T}\right)-\left(f_s^b-\overline{\widetilde{f}_T}\right)\right\}^\top\right\|_F\right),\end{aligned}$$

where the last line follows from the fact that $\|\boldsymbol{c}^\top\boldsymbol{Q}\|^2$ is bounded when $N\to\infty$.

To proceed, first note that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\{\left(f_t^*-\overline{\widehat{f}_T}\right)-\left(f_t^b-\overline{\widetilde{f}_T}\right)\right\}\left\{\left(f_s^*-\overline{\widehat{f}_T}\right)-\left(f_s^b-\overline{\widetilde{f}_T}\right)\right\}^\top\\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\{\sum_{l_1=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^*-\widetilde{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^b\right\}\left\{\sum_{l_2=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*-\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^b\right\}^\top\\
=&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\{\sum_{l_1=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^*\right\}\left\{\sum_{l_2=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*-\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^b\right\}^\top\\
+&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}^*\left\{\sum_{l_1=0}^{\infty}\widetilde{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^b\right\}\left\{\sum_{l_2=0}^{\infty}\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^b-\widehat{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*\right\}^\top\\
=:&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\mathcal{H}_1+\mathcal{H}_2\right),
\end{aligned}
$$

with an obvious notation for $\mathcal{H}_1$ and $\mathcal{H}_2$. Then, we only consider $\mathcal{H}_1$ as $\mathcal{H}_2$ can be dealt with similarly.

For $\mathcal{H}_1$, we can further decompose it as

$$
\begin{aligned}
\mathcal{H}_1=&\sum_{s=1}^{T}\mathbb{E}^*\left\{\sum_{l_1=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^*\right\}\left\{\sum_{l_2=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*-\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*\right\}^\top\\
&+\sum_{s=1}^{T}\mathbb{E}^*\left\{\sum_{l_1=0}^{\infty}\widehat{\mathbf{\Psi}}_{l_1,p}\boldsymbol{e}_{t-l_1,p}^*\right\}\left\{\sum_{l_2=0}^{\infty}\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^*-\widetilde{\mathbf{\Psi}}_{l_2,p}\boldsymbol{e}_{s-l_2,p}^b\right\}^\top\\
=&\sum_{s=1}^{T}\sum_{l=0}^{\infty}\widehat{\mathbf{\Psi}}_{l,p}\mathbb{E}^*\left\{\boldsymbol{e}_{t-l,p}^*\boldsymbol{e}_{t-l,p}^{*\top}\right\}\left\{\widehat{\mathbf{\Psi}}_{l+s-t,p}-\widetilde{\mathbf{\Psi}}_{l+s-t,p}\right\}^\top\\
&+\sum_{s=1}^{T}\sum_{l=0}^{\infty}\widehat{\mathbf{\Psi}}_{l,p}\mathbb{E}^*\left\{\boldsymbol{e}_{t-l,p}^*(\boldsymbol{e}_{t-l,p}^*-\boldsymbol{e}_{t-l,p}^b)^\top\right\}\widetilde{\mathbf{\Psi}}_{l+s-t,p}^\top\\
=:&\mathcal{H}_{11}+\mathcal{H}_{12}.
\end{aligned}
$$

where the second last line follows from the bootstrap independence for $l_1\neq l_2$.

Hence we can conclude for $\mathcal{H}_{11}$ that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\mathcal{H}_{11}\|_F =& \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|\sum_{l=0}^{\infty}\widehat{\mathbf{\Psi}}_{l,p}\Sigma^*_{e,p}\left\{\widehat{\mathbf{\Psi}}_{l+s-t,p}-\widetilde{\mathbf{\Psi}}_{l+s-t,p}\right\}^\top\right\|_F \\
\leq& \left\|\Sigma^*_{e,p}\right\|_F\frac{1}{T}\sum_{l=0}^{\infty}\left\|\widehat{\mathbf{\Psi}}_{l,p}\right\|_F\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{\Psi}}_{l+s-t,p}-\widetilde{\mathbf{\Psi}}_{l+s-t,p}\right\|_F \\
=& O_P\left(p^{\frac{3}{2}}\left\|\widehat{\boldsymbol{A}}_p-\widetilde{\boldsymbol{A}}_p\right\|_F\right) \\
=& o_P(1),
\end{aligned}$$

where the second last line follows from the results in Lemmas 2.6 and 2.8, and the last line follows the result in Lemma 2.8.

For $\mathcal{H}_{12}$ we can show that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\mathcal{H}_{12}\|_F \leq& \sqrt{\mathbb{E}^*\left\|\boldsymbol{e}^*_{t,p}\right\|^2}\sqrt{\mathbb{E}^*\left\|\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right\|^2}\frac{1}{T}\sum_{l=0}^{\infty}\left\|\widehat{\mathbf{\Psi}}_{l,p}\right\|_F\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|\widehat{\mathbf{\Psi}}_{l+s-t,p}\right\|_F \\
=& O_P\left(\sqrt{\mathbb{E}^*\left\|\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right\|^2}\right),
\end{aligned}$$

where the last line follows from the same arguments on summability properties in Lemmas 2.6. Hence it remains to show $\mathbb{E}^*\left\|\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right\|^2\to 0$ in probability. Recall that $\mathbb{E}^*$ defines expectation with respect to the measure assigning probability $1/(T-p)$ to each observation, this follows as

$$\begin{aligned}
\mathbb{E}^*\left\|\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right\|^2 =& \mathbb{E}^*\left\{\left(\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right)\left(\boldsymbol{e}^*_{t,p}-\boldsymbol{e}^b_{t,p}\right)^\top\right\} \\
=& \frac{1}{T-p}\sum_{t=p+1}^{T}\left\{(\widehat{\boldsymbol{e}}_{t,p}-\bar{\widehat{\boldsymbol{e}}}_{T,p})-(\widetilde{\boldsymbol{e}}_{t,p}-\bar{\widetilde{\boldsymbol{e}}}_{T,p})\right\}\left\{(\widehat{\boldsymbol{e}}_{t,p}-\bar{\widehat{\boldsymbol{e}}}_{T,p})-(\widetilde{\boldsymbol{e}}_{t,p}-\bar{\widetilde{\boldsymbol{e}}}_{T,p})\right\}^\top \\
=& \frac{1}{T-p}\sum_{t=p+1}^{T}\left\{(\widehat{\boldsymbol{e}}_{t,p}-\widetilde{\boldsymbol{e}}_{t,p})-(\bar{\widehat{\boldsymbol{e}}}_{T,p}-\bar{\widetilde{\boldsymbol{e}}}_{T,p})\right\}\left\{(\widehat{\boldsymbol{e}}_{t,p}-\widetilde{\boldsymbol{e}}_{t,p})-(\bar{\widehat{\boldsymbol{e}}}_{T,p}-\bar{\widetilde{\boldsymbol{e}}}_{T,p})\right\}^\top \\
\leq& \frac{2}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{e}}_{t,p}-\widetilde{\boldsymbol{e}}_{t,p}\right\|^2+2\left\{\left\|\bar{\widetilde{\boldsymbol{e}}}_{T,p}\right\|^2+\left\|\bar{\widehat{\boldsymbol{e}}}_{T,p}\right\|^2-2\left\|\bar{\widetilde{\boldsymbol{e}}}_{T,p}\right\|\left\|\bar{\widehat{\boldsymbol{e}}}_{T,p}\right\|\right\} \\
\leq& \frac{2}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{e}}_{t,p}-\widetilde{\boldsymbol{e}}_{t,p}\right\|^2+4\left\{\left\|\bar{\widetilde{\boldsymbol{e}}}_{T,p}\right\|^2+\left\|\bar{\widehat{\boldsymbol{e}}}_{T,p}\right\|^2\right\}.
\end{aligned}$$

Recall that when $\{f_t\}$ and $\{\widehat{f}_t\}$ have non-zero means, $\widetilde{e}_{t,p} = \left(f_t - \overline{f_T}\right) - \sum_{l=1}^{p} \widetilde{A}_{l,p}\left(f_{t-l} - \overline{f}_T\right)$ and $\widehat{e}_{t,p} = \left(\widehat{f}_t - \overline{\widehat{f}_T}\right) - \sum_{l=1}^{p} \widehat{A}_{l,p}\left(\widehat{f}_{t-l} - \overline{\widehat{f}}_T\right)$. Without altering the idea of proof, to simplify the notations used, we use $\{f_t\}$ and $\{\widehat{f}_t\}$ to denote the demeaned factors $\left(f_t - \overline{f_T}\right)$ and their sample counterparts $\left(\widehat{f}_t - \overline{\widehat{f}_T}\right)$, respectively. Therefore, with the same arguments in the proof of Lemma 2.9, we have

$$\begin{aligned}
\frac{2}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{e}_{t,p}-\widetilde{e}_{t,p}\right\|^2 =& \frac{2}{T-p}\sum_{t=p+1}^{T}\left\|(\widehat{f}_t-f_t)+\sum_{l=1}^{p}(\widetilde{A}_{l,p}f_{t-l}-\widehat{A}_{l,p}\widehat{f}_{t-l})\right\|^2 \\
\leq& \frac{4}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{f}_t-f_t\right\|^2+\frac{4}{T-p}\sum_{t=p+1}^{T}\left\|\sum_{l=1}^{p}\widetilde{A}_{l,p}f_{t-l}-\widehat{A}_{l,p}\widehat{f}_{t-l}\right\|^2 \\
\leq& \frac{4}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{f}_t-f_t\right\|^2+8\sum_{l=1}^{p}\left\|\widehat{A}_{l,p}\right\|_F^2\frac{1}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{f}_{t-l}-f_{t-l}\right\|^2 \\
&+8\left\|\sum_{l=1}^{p}\left(\widehat{A}_{l,p}-\widetilde{A}_{l,p}\right)\frac{1}{T-p}\sum_{t=p+1}^{T}f_{t-l}\right\|_F^2 \\
=& O_P\left(\sup_{p+1\leq t\leq T}\left\|\widehat{f}_t-f_t\right\|^2\right)+O_P\left(\left\|\sum_{l=1}^{p}\left(\widehat{A}_p-\widetilde{A}_p\right)\right\|_F^2\right) \\
=& O_P\left(\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right)^2\right)+O_P\left(p^8\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right)^2\right) \\
=& o_P(1), \qquad (2.11)
\end{aligned}$$

where the third last line follows from the fact that $\left\|\widehat{A}_{l,p}\right\|_F^2$ is summable, which is implied by Assumption 2.3 and Lemma 2.4. The second last line is then a direct result of Lemmas 2.4 and 2.5, and Assumption 2.4 implies the last line.

Furthermore, $\overline{\widehat{e}}_{T,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{e}_{t,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widehat{f}_t - \sum_{l=1}^{p}\widehat{A}_{l,p}\widehat{f}_{t-l}\right)$ and we can show that

$$\left\|\overline{\widehat{e}}_{T,p}\right\|^2 \leq 2\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{f}_t\right\|^2+2\left\|\sum_{l=1}^{p}\widehat{A}_{l,p}\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{f}_{t-l}\right\|^2 = o_P(1). \qquad (2.12)$$

This is because, firstly

$$\begin{aligned}\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\boldsymbol{f}}_t\right\|^2 \leq& 2\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\boldsymbol{f}_t\right\|^2+2\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widehat{\boldsymbol{f}}_t-\boldsymbol{f}_t\right)\right\|^2\\ =&O_P\left(\frac{1}{T-p}\right)+O_P\left(\frac{1}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{f}}_t-\boldsymbol{f}_t\right\|^2\right)\\ =&O_P\left(\frac{1}{T-p}\right)+O_P\left(\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right)^2\right)=o_P(1),\end{aligned}$$

where the second last line follows as we have assumed the population mean of $\{\boldsymbol{f}_t\}$ is 0 for technical convenience. Moreover,

$$\begin{aligned}\left\|\sum_{l=1}^{p}\widehat{\boldsymbol{A}}_{l,p}\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\boldsymbol{f}}_{t-l}\right\| \leq& \sum_{l=1}^{p}\left\|\widehat{\boldsymbol{A}}_{l,p}\right\|_F\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\boldsymbol{f}}_{t-l}\right\|\\ =&O_P(1)\times O_P\left(\frac{1}{\sqrt{T-p}}+\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{N}}\right)=o_P(1),\end{aligned}$$

where the second last line follows from the summability conditions in Lemma 2.6, the order of $\left\|\widehat{\boldsymbol{f}}_t-\boldsymbol{f}_t\right\|$ in Lemma 2.4 and the fact that the mean of $\{\widehat{\boldsymbol{f}}_t\}$ is assumed to be 0 for technical convenience.

Lastly, we can show that $\left\|\bar{\widetilde{\boldsymbol{e}}}_T\right\|^2 \to 0$ in probability with the same technique as stated above for $\left\|\bar{\widetilde{\boldsymbol{e}}}_T\right\|$. Hence with (2.11) and (2.12), we can conclude that $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\mathcal{H}_{12}\|_F \to 0$ in probability. Together with the result that $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\mathcal{H}_{11}\|_F \to 0$ in probability, we have $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\|\mathcal{H}_{1}\|_F \to 0$ in probability. Therefore, it suffices to conclude that $\mathcal{M}_3 \to 0$ in probability conditional on the sample.

Consequently, by utilizing Slutsky's theorem conditional on the sample, we can conclude that

$$d_K\left(\mathcal{L}^*\left(\sqrt{T}\boldsymbol{c}^\top\widehat{\boldsymbol{Q}}\left(\overline{\boldsymbol{f}_T^*}-\mathbb{E}^*\overline{\boldsymbol{f}_T^*}\right)\right),\mathcal{L}\left(\sqrt{T}\boldsymbol{c}^\top\boldsymbol{Q}\left(\overline{\boldsymbol{f}_T}-\mathbb{E}\overline{\boldsymbol{f}_T}\right)\right)\right)\xrightarrow{p}0,$$

□

Proof of Theorem 2.2. Without loss of generality, we again assume $\{\boldsymbol{f}_t\}$ are the

demeaned factors (or the mean of factors are all 0) in this proof to simplify the notations.

Firstly, notice that $\boldsymbol{f}_t^* = \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p} \boldsymbol{f}_{t-l}^* + \boldsymbol{e}_t^* = \sum_{l=1}^{\infty} \widehat{\boldsymbol{\Psi}}_{l,p} \boldsymbol{e}_{t-l}^* + \boldsymbol{e}_t^* = \sum_{l=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l,p} \boldsymbol{e}_{t-l}^*$. We can then represent $\boldsymbol{\Gamma}_f^*(k)$ as

$$\begin{aligned}
\boldsymbol{\Gamma}_f^*(k) &= \mathrm{Cov}^*(\boldsymbol{f}_t^*, \boldsymbol{f}_{t+k}^*) \\
&= \mathrm{Cov}^* \left(\sum_{l_1=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l_1,p} \boldsymbol{e}_{t-l_1}^*, \sum_{l_2=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l_2,p} \boldsymbol{e}_{t+k-l_2}^* \right) \\
&= \sum_{l_1=0}^{\infty} \sum_{l_2=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l_1,p} \mathrm{Cov}^*(\boldsymbol{e}_{t-l_1}^*, \boldsymbol{e}_{t+k-l_2}^*) \widehat{\boldsymbol{\Psi}}_{l_2,p}^\top \\
&= \sum_{l_1=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l_1,p} \mathrm{Cov}^*(\boldsymbol{e}_{t-l_1}^*, \boldsymbol{e}_{t-l_1}^*) \widehat{\boldsymbol{\Psi}}_{l_1+k,p}^\top \\
&= \sum_{l=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l,p} \widehat{\boldsymbol{\Sigma}}_{e,p} \widehat{\boldsymbol{\Psi}}_{l+k,p}^\top,
\end{aligned}$$

where we stress the fact that $\mathrm{Cov}^*(\boldsymbol{e}_{t-l_1}^*, \boldsymbol{e}_{t-l_2}^*) = 0$ for $l_1 \neq l_2$ and $\mathrm{Cov}^*(\boldsymbol{e}_{t-l_1}^*, \boldsymbol{e}_{t-l_1}^*) = \mathbb{E}^*(\boldsymbol{e}_t^* \boldsymbol{e}_t^{*\top}) = \widehat{\boldsymbol{\Sigma}}_{e,p}$ for all $l_1 \in \mathbf{Z}$, since $\boldsymbol{e}_t^*$ is uniformly distributed on the set of centered residuals $(\widehat{\boldsymbol{e}}_{t,p} - \overline{\widehat{\boldsymbol{e}}}_T)$. Similarly,

$$\begin{aligned}
\boldsymbol{\Gamma}_f(k) &= \mathrm{Cov}\left(\boldsymbol{f}_t, \boldsymbol{f}_{t+k}\right) \\
&= \mathrm{Cov} \left(\sum_{l_1=0}^{\infty} \boldsymbol{\Psi}_{l_1} \boldsymbol{e}_{t-l_1}, \sum_{l_2=0}^{\infty} \boldsymbol{\Psi}_{l_2} \boldsymbol{e}_{t+k-l_2} \right) \\
&= \sum_{l_1=0}^{\infty} \sum_{l_2=0}^{\infty} \boldsymbol{\Psi}_{l_1} \mathrm{Cov}\left(\boldsymbol{e}_{t-l_1}, \boldsymbol{e}_{t+k-l_2}\right) \boldsymbol{\Psi}_{l_2}^\top \\
&= \sum_{l_1=0}^{\infty} \boldsymbol{\Psi}_{l_1} \mathrm{Cov}\left(\boldsymbol{e}_{t-l_1}, \boldsymbol{e}_{t-l_1}\right) \boldsymbol{\Psi}_{l_1+k}^\top \\
&= \sum_{l=0}^{\infty} \boldsymbol{\Psi}_l \boldsymbol{\Sigma}_e \boldsymbol{\Psi}_{l+k}^\top,
\end{aligned}$$

where we write $\boldsymbol{\Sigma}_e = \mathrm{Cov}(\boldsymbol{e}_t, \boldsymbol{e}_t)$ and use the fact that $\boldsymbol{f}_t = \sum_{l=1}^{\infty} \boldsymbol{A}_l \boldsymbol{f}_{t-l} + \boldsymbol{e}_t = \sum_{l=1}^{\infty} \boldsymbol{\Psi}_l \boldsymbol{e}_{t-l} + \boldsymbol{e}_t = \sum_{l=0}^{\infty} \boldsymbol{\Psi}_l \boldsymbol{e}_{t-l}$.

To see the assertion in this theorem, we first of all define an intermediate term $\boldsymbol{\Gamma}_{f,p}(k) := \sum_{l=0}^{\infty} \boldsymbol{\Psi}_{l,p} \boldsymbol{\Sigma}_{e,p} \boldsymbol{\Psi}_{l+k,p}^\top$, where $\{\boldsymbol{\Psi}_{l,p},\ l \in \mathbf{N}\}$ are the power series coefficients matrices of $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \boldsymbol{A}_{l,p} z^l\right)^{-1}$ for $|z| \leq 1$, and $\boldsymbol{\Sigma}_{e,p} = \mathrm{Cov}(\boldsymbol{e}_{t,p}, \boldsymbol{e}_{t,p})$

where $\boldsymbol{e}_{t,p} = \boldsymbol{f}_t - \sum_{l=1}^{p} \boldsymbol{A}_{l,p}\boldsymbol{f}_{t-l}$ with $\{\boldsymbol{A}_{l,p},\ l \in \mathbf{N}\}$ the finite predictor coefficients matrices of $\{\boldsymbol{A}_l,\ l \in \mathbf{N}\}$. Hence by triangular inequality, we have

$$\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_f(k)\right\|_2 \leq \left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_{f,p}(k)\right\|_2 + \left\|\boldsymbol{\Gamma}_{f,p}(k) - \boldsymbol{\Gamma}_f(k)\right\|_2.$$

It is then sufficient to show both terms on the right side converge to 0 in probability. For $\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_{f,p}(k)\right\|_2$, we have

$$\begin{aligned}
\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_{f,p}(k)\right\|_2 &= \left\|\sum_{l=0}^{\infty} \widehat{\boldsymbol{\Psi}}_{l,p}\widehat{\boldsymbol{\Sigma}}_{e,p}\widehat{\boldsymbol{\Psi}}_{l+k,p}^{\top} - \sum_{l=0}^{\infty} \boldsymbol{\Psi}_{l,p}\boldsymbol{\Sigma}_{e,p}\boldsymbol{\Psi}_{l+k,p}^{\top}\right\|_2 \\
&= \left\|\sum_{l=0}^{\infty} \left[\left(\widehat{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right)\widehat{\boldsymbol{\Sigma}}_{e,p}\widehat{\boldsymbol{\Psi}}_{l+k,p}^{\top} + \boldsymbol{\Psi}_{l,p}\left(\widehat{\boldsymbol{\Sigma}}_{e,p} - \boldsymbol{\Sigma}_{e,p}\right)\widehat{\boldsymbol{\Psi}}_{l+k,p}^{\top}\right.\right. \\
&\quad \left.\left. + \boldsymbol{\Psi}_{l,p}\boldsymbol{\Sigma}_{e,p}\left(\widehat{\boldsymbol{\Psi}}_{l+k,p} - \boldsymbol{\Psi}_{l+k,p}\right)^{\top}\right]\right\|_2 \\
&= O_P\left(\sum_{l=1}^{\infty}\left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F\right) + O_P\left(\left\|\widehat{\boldsymbol{\Sigma}}_{e,p} - \boldsymbol{\Sigma}_{e,p}\right\|_F\right),
\end{aligned}$$

where the second last line follows from the norm summable conditions on $\widehat{\boldsymbol{\Psi}}_{l,p}$ and $\boldsymbol{\Psi}_{l,p}$. Hence we can use the results of Lemma 2.8 and 2.9 to conclude that $\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_{f,p}(k)\right\|_2 \to 0$ in probability. Similarly, we have

$$\begin{aligned}
\left\|\boldsymbol{\Gamma}_{f,p}(k) - \boldsymbol{\Gamma}_f(k)\right\|_2 &= \left\|\sum_{l=0}^{\infty} \boldsymbol{\Psi}_{l,p}\boldsymbol{\Sigma}_{e,p}\boldsymbol{\Psi}_{l+k,p}^{\top} - \sum_{l=0}^{\infty} \boldsymbol{\Psi}_{l}\boldsymbol{\Sigma}_{e}\boldsymbol{\Psi}_{l+k}^{\top}\right\|_2 \\
&= \left\|\sum_{l=0}^{\infty} \left[\left(\boldsymbol{\Psi}_{l,p} - \boldsymbol{\Psi}_{l}\right)\boldsymbol{\Sigma}_{e,p}\boldsymbol{\Psi}_{l+k,p}^{\top} + \boldsymbol{\Psi}_{l}\left(\boldsymbol{\Sigma}_{e,p} - \boldsymbol{\Sigma}_{e}\right)\boldsymbol{\Psi}_{l+k,p}^{\top}\right.\right. \\
&\quad \left.\left. + \boldsymbol{\Psi}_{l}\boldsymbol{\Sigma}_{e}\left(\boldsymbol{\Psi}_{l+k,p} - \boldsymbol{\Psi}_{l+k}\right)^{\top}\right]\right\|_2 \\
&= O_P\left(\sum_{l=1}^{\infty}\left\|\boldsymbol{\Psi}_{l,p} - \boldsymbol{\Psi}_{l}\right\|_F\right) + O_P\left(\left\|\boldsymbol{\Sigma}_{e,p} - \boldsymbol{\Sigma}_{e}\right\|_F\right),
\end{aligned}$$

since $\boldsymbol{\Psi}_{l,p}$ and $\boldsymbol{\Psi}_l$ are norm summable. Hence $\left\|\boldsymbol{\Gamma}_{f,p}(k) - \boldsymbol{\Gamma}_f(k)\right\|_2 \to 0$ in probability by Lemmas 2.8 and 2.9. Therefore we can conclude that $\left\|\boldsymbol{\Gamma}_f^*(k) - \boldsymbol{\Gamma}_f(k)\right\|_2 \to 0$ in probability. □

Proof of Proposition 2.1. To see the assertions, we first note that,

$$\begin{aligned}\left\|\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right\|_2 &= \left\|\widehat{\boldsymbol{Q}}\mathbf{\Gamma}_f^*(k)\widehat{\boldsymbol{Q}}^T-\boldsymbol{Q}\mathbf{\Gamma}_f(k)\boldsymbol{Q}^T\right\|_2\\ &\leq \left\|\left(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\right)\mathbf{\Gamma}_f^*(k)\widehat{\boldsymbol{Q}}^T\right\|_2+\left\|\boldsymbol{Q}\left(\mathbf{\Gamma}_f^*(k)-\mathbf{\Gamma}_f(k)\right)\widehat{\boldsymbol{Q}}^T\right\|_2\\ &+\left\|\boldsymbol{Q}\mathbf{\Gamma}_f(k)\left(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\right)^T\right\|_2\\ &= O_P\left(N^{1/2}\left\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\right\|_2\right)+O_P\left(N\left\|\mathbf{\Gamma}_f^*(k)-\mathbf{\Gamma}_f(k)\right\|_2\right)=o_P(1),\end{aligned}$$

where the last line follows from Assumption 2.1, Lemma 2.2 and Theorem 2.2. To see (2.10), we can apply Weyl's Eigenvalue Theorem (Fan et al., 2013b), that is

$$|\delta_i^*(k)-\delta_i(k)|\leq\frac{1}{N^2}\left\|\mathbf{\Gamma}_y^*(k)\mathbf{\Gamma}_y^*(k)^\top-\mathbf{\Gamma}_y(k)\mathbf{\Gamma}_y(k)^\top\right\|_2.$$

Furthermore,

$$\begin{aligned}\frac{1}{N^2}\left\|\mathbf{\Gamma}_y^*(k)\mathbf{\Gamma}_y^*(k)^\top-\mathbf{\Gamma}_y(k)\mathbf{\Gamma}_y(k)^\top\right\|_2 &= \frac{1}{N^2}\left\|\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]\mathbf{\Gamma}_y^*(k)^\top+\mathbf{\Gamma}_y(k)\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]^\top\right\|_2\\ &\leq \frac{1}{N^2}\left\|\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]\mathbf{\Gamma}_y^*(k)^\top\right\|_2\\ &+\frac{1}{N^2}\left\|\mathbf{\Gamma}_y(k)\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]^\top\right\|_2.\end{aligned}$$

It is then sufficient to consider one of the two terms on the right side since the other one can be dealt with similarly. To study $\frac{1}{N^2}\left\|\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]\mathbf{\Gamma}_y^*(k)^\top\right\|_2$, we first notice that from Assumption 2.1, Lemma 2.2 and Theorem 2.2, $\left\|\mathbf{\Gamma}_y^*(k)\right\|_2=\left\|\widehat{\boldsymbol{Q}}\mathbf{\Gamma}_f^*(k)\widehat{\boldsymbol{Q}}^T\right\|_2\asymp N$. Therefore, we have

$$\begin{aligned}\frac{1}{N^2}\left\|\left[\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right]\mathbf{\Gamma}_y^*(k)^\top\right\|_2 &= O_P\left(\frac{1}{N}\left\|\mathbf{\Gamma}_y^*(k)-\mathbf{\Gamma}_y(k)\right\|_2\right)\\ &= O_P\left(N^{-1/2}\left\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\right\|_2\right)+O_P\left(\left\|\mathbf{\Gamma}_f^*(k)-\mathbf{\Gamma}_f(k)\right\|_2\right),\end{aligned}$$

where both terms on the right side converge to 0 in probability as shown in Lemma 2.3 and Theorem 2.2. □

2.B Appendix B: Auxiliary lemmas and proofs

In this section, we present some auxiliary results that facilitate the proofs of theorems in this chapter. Those auxiliary results are divided into two subsections according to the related topics. In the first subsection, we present some results for factor models' estimates, and in the second subsection, the results for sieve bootstrap of factor models are summarised.

2.B.1 Auxiliary results for estimates of factor models

Lemma 2.2. *Denoted by* $\|\mathbf{V}\|_{min}$ *the positive square root of the minimum eigenvalue of* $\mathbf{V}\mathbf{V}^\top$ *or* $\mathbf{V}^\top\mathbf{V}$*, under Assumption 2.1, we have*

$$\left\|\mathbf{\Gamma}_f(k)\right\|_2 \asymp 1 \asymp \left\|\mathbf{\Gamma}_f(k)\right\|_{min}, \tag{2.13}$$

and

$$\left\|\widetilde{\mathbf{\Gamma}}_f(k) - \mathbf{\Gamma}_f(k)\right\|_2 = O_P\left(T^{-1/2}\right). \tag{2.14}$$

Lemma 2.2 is a modification of the results in Lemma 1 and 2 of Lam et al. (2011) for the strong factors' case, since we have assumed $\boldsymbol{Q}^\top\boldsymbol{Q} = N\boldsymbol{I}_r$ but not $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r$ as in Lam et al. (2011). Therefore, the proof of Lemma 2.2 is similar to the proofs of Lemma 1 and 2 in Lam et al. (2011), hence omitted.

Lemma 2.3. *Under Assumption 2.1,*

$$\left\|\widehat{\boldsymbol{Q}} - \boldsymbol{Q}\right\|_2 = O_P\left(N^{1/2}T^{-1/2}\right),$$

and

$$N^{-1/2}\left\|\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{f}}_t - \boldsymbol{Q}\boldsymbol{f}_t\right\|_2 = O_P\left(T^{-1/2} + N^{-1/2}\right).$$

Although compared with the model introduced in Lam et al. (2011), we scale the columns in $\boldsymbol{Q}$ by $\sqrt{N}$ in our factor models' setting, the above convergence rate is the same as that of strong factors' case in Theorem 3 of Lam et al. (2011). Besides, the proof of Lemma 2.3 is the case for strong factors in the proof of

Theorem 3 in Lam et al. (2011) with the only difference on scaled factor loading matrix $\boldsymbol{Q}$ and factors $\boldsymbol{f}$. Therefore, the proof is omitted here.

Lemma 2.4. *Define* $\widehat{\boldsymbol{\Gamma}}_f(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}\widehat{\boldsymbol{f}}_t\widehat{\boldsymbol{f}}_{t+k}$ *and* $\widetilde{\boldsymbol{\Gamma}}_f(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}\boldsymbol{f}_t\boldsymbol{f}_{t+k}$*, for some* $k \leq p$*, where* p *fulfils Assumption 2.4. It then holds that*

$$\left\|\widehat{\boldsymbol{\Gamma}}_f(k) - \widetilde{\boldsymbol{\Gamma}}_f(k)\right\|_2 = O_P\left(N^{-1/2} + T^{-1/2}\right).$$

Lemma 2.4 illustrates the convergence rate on autocovariance matrices of estimated factors under strong factors' case, which is an extension to the convergence rate of estimated factors obtained in Theorem 3 in Lam et al. (2011).

Proof of Lemma 2.4. First of all, we notice that

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_f(k) - \widetilde{\boldsymbol{\Gamma}}_f(k) &= \frac{1}{T-k}\sum_{t=1}^{T-k}\left(\widehat{\boldsymbol{f}}_t\widehat{\boldsymbol{f}}_{t+k} - \boldsymbol{f}_t\boldsymbol{f}_{t+k}\right)\\ &= \frac{1}{T-k}\sum_{t=1}^{T-k}\left[\left(\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right)\widehat{\boldsymbol{f}}_{t+k} + \boldsymbol{f}_t\left(\widehat{\boldsymbol{f}}_{t+k} - \boldsymbol{f}_{t+k}\right)\right].\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Gamma}}_f(k) - \widetilde{\boldsymbol{\Gamma}}_f(k)\right\|_2 &\leq \left\|\frac{1}{T-k}\sum_{t=1}^{T-k}\left(\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right)\widehat{\boldsymbol{f}}_{t+k}\right\|_2 + \left\|\frac{1}{T-k}\sum_{t=1}^{T-k}\boldsymbol{f}_t\left(\widehat{\boldsymbol{f}}_{t+k} - \boldsymbol{f}_{t+k}\right)\right\|_2\\ &\leq \frac{1}{T-k}\sum_{t=1}^{T-k}\left\|\left(\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right)\widehat{\boldsymbol{f}}_{t+k}\right\|_2 + \frac{1}{T-k}\sum_{t=1}^{T-k}\left\|\boldsymbol{f}_t\left(\widehat{\boldsymbol{f}}_{t+k} - \boldsymbol{f}_{t+k}\right)\right\|_2.\end{aligned}$$

And it is sufficient to consider only one of the two terms on the right-hand side above since the other one can be dealt with in precisely the same way. For the first term on the right-hand side above, notice that under the factor model

defined in (2.3), we have

$$\begin{aligned}\widehat{f}_t - f_t &= \frac{1}{N}\widehat{Q}^\top y_t - f_t \\ &= \frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t + \frac{1}{N}Q^\top y_t - f_t \\ &= \frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t + \frac{1}{N}Q^\top y_t - \frac{1}{N}Q^\top Q f_t \\ &= \frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t + \frac{1}{N}Q^\top u_t.\end{aligned}$$

Hence

$$\left\|\widehat{f}_t - f_t\right\|_2 \leq \left\|\frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t\right\|_2 + \left\|\frac{1}{N}Q^\top u_t\right\|_2,$$

by triangular inequality. To study $\left\|\frac{1}{N}Q^\top u_t\right\|_2$, first consider the random variables $\frac{1}{\sqrt{N}}q_i^\top u_t$ for each $\frac{1}{\sqrt{N}}q_i$ in $\frac{1}{\sqrt{N}}Q = \left(\frac{1}{\sqrt{N}}q_1, \frac{1}{\sqrt{N}}q_2, ..., \frac{1}{\sqrt{N}}q_r\right)$, where $\frac{1}{\sqrt{N}}q_i$ for $i = 1, 2, ..., r$ are unscaled eigenvectors estimated from $\widehat{L}$. Observe that $\mathbb{E}\left(\frac{1}{\sqrt{N}}q_i^\top u_t\right) = 0$ and $\mathbb{V}\left(\frac{1}{\sqrt{N}}q_i^\top u_t\right) = \frac{1}{N}q_i^\top \Sigma_u q_i \leq \lambda_{\max}(\Sigma_u) < \infty$, since $\left\|\frac{1}{\sqrt{N}}q_i\right\|_2 = 1$ and $\lambda_{\max}(\Sigma_u)$ is the largest eigenvalue of Σ_u. Consequently, $\frac{1}{\sqrt{N}}q_i^\top u_t = O_P(1)$ and $\left\|\frac{1}{N}Q^\top u_t\right\|_2 = \sqrt{\frac{1}{N}\sum_{i=1}^r\left(\frac{1}{\sqrt{N}}q_i^\top u_t\right)^2} = O_P\left(N^{-1/2}\right)$, as the eigenvalues of Σ_u are assumed to be bounded when $N \to \infty$ under Assumption 2.1.

Recall that $\left\|\widehat{Q} - Q\right\|_2 = O_P\left(N^{1/2}T^{-1/2}\right)$ by Lemma 2.3, we then have $\left\|\frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t\right\|_2 \leq \frac{1}{N}\left\|\left(\widehat{Q} - Q\right)^\top\right\|_2 \|y_t\|_2 = O_P\left(T^{-1/2}\right)$, and

$$\begin{aligned}\left\|\widehat{f}_t - f_t\right\|_2 &\leq \left\|\frac{1}{N}\left(\widehat{Q} - Q\right)^\top y_t\right\|_2 + \left\|\frac{1}{N}Q^\top u_t\right\|_2 \\ &= O_P\left(N^{-1/2} + T^{-1/2}\right),\end{aligned}$$

uniformly for t. Finally, we can conclude that

$$\begin{aligned}\left\|\widehat{\mathbf{\Gamma}}_f(k)-\widetilde{\mathbf{\Gamma}}_f(k)\right\|_2 &\leq \frac{1}{T-k}\sum_{t=1}^{T-k}\left\|\left(\widehat{\boldsymbol{f}}_t-\boldsymbol{f}_t\right)\widehat{\boldsymbol{f}}_{t+k}\right\|_2+\frac{1}{T-k}\sum_{t=1}^{T-k}\left\|\boldsymbol{f}_t\left(\widehat{\boldsymbol{f}}_{t+k}-\boldsymbol{f}_{t+k}\right)\right\|_2\\ &=O_P\left(N^{-1/2}+T^{-1/2}\right).\end{aligned}$$

□

2.B.2 Auxiliary results for sieve bootstrap of factor models

Lemma 2.5. *Let* $\widetilde{\boldsymbol{A}}_p=\left(\widetilde{\boldsymbol{A}}_{1,p},\widetilde{\boldsymbol{A}}_{2,p},...,\widetilde{\boldsymbol{A}}_{p,p}\right)$ *be the matrix of the Yule-Walker estimators of the finite predictor coefficients on true factors* $\{\boldsymbol{f}_t\}$*, and* $\widehat{\boldsymbol{A}}_p=\left(\widehat{\boldsymbol{A}}_{1,p},\widehat{\boldsymbol{A}}_{2,p},...,\widehat{\boldsymbol{A}}_{p,p}\right)$ *be the matrix of the Yule-Walker estimators of the finite predictor coefficients on estimated factors* $\{\widehat{\boldsymbol{f}}_t\}$*, then*

$$\left\|\widehat{\boldsymbol{A}}_p-\widetilde{\boldsymbol{A}}_p\right\|_F=O_P\left(p^4\left(N^{-1/2}+T^{-1/2}\right)\right).$$

Proof of Lemma 2.5. Recall that the Yule-Walker estimators are solved from the Yule-Walker equations on the finite predictors' coefficients matrices as follows,

$$\boldsymbol{A}_p=\left(\boldsymbol{A}_{1,p},\boldsymbol{A}_{2,p},...,\boldsymbol{A}_{p,p}\right)=\mathbf{\Pi}_1\mathbf{\Pi}_{0,p}^{-1},$$

where $\mathbf{\Pi}_1=\left(\mathbf{\Gamma}_f(1),\mathbf{\Gamma}_f(2),...,\mathbf{\Gamma}_f(p)\right)$ is an $r\times(rp)$ block matrix of autocovariance matrices and

$$\mathbf{\Pi}_{0,p}=\begin{pmatrix}\mathbf{\Gamma}_f(0) & \mathbf{\Gamma}_f(1) & \cdots & \mathbf{\Gamma}_f(p-1)\\ \mathbf{\Gamma}_f(-1) & \mathbf{\Gamma}_f(0) & \cdots & \mathbf{\Gamma}_f(p-2)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{\Gamma}_f(-p+1) & \mathbf{\Gamma}_f(-p+2) & \cdots & \mathbf{\Gamma}_f(0)\end{pmatrix},$$

is then an $(rp)\times(rp)$ block matrix of autocovariance matrices (Brockwell and Davis, 1991). Write $\widehat{\boldsymbol{A}}_p=\left(\widehat{\boldsymbol{A}}_{1,p},\widehat{\boldsymbol{A}}_{2,p},...,\widehat{\boldsymbol{A}}_{p,p}\right)=\widehat{\mathbf{\Pi}}_1\widehat{\mathbf{\Pi}}_{0,p}^{-1}$ with $\widehat{\mathbf{\Pi}}_1$ and $\widehat{\mathbf{\Pi}}_{0,p}$ the same matrices as $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_{0,p}$ but defined based on $\widehat{\mathbf{\Gamma}}_f$ rather than $\mathbf{\Gamma}_f$. Similarly,

$\widetilde{\boldsymbol{A}}_p = \left(\widetilde{\boldsymbol{A}}_{1,p}, \widetilde{\boldsymbol{A}}_{2,p}, ..., \widetilde{\boldsymbol{A}}_{p,p}\right) = \widetilde{\boldsymbol{\Pi}}_1 \widetilde{\boldsymbol{\Pi}}_{0,p}^{-1}$ with $\widetilde{\boldsymbol{\Pi}}_1$ and $\widetilde{\boldsymbol{\Pi}}_{0,p}$ defined based on $\widetilde{\boldsymbol{\Gamma}}_f$ rather than $\boldsymbol{\Gamma}_f$. Recall that $\widehat{\boldsymbol{\Gamma}}_f$ and $\widetilde{\boldsymbol{\Gamma}}_f$ are sample lag-k autocovariance matrices defined in Lemma 2.4, then we have

$$\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F \leq \left\|\widehat{\boldsymbol{\Pi}}_{0,p}^{-1} - \widetilde{\boldsymbol{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widehat{\boldsymbol{\Pi}}_1\right\|_F + \left\|\widetilde{\boldsymbol{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widehat{\boldsymbol{\Pi}}_1 - \widetilde{\boldsymbol{\Pi}}_1\right\|_F. \tag{2.15}$$

To find $\left\|\widetilde{\boldsymbol{\Pi}}_{0,p}^{-1}\right\|_F$, we first compute $\left\|\boldsymbol{\Pi}_{0,p}^{-1}\right\|_F$. Recall the recursive derivation based on the partitioned inverse formula for $\boldsymbol{\Pi}_{0,p+1}^{-1}$ as in Sowell (1989),

$$\begin{aligned}\boldsymbol{\Pi}_{0,p+1}^{-1} &= \begin{pmatrix} \boldsymbol{\Pi}_{0,p}^{-1} + \mathcal{J}_p \overline{\boldsymbol{A}}_p \overline{\boldsymbol{v}}_p^{-1} \overline{\boldsymbol{A}}_p^\top \mathcal{J}_p & -\mathcal{J}_p \overline{\boldsymbol{A}}_p \overline{\boldsymbol{v}}_p^{-1} \\ -\overline{\boldsymbol{v}}_p^{-1} \overline{\boldsymbol{A}}_p^\top \mathcal{J}_p & \overline{\boldsymbol{v}}_p^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Pi}_{0,p}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -\mathcal{J}_p \overline{\boldsymbol{A}}_p \overline{\boldsymbol{v}}_p^{-1/2} \\ 0 & \overline{\boldsymbol{v}}_p^{-1/2} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ -\overline{\boldsymbol{v}}_p^{-1/2} \overline{\boldsymbol{A}}_p^\top \mathcal{J}_p & \overline{\boldsymbol{v}}_p^{-1/2} \end{pmatrix}, \end{aligned} \tag{2.16}$$

where $\mathcal{J}_p = \boldsymbol{J}_p \otimes \boldsymbol{I}_r$ with $\boldsymbol{J}_p$ the $p \times p$ matrix with ones on the anti-diagonal and $\boldsymbol{I}_r$ the $r \times r$ identity matrix, $\overline{\boldsymbol{v}} = \mathbb{E}\left(\boldsymbol{f}_t - \sum_{l=1}^p \overline{\boldsymbol{A}}_{l,p} \boldsymbol{f}_{t+l}\right)\left(\boldsymbol{f}_t - \sum_{l=1}^p \overline{\boldsymbol{A}}_{l,p} \boldsymbol{f}_{t+l}\right)^\top$ and $\overline{\boldsymbol{A}}_p = \left(\overline{\boldsymbol{A}}_{1,p}^\top, \overline{\boldsymbol{A}}_{2,p}^\top, ..., \overline{\boldsymbol{A}}_{p,p}^\top\right)$ the coefficient matrices minimizing the forward prediction variance $\mathbb{E}\left(\boldsymbol{f}_t - \sum_{l=1}^p \mathbf{F}_{l,p} \boldsymbol{f}_{t+l}\right)\left(\boldsymbol{f}_t - \sum_{l=1}^p \mathbf{F}_{l,p} \boldsymbol{f}_{t+l}\right)^\top$. Denoted by $\mathcal{S}_p$ the second term on the right-hand side of (2.16), we can then get the recursive expression of $\boldsymbol{\Pi}_{0,p}^{-1}$ as

$$\boldsymbol{\Pi}_{0,p}^{-1} = \begin{pmatrix} \boldsymbol{\Gamma}_f(0)^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \sum_{l=1}^{p-1} \mathcal{S}_l.$$

For $\mathcal{S}_l$, note that

$$\begin{aligned}\|\mathcal{S}_l\|_F &\leq \left\|\overline{\boldsymbol{v}}_l^{-1/2}\right\|_F^2 \left(1 + \left\|\mathcal{J}_l \overline{\boldsymbol{A}}_l\right\|_F\right)^2 \\ &\leq \left\|\overline{\boldsymbol{v}}_l^{-1/2}\right\|_F^2 \left(1 + \sum_{j=1}^{l} \left\|\overline{\boldsymbol{A}}_{j,l}\right\|_F\right)^2 \\ &= O(1),\end{aligned}$$

uniformly for $l = 1, 2, ..., p$, where we use the definition of $\overline{\boldsymbol{v}}_l$ and Lemma 2.6.

Hence $\left\|\sum_{l=1}^{p-1}\mathcal{S}_l\right\|_F \leq \sum_{l=1}^{p-1}\|\mathcal{S}_l\|_F = O(p)$. Besides, $\left\|\mathbf{\Gamma}_f(0)^{-1}\right\|_F = \sqrt{\sum_{i=1}^{r}\lambda_i^{-2}} \leq \sqrt{r}\lambda_{\min}^{-1} = O(1)$, where λ_i is the i-th eigenvalue of $\mathbf{\Gamma}_f(0)$, $\lambda_{\min}$ is the smallest eigenvalue of $\mathbf{\Gamma}_f(0)$ and we use Assumption 2.2 that $\mathbf{\Gamma}_f(0)$ is full rank. Thus, we have shown $\left\|\mathbf{\Pi}_{0,p}^{-1}\right\|_F = O(p)$.

To find $\|\widehat{\mathbf{\Pi}}_{0,p}^{-1} - \widetilde{\mathbf{\Pi}}_{0,p}^{-1}\|_F$, note that for invertible matrices $\widehat{\mathbf{\Pi}}_{0,p}$ and $\widetilde{\mathbf{\Pi}}_{0,p}$,

$$\begin{aligned}
\left\|\widehat{\mathbf{\Pi}}_{0,p}^{-1} - \widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F &= \left\|\widehat{\mathbf{\Pi}}_{0,p}^{-1}(\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p})\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \\
&= \left\|(\widehat{\mathbf{\Pi}}_{0,p}^{-1} - \widetilde{\mathbf{\Pi}}_{0,p}^{-1})(\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p})\widetilde{\mathbf{\Pi}}_{0,p}^{-1} + \widetilde{\mathbf{\Pi}}_{0,p}^{-1}(\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p})\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \\
&\leq \left\|\widehat{\mathbf{\Pi}}_{0,p}^{-1} - \widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F + \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F^2.
\end{aligned}$$

And for large enough N and T such as $\left\|\widehat{\mathbf{\Gamma}}_f(k) - \widetilde{\mathbf{\Gamma}}_f(k)\right\|_2 \to 0$ and $\left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F \to 0$ in probability, we can write

$$\begin{aligned}
\left\|\widehat{\mathbf{\Pi}}_{0,p}^{-1} - \widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F &\leq \frac{\left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F^2 \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F}{1 - \left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F} \\
&\leq \frac{\left\|\mathbf{\Pi}_{0,p}^{-1}\right\|_F^2 \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F}{1 - \left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F} + \frac{\left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1} - \mathbf{\Pi}_{0,p}^{-1}\right\|_F^2 \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F}{1 - \left\|\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F} \\
&= O_P\left(\left\|\mathbf{\Pi}_{0,p}^{-1}\right\|_F^2 \left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F\right),
\end{aligned}$$

where the last line follows since when $N, T \to \infty$, $\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\|_F \to 0$ in probability and the first term in the second last line is the leading term. In addition, we have

$$\begin{aligned}
\left\|\widetilde{\mathbf{\Pi}}_{0,p} - \widehat{\mathbf{\Pi}}_{0,p}\right\|_F &\leq \sum_{l=1}^{p}\sum_{j=1}^{p}\left\|\widehat{\mathbf{\Gamma}}_f(l-j) - \widetilde{\mathbf{\Gamma}}_f(l-j)\right\|_F \\
&\leq p^2 \max_{|k|\leq p-1}\left\|\widehat{\mathbf{\Gamma}}_f(k) - \widetilde{\mathbf{\Gamma}}_f(k)\right\|_F \\
&= O_P\left(p^{5/2}\left(N^{-1/2} + T^{-1/2}\right)\right), \qquad (2.17)
\end{aligned}$$

where for $r \times r$ matrices $\widehat{\mathbf{\Gamma}}_f(k)$ and $\widetilde{\mathbf{\Gamma}}_f(k)$, $\left\|\widehat{\mathbf{\Gamma}}_f(k) - \widetilde{\mathbf{\Gamma}}_f(k)\right\|_F \asymp \left\|\widehat{\mathbf{\Gamma}}_f(k) - \widetilde{\mathbf{\Gamma}}_f(k)\right\|_2 = O_P\left(N^{-1/2} + T^{-1/2}\right)$ as shown in Lemma 2.4. Therefore, with (2.17) we can con-

clude that

$$\begin{aligned}\left\|\widehat{\mathbf{\Pi}}_{0,p}^{-1}-\widetilde{\mathbf{\Pi}}_{0,p}^{-1}\right\|_F &= O_P\left(\left\|\mathbf{\Pi}_{0,p}^{-1}\right\|_F^2\left\|\widetilde{\mathbf{\Pi}}_{0,p}-\widehat{\mathbf{\Pi}}_{0,p}\right\|_F\right)\\ &= O_P\left(p^4\left(N^{-1/2}+T^{-1/2}\right)\right). \end{aligned} \tag{2.18}$$

Lastly,

$$\begin{aligned}\left\|\widehat{\mathbf{\Pi}}_1\right\|_F &\leq \sum_{k=1}^{p}\left\|\widehat{\mathbf{\Gamma}}_f(k)\right\|_F\\ &\leq \sum_{k=1}^{p}\left\|\mathbf{\Gamma}_f(k)\right\|_F+\sum_{k=1}^{p}\left\|\widehat{\mathbf{\Gamma}}_f(k)-\mathbf{\Gamma}_f(k)\right\|_F\\ &= O(1)+O_P\left(p\left(N^{-1/2}+T^{-1/2}\right)\right), \end{aligned} \tag{2.19}$$

where the first term follows from the summability condition in Assumption 2.2. Moreover,

$$\begin{aligned}\left\|\widehat{\mathbf{\Pi}}_1-\widetilde{\mathbf{\Pi}}_1\right\|_F &\leq \sum_{k=1}^{p}\left\|\widehat{\mathbf{\Gamma}}_f(k)-\widetilde{\mathbf{\Gamma}}_f(k)\right\|_F\\ &= O_P\left(p\left(N^{-1/2}+T^{-1/2}\right)\right). \end{aligned}$$

Hence we can conclude that the first term in (2.15) is the leading term, and

$$\left\|\widehat{\boldsymbol{A}}_p-\widetilde{\boldsymbol{A}}_p\right\|_F = O_P\left(p^4\left(N^{-1/2}+T^{-1/2}\right)\right),$$

by (2.18) and (2.19). □

Lemma 2.6. *Let $\{\boldsymbol{f}_t\}$ be factor processes fulfilling Assumptions 2.1 and 2.2 for some $\gamma \geq 0$. Write $\{\boldsymbol{A}_{l,p}, l = 1,2,...,p\}$ and $\{\mathbf{\Psi}_{l,p}, l = 1,2,...,p\}$ as the finite predictor coefficients matrices of the AR coefficients $\{\boldsymbol{A}_l, l \in \mathbb{N}\}$ and the MA coefficients $\{\mathbf{\Psi}_l, l \in \mathbb{N}\}$ as in (2.3) and (2.4), respectively.*

(i) Norm summability: the coefficients matrices $\boldsymbol{A}_l$ and $\mathbf{\Psi}_l$ fulfil the following summability properties: $\sum_{l=1}^{\infty}(1+l)^{\gamma}\left\|\boldsymbol{A}_l\right\|_F < \infty$ and $\sum_{l=1}^{\infty}(1+l)^{\gamma}\left\|\mathbf{\Psi}_l\right\|_F < \infty$.

(ii) (Lemma 3.1 of Meyer and Kreiss (2015)) For some $\gamma \geq 0$ as in Assumption 2.2,

there exist $p_0 \in \mathbb{N}$ *and* $d < \infty$ *such that*

$$\sum_{l=1}^{p}(1+l)^{\gamma}\left\|\boldsymbol{A}_{l,p}-\boldsymbol{A}_{l}\right\|_{F} \leq d \sum_{l=p+1}^{\infty}(1+l)^{\gamma}\left\|\boldsymbol{A}_{l}\right\|_{F}, \text{ for } p \geq p_0,$$

and the right side converges to 0 when $p \to \infty$.

(iii) (Lemma 3.2 of Meyer and Kreiss (2015)) Let $\boldsymbol{A}_p(z) := \boldsymbol{I}_r - \sum_{l=1}^{p} \boldsymbol{A}_{l,p} z^l$, *then there exist* $p_1 \in \mathbb{N}$ *and* $c < \infty$ *such that*

$$\inf_{|z| \leq 1+1/p} \left|\det\left(\boldsymbol{A}_p(z)\right)\right| \geq c, \text{ for } p \geq p_1.$$

(iv) (Lemma 3.3 of Meyer and Kreiss (2015)) Let $\{\boldsymbol{\Psi}_{l,p}, l \in \mathbb{N}\}$ *be the power series coefficients matrices of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \boldsymbol{A}_{l,p} z^l\right)^{-1}$, *for* $|z| \leq 1$. *For* p_1 *as defined in (iii) and some* $\gamma \geq 0$ *in Assumption 2.2, there exist* $p_2 \geq p_1$ *and* $d < \infty$ *such that*

$$\sum_{l=1}^{\infty}(1+l)^{\gamma}\left\|\boldsymbol{\Psi}_{l,p}-\boldsymbol{\Psi}_{l}\right\|_{F} \leq d \sum_{l=p+1}^{\infty}(1+l)^{\gamma}\left\|\boldsymbol{A}_{l}\right\|_{F}, \text{ for } p \geq p_2,$$

and the right side converges to 0 when $p \to \infty$.

Lemma 2.6 (ii) is the vector form of Baxter's inequality on the AR coefficients matrices $\{\boldsymbol{A}_l\}$ and its finite predictor coefficients matrices $\{\boldsymbol{A}_{l,p}\}$, whereas Lemma 2.6 (iv) relates Baxter's inequality of AR coefficients to the MA coefficients matrices $\{\boldsymbol{\Psi}_l\}$ and its finite predictor coefficients matrices $\{\boldsymbol{\Psi}_{l,p}\}$. The proofs of Lemma 2.6 can be found in Meyer and Kreiss (2015), hence it is omitted here.

Lemma 2.7. *(Lemma 3.5 of Meyer and Kreiss (2015)) Let* $\{\boldsymbol{f}_t\}$ *be factor processes defined under the assumptions of Lemma 2.6 and also fulfil Assumption 2.3. Define* $\boldsymbol{\Psi}_{l,p}$ *as the coefficients matrices in the power series of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \boldsymbol{A}_{l,p} z^l\right)^{-1}$, *for* $|z| \leq 1$ *with* $\boldsymbol{\Psi}_{0,q} := \boldsymbol{I}_r$ *and* $\widetilde{\boldsymbol{\Psi}}_{l,p}$ *as the power series coefficients matrices of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \widetilde{\boldsymbol{A}}_{l,p} z^l\right)^{-1}$, *for* $|z| \leq 1$ *with* $\widetilde{\boldsymbol{\Psi}}_{0,q} := \boldsymbol{I}_r$. *Then, there exists* $p_3 \in \mathbb{N}$ *such that it holds uniformly in* $l \in \mathbb{N}$ *and*

for all $p \geq p_3$,

$$\left\|\widetilde{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F \leq \left(1+\frac{1}{p}\right)^{-l} \frac{1}{p^2} O_P(1).$$

The proof of Lemma 2.7 can be found in Meyer and Kreiss (2015).

Lemma 2.8. *Let* $\{\boldsymbol{f}_t\}$ *be factor processes fulfilling Assumptions 2.1, 2.2* ($\gamma = 1$)*, 2.3 and 2.4. Define* $\{\boldsymbol{\Psi}_{l,p}\}$ *as the coefficients matrices in the power series of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \boldsymbol{A}_{l,p} z^l\right)^{-1}$*, for* $|z| \leq 1$ *with* $\boldsymbol{\Psi}_{0,q} := \boldsymbol{I}_r$*. Similarly, define* $\{\widetilde{\boldsymbol{\Psi}}_{l,p}\}$ *as the power series coefficients matrices of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \widetilde{\boldsymbol{A}}_{l,p} z^l\right)^{-1}$*, for* $|z| \leq 1$ *with* $\widetilde{\boldsymbol{\Psi}}_{0,q} := \boldsymbol{I}_r$*, and* $\{\widehat{\boldsymbol{\Psi}}_{l,p}\}$ *as the power series coefficients matrices of* $\left(\boldsymbol{I}_r - \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p} z^l\right)^{-1}$*, for* $|z| \leq 1$ *with* $\widehat{\boldsymbol{\Psi}}_{0,q} := \boldsymbol{I}_r$*. Then, there exists* $p_3 \in \mathbb{N}$ *such that for all* $p \geq p_3$ *as in Lemma 2.7,*

$$\begin{aligned}
&\sum_{l=1}^{\infty} \left\|\widetilde{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F = O_P\left(\frac{1}{p}\right) = o_P(1),\\
&\sum_{l=1}^{\infty} \left\|\boldsymbol{\Psi}_{l,p} - \boldsymbol{\Psi}_{l}\right\|_F = o(1),\\
&\sum_{l=1}^{\infty} \left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \widetilde{\boldsymbol{\Psi}}_{l,p}\right\|_F = O_P\left(p^{3/2}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right) = o_P(1),\\
&\sum_{l=1}^{\infty} \left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F = o_P(1),
\end{aligned}$$

when $N \to \infty$ *and* $T \to \infty$.

Proof of Lemma 2.8. For large enough N, T and $p > p_3$ as in Lemma 2.7, $\sum_{l=1}^{\infty} \left\|\widetilde{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F$ follows directly from Lemma 2.7 as

$$\begin{aligned}
\sum_{l=1}^{\infty} \left\|\widetilde{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F &\leq \frac{1}{p^2} \sum_{l=1}^{\infty} \left(1+\frac{1}{p}\right)^{-l} O_P(1)\\
&\leq \frac{1}{p^2} \frac{p}{1+p} (1+p) O_P(1)\\
&= O_P\left(\frac{1}{p}\right).
\end{aligned}$$

The order of $\sum_{l=1}^{\infty} \left\| \mathbf{\Psi}_{l,p} - \mathbf{\Psi}_l \right\|_F$ follows directly from Lemma 2.6 (i) and (iv), as

$$\begin{aligned}\sum_{l=1}^{\infty} \left\| \mathbf{\Psi}_{l,p} - \mathbf{\Psi}_l \right\|_F &\leq \sum_{l=1}^{\infty} (1+l)^{\gamma} \left\| \mathbf{\Psi}_{l,p} - \mathbf{\Psi}_l \right\|_F \\ &\leq d \sum_{l=p+1}^{\infty} (1+l)^{\gamma} \left\| \boldsymbol{A}_l \right\|_F \\ &= o(1).\end{aligned}$$

To show $\sum_{l=1}^{\infty} \left\| \widehat{\mathbf{\Psi}}_{l,p} - \widetilde{\mathbf{\Psi}}_{l,p} \right\|_F = o_P(1)$, first notice that

$$\sum_{l=1}^{\infty} \left\| \widehat{\mathbf{\Psi}}_{l,p} - \widetilde{\mathbf{\Psi}}_{l,p} \right\|_F \leq \sum_{l=1}^{\infty} \sum_{u=1}^{r} \sum_{v=1}^{r} \left| \widehat{\Psi}_{l,p}^{(u,v)} - \widetilde{\Psi}_{l,p}^{(u,v)} \right|,$$

where $\widehat{\Psi}_{l,p}^{(u,v)}$ and $\widetilde{\Psi}_{l,p}^{(u,v)}$ are the (u,v)-th elements of the matrices $\widehat{\mathbf{\Psi}}_{l,p}$ and $\widetilde{\mathbf{\Psi}}_{l,p}$, respectively. We then apply Cauchy's inequality for holomorphic functions on the (u,v)-th element of $\widetilde{\mathbf{\Psi}}_{l,p}$ and $\mathbf{\Psi}_{l,p}$, that is

$$\begin{aligned}\left| \widehat{\Psi}_{l,p}^{(u,v)} - \widetilde{\Psi}_{l,p}^{(u,v)} \right| &\leq \left(1+\frac{1}{p}\right)^{-l} \max_{|z|=1+\frac{1}{p}} \left\| \widehat{\boldsymbol{A}}_p^{-1}(z) - \widetilde{\boldsymbol{A}}_p^{-1}(z) \right\|_F \\ &\leq \left(1+\frac{1}{p}\right)^{-l} \left[\max_{|z|=1+\frac{1}{p}} \frac{1}{|\det(\widehat{\boldsymbol{A}}_p(z))|} \left\| \widehat{\boldsymbol{A}}_p^{adj}(z) - \widetilde{\boldsymbol{A}}_p^{adj}(z) \right\|_F \right. \\ &\quad \left. + \max_{|z|=1+\frac{1}{p}} \left| \frac{1}{\det(\widehat{\boldsymbol{A}}_p(z))} - \frac{1}{\det(\widetilde{\boldsymbol{A}}_p(z))} \right| \left\| \widetilde{\boldsymbol{A}}_p^{adj}(z) \right\|_F \right] \\ &=: \left(1+\frac{1}{p}\right)^{-l} \left[\max_{|z|=1+\frac{1}{p}} \mathcal{K}_{1,z} + \max_{|z|=1+\frac{1}{p}} \mathcal{K}_{2,z} \right],\end{aligned}$$

where we use $\boldsymbol{A}^{adj}$ to denote the adjugate matrix of $\boldsymbol{A}$, and write the two terms above as $\mathcal{K}_{1,z}$ and $\mathcal{K}_{2,z}$.

To study $\mathcal{K}_{1,z}$, with Assumption 2.3, Lemmas 2.3 and 2.5, we show that with sufficiently large N and T, we can choose $p > p_3$ such that $\left\| \widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p \right\|_F = o_P(1)$ and $\sup_{|z|\leq 1+\frac{1}{p}} \left\| \widehat{\boldsymbol{A}}_p(z) - \widetilde{\boldsymbol{A}}_p(z) \right\|_F = o_P(1)$.

Furthermore, since determinants are continuous functions of the elements, it

can be extended to $\sup_{|z|\leq 1+\frac{1}{p}} \left|\det \widehat{\boldsymbol{A}}_p(z) - \det \widetilde{\boldsymbol{A}}_p(z)\right| \to 0$ in probability, with

$$\left|\det\left(\widetilde{\boldsymbol{A}}_p(z)\right)\right| \geq c \text{ and } \left|\det\left(\widehat{\boldsymbol{A}}_p(z)\right)\right| \geq c \text{ in probability, for } |z| \leq 1+\frac{1}{p},$$

and for some $c > 0$ as in Lemma 2.6. Then, for $p > p_3$ and any $|z| = 1 + 1/p$ we can show that

$$\begin{aligned}
\mathcal{K}_{1,z} &\leq \frac{1}{c}\left\|\widehat{\boldsymbol{A}}_p^{adj}(z) - \widetilde{\boldsymbol{A}}_p^{adj}(z)\right\|_F \\
&\leq \frac{1}{c}\sum_{u=1}^{r}\sum_{v=1}^{r}\left|\widehat{\boldsymbol{A}}_p^{adj}(z)^{(u,v)} - \widetilde{\boldsymbol{A}}_p^{adj}(z)^{(u,v)}\right| \\
&\leq \frac{1}{c}\sum_{u=1}^{r}\sum_{v=1}^{r}\sup_{|z|\leq 1+\frac{1}{p}}\left|\det \widehat{\boldsymbol{A}}_p^{(-v,-u)}(z) - \det \widetilde{\boldsymbol{A}}_p^{(-v,-u)}(z)\right| \\
&\leq \frac{1}{c}\sum_{u=1}^{r}\sum_{v=1}^{r}\sup_{|z|\leq 1+\frac{1}{p}} r\left\|\widehat{\boldsymbol{A}}_p(z) - \widetilde{\boldsymbol{A}}_p(z)\right\|_F O_P(1) \\
&\leq \sup_{|z|\leq 1+\frac{1}{p}}\left\|\widehat{\boldsymbol{A}}_p(z) - \widetilde{\boldsymbol{A}}_p(z)\right\|_F,
\end{aligned}$$

where $\widetilde{\boldsymbol{A}}_p^{(-v,-u)}(z)$ is a matrix generated by removing the v-th row and the u-th column of $\widetilde{\boldsymbol{A}}_p(z)$.

And for $\sup_{|z|\leq 1+\frac{1}{p}}\left\|\widehat{\boldsymbol{A}}_p(z) - \widetilde{\boldsymbol{A}}_p(z)\right\|_F$, we have

$$\begin{aligned}
\sup_{|z|\leq 1+\frac{1}{p}}\left\|\widehat{\boldsymbol{A}}_p(z) - \widetilde{\boldsymbol{A}}_p(z)\right\|_F &\leq \sup_{|z|\leq 1+\frac{1}{p}}\sum_{l=1}^{p}\left\|\widehat{\boldsymbol{A}}_{l,p} - \widetilde{\boldsymbol{A}}_{l,p}\right\|_F |Z|^l \\
&\leq \left(1+\frac{1}{p}\right)^p \sum_{l=1}^{p}\left\|\widehat{\boldsymbol{A}}_{l,p} - \widetilde{\boldsymbol{A}}_{l,p}\right\|_F \\
&= O_P\left(\sqrt{p}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right).
\end{aligned}$$

Hence we can conclude that for $\mathcal{K}_{1,z}$,

$$\max_{|z|=1+\frac{1}{p}} \mathcal{K}_{1,z} = O_P\left(\sqrt{p}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right),$$

since the bound does not depend on z.

For $\mathcal{K}_{2,z}$, note that $\max_{|z|=1+\frac{1}{p}} \left\|\boldsymbol{A}_p(z)\right\|_F \leq (1+1/p)^p \sum_{l=1}^p \left\|\boldsymbol{A}_{l,p}\right\|_F = O_P(1)$ by Lemma 2.6, therefore, $\max_{|z|=1+\frac{1}{p}} \left\|\widetilde{\boldsymbol{A}}_p(z)\right\|_F = O_P(1)$ by Assumption 2.3. Similarly, for some constants c,

$$\begin{aligned}\max_{|z|=1+\frac{1}{p}} \mathcal{K}_{2,z} &\leq \frac{1}{c^2} \max_{|z|=1+\frac{1}{p}} \left|\det \widehat{\boldsymbol{A}}_p(z) - \det \widetilde{\boldsymbol{A}}_p(z)\right| \left\|\widetilde{\boldsymbol{A}}_p^{adj}(z)\right\|_F \\ &= O_P\left(\sqrt{p}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right).\end{aligned}$$

As a result,

$$\begin{aligned}\sum_{l=1}^{\infty}\left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \widetilde{\boldsymbol{\Psi}}_{l,p}\right\|_F &\leq \sum_{l=1}^{\infty}\sum_{u=1}^{r}\sum_{v=1}^{r} |\widehat{\Psi}_{l,p}^{(u,v)} - \widetilde{\Psi}_{l,p}^{(u,v)}| \\ &= O_P\left(p^{3/2}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right).\end{aligned}$$

Then, we can conclude that

$$\begin{aligned}\sum_{l=1}^{\infty}\left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F &\leq \sum_{l=1}^{\infty}\left\|\widetilde{\boldsymbol{\Psi}}_{l,p} - \boldsymbol{\Psi}_{l,p}\right\|_F + \sum_{l=1}^{\infty}\left\|\widehat{\boldsymbol{\Psi}}_{l,p} - \widetilde{\boldsymbol{\Psi}}_{l,p}\right\|_F \\ &= O_P\left(\frac{1}{p}\right) + O_P\left(p^{3/2}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right).\end{aligned}$$

□

Lemma 2.9. *Let $\{\boldsymbol{f}_t\}$ be factor processes defined under the assumptions of Lemma 2.8. Write $\boldsymbol{e}_t = \boldsymbol{f}_t - \sum_{l=1}^{\infty} \boldsymbol{A}_l \boldsymbol{f}_{t-l}$, $\boldsymbol{e}_{t,p} = \boldsymbol{f}_t - \sum_{l=1}^{p} \boldsymbol{A}_{l,p}\boldsymbol{f}_{t-l}$, $\widetilde{\boldsymbol{e}}_{t,p} = \boldsymbol{f}_t - \sum_{l=1}^{p} \widetilde{\boldsymbol{A}}_{l,p}\boldsymbol{f}_{t-l}$ and $\widehat{\boldsymbol{e}}_{t,p} = \widehat{\boldsymbol{f}}_t - \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_{l,p}\widehat{\boldsymbol{f}}_{t-l}$. Furthermore, define the corresponding covariance $\widetilde{\boldsymbol{\Sigma}}_{e,p} = \mathbb{E}^*(\widetilde{\boldsymbol{e}}_{t,p} - \overline{\widetilde{\boldsymbol{e}}}_{T,p})(\widetilde{\boldsymbol{e}}_{t,p} - \overline{\widetilde{\boldsymbol{e}}}_{T,p})^\top$ with $\overline{\widetilde{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widetilde{\boldsymbol{e}}_{t,p}$, and $\widehat{\boldsymbol{\Sigma}}_{e,p} = \mathbb{E}^*(\widehat{\boldsymbol{e}}_{t,p} - \overline{\widehat{\boldsymbol{e}}}_{T,p})(\widehat{\boldsymbol{e}}_{t,p} - \overline{\widehat{\boldsymbol{e}}}_{T,p})^\top$ with $\overline{\widehat{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\boldsymbol{e}}_{t,p}$, where $\mathbb{E}^*$ is the expectation defined on the measure of assigning probability $\frac{1}{T-p}$ to each observation.*

If we additionally assume that the empirical distribution of $\{\boldsymbol{e}_t\}$ converges weakly to the distribution function of $\mathcal{L}(\boldsymbol{e}_t)$, then, there exists $p_3 \in \mathbb{N}$ such that for all $p \geq p_3$ as

in Lemma 2.7,

$$\begin{aligned}
&\|\widetilde{\mathbf{\Sigma}}_{e,p} - \mathbf{\Sigma}_{e,p}\|_F = o_P(1),\\
&\|\mathbf{\Sigma}_{e,p} - \mathbf{\Sigma}_e\|_F = o(1),\\
&\|\widehat{\mathbf{\Sigma}}_{e,p} - \widetilde{\mathbf{\Sigma}}_{e,p}\|_F = O_P\left(p^{3/2}\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F\right) = o_P(1),\\
&\|\widehat{\mathbf{\Sigma}}_{e,p} - \mathbf{\Sigma}_{e,p}\|_F = o_P(1),
\end{aligned}$$

when $N \to \infty$ and $T \to \infty$.

Proof of Lemma 2.9. To show $\left\|\widetilde{\mathbf{\Sigma}}_{e,p} - \mathbf{\Sigma}_{e,p}\right\|_F \to 0$ in probability, first note that by definition,

$$\begin{aligned}
\left\|\widetilde{\mathbf{\Sigma}}_{e,p} - \mathbf{\Sigma}_{e,p}\right\|_F = &\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widetilde{\boldsymbol{e}}_{t,p}\widetilde{\boldsymbol{e}}_{t,p}^\top - \boldsymbol{e}_{t,p}\boldsymbol{e}_{t,p}^\top\right)\right\|_F\\
&+ \left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\boldsymbol{e}_{t,p}\boldsymbol{e}_{t,p}^\top - \mathbb{E}\left(\boldsymbol{e}_{t,p}\boldsymbol{e}_{t,p}^\top\right)\right\|_F\\
&+ \left\|\widetilde{\boldsymbol{e}}_{T,p}\widetilde{\boldsymbol{e}}_{T,p}^\top\right\|_F\\
&=: \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3,
\end{aligned}$$

with straightforward notations for $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$. Next, we show that the three terms above converge to zero in probability. For $\mathcal{E}_1$, we know that by triangular inequality,

$$\begin{aligned}
\mathcal{E}_1 &\leq \left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\left(\widetilde{\boldsymbol{e}}_{t,p} - \boldsymbol{e}_{t,p}\right)\widetilde{\boldsymbol{e}}_{t,p}^\top\right\|_F + \left\|\frac{1}{T-p}\sum_{t=p+1}^{T}\boldsymbol{e}_{t,p}\left(\widetilde{\boldsymbol{e}}_{t,p} - \boldsymbol{e}_{t,p}\right)^\top\right\|_F\\
&=: \mathcal{E}_{1,1} + \mathcal{E}_{1,2},
\end{aligned}$$

with obvious notations for $\mathcal{E}_{1,1}$ and $\mathcal{E}_{1,2}$. It is then sufficient to show $\mathcal{E}_{1,1} \to 0$ in probability since $\mathcal{E}_{1,2}$ can be dealt with in a similar way. We can now bound $\mathcal{E}_{1,1}$

by

$$\begin{aligned}\mathcal{E}_{1,1} \leq & \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=1}^{p} \left(\widetilde{\boldsymbol{A}}_{l,p} - \boldsymbol{A}_{l,p} \right) \boldsymbol{f}_{t-l} \widetilde{\boldsymbol{e}}_{t,p}^{\top} \right\|_F \\ & + \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=1}^{p} \left(\boldsymbol{A}_{l,p} - \boldsymbol{A}_{l} \right) \boldsymbol{f}_{t-l} \widetilde{\boldsymbol{e}}_{t,p}^{\top} \right\|_F \\ & + \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=p+1}^{\infty} \boldsymbol{A}_{l} \boldsymbol{f}_{t-l} \widetilde{\boldsymbol{e}}_{t,p}^{\top} \right\|_F .\end{aligned}$$

Since both $\{\boldsymbol{f}_t\}$ and $\{\widetilde{\boldsymbol{e}}_{t,p}\}$ are $r \times 1$ vectors, by Assumption 2.3 and Lemma 2.6, we have

$$\mathcal{E}_{1,1} = O_P \left(\left\| \sum_{l=1}^{p} \left(\widetilde{\boldsymbol{A}}_{l,p} - \boldsymbol{A}_{l,p} \right) \right\|_F + \sum_{l=p+1}^{\infty} (1+l) \left\| \boldsymbol{A}_l \right\|_F \right),$$

which tends to zero in probability.

$\mathcal{E}_2 \to 0$ in probability can be shown similarly, since $\{\boldsymbol{f}_t\}$ is stationary. For $\mathcal{E}_3$, first write that

$$\begin{aligned}\mathcal{E}_3 &= \left\| \bar{\widetilde{\boldsymbol{e}}}_{T,p} \bar{\widetilde{\boldsymbol{e}}}_{T,p}^{\top} \right\|_F \\ &\leq \left\| \left(\bar{\widetilde{\boldsymbol{e}}}_{T,p} - \bar{\boldsymbol{e}}_{T,p} \right) \left(\bar{\widetilde{\boldsymbol{e}}}_{T,p} - \bar{\boldsymbol{e}}_{T,p} \right)^{\top} \right\|_F + 2 \left\| \left(\bar{\widetilde{\boldsymbol{e}}}_{T,p} - \bar{\boldsymbol{e}}_{T,p} \right) \bar{\boldsymbol{e}}_{T,p}^{\top} \right\|_F + \left\| \bar{\boldsymbol{e}}_{T,p} \bar{\boldsymbol{e}}_{T,p}^{\top} \right\|_F ,\end{aligned}$$

where $\left\| \bar{\boldsymbol{e}}_{T,p} \right\| = O_P \left((T-p)^{-1/2} \right)$. Hence it is sufficient to consider $\left\| \bar{\widetilde{\boldsymbol{e}}}_{T,p} - \bar{\boldsymbol{e}}_{T,p} \right\|$

as

$$\begin{aligned}
\left\|\widetilde{\bar{e}}_{T,p} - \bar{e}_{T,p}\right\| &= \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \left(\widetilde{e}_{T,p} - e_{T,p}\right) \right\| \\
&= \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \left(\sum_{l=1}^{p} \widetilde{A}_{l,p} f_{t-l} - \sum_{l=1}^{\infty} A_l f_{t-l} \right) \right\| \\
&\leq \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=1}^{p} \left(\widetilde{A}_{l,p} - A_{l,p} \right) f_{t-l} \right\| \\
&+ \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=1}^{p} \left(A_{l,p} - A_l \right) f_{t-l} \right\| + \left\| \frac{1}{T-p} \sum_{t=p+1}^{T} \sum_{l=p+1}^{\infty} A_l f_{t-l} \right\| \\
&= O_P\left(\left\| \sum_{l=1}^{p} \left(\widetilde{A}_{l,p} - A_{l,p} \right) \right\|_F \right) + O_P\left(\sum_{l=p+1}^{\infty} (1+l) \left\| A_l \right\|_F \right) \xrightarrow{p} 0,
\end{aligned}$$

where the last line follows from Assumption 2.3 and Lemma 2.6, and we use the same arguments for $\mathcal{E}_{1,1}$ as above. Therefore, we can conclude that $\left\|\widetilde{\Sigma}_{e,p} - \Sigma_{e,p}\right\|_F \to 0$ in probability.

To see $\left\|\Sigma_{e,p} - \Sigma_e\right\|_F \to 0$, note that

$$\begin{aligned}
\left\|\Sigma_{e,p} - \Sigma_e\right\|_F &= \left\| \mathbb{E}\left(e_{t,p} e_{t,p}^\top - e_t e_t^\top \right) \right\|_F \\
&\leq \left\| \mathbb{E}\left\{ \left(e_{t,p} - e_t \right) e_{t,p}^\top \right\} \right\|_F + \left\| \mathbb{E}\left\{ e_{t,p} \left(e_{t,p} - e_t \right)^\top \right\} \right\|_F .
\end{aligned}$$

Hence it suffices to show $\left\| \mathbb{E}\left\{ \left(e_{t,p} - e_t \right) e_{t,p}^\top \right\} \right\|_F \to 0$. For this, by triangular inequality, we have

$$\begin{aligned}
\left\| \mathbb{E}\left\{ \left(e_{t,p} - e_t \right) e_{t,p}^\top \right\} \right\|_F &\leq \left\| \mathbb{E} \sum_{l=1}^{p} \left(A_{l,p} - A_l \right) f_{t-l} e_{t,p}^\top \right\|_F + \left\| \mathbb{E} \sum_{l=p+1}^{\infty} A_l f_{t-l} e_{t,p}^\top \right\|_F \\
&= O\left(\sum_{l=1}^{p} \left\| A_{l,p} - A_l \right\|_F \right) + O\left(\sum_{l=p+1}^{\infty} \left\| A_l \right\|_F \right) \to 0,
\end{aligned}$$

where we stress the fact that $\|f_t\| \asymp \|e_{t,p}\| \asymp 1$ and use the results in Lemma 2.6.

With similar arguments, we can show that $\left\|\widehat{\Sigma}_{e,p} - \widetilde{\Sigma}_{e,p}\right\|_F \to 0$ in probability.

Firstly, notice that $\left(\widehat{\boldsymbol{\Sigma}}_{e,p} - \widetilde{\boldsymbol{\Sigma}}_{e,p}\right)$ can be expressed as

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{e,p} - \widetilde{\boldsymbol{\Sigma}}_{e,p} &= \frac{1}{T-p}\sum_{t=p+1}^{T}\left[\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right)\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right)^{\top} - \left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)\left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)^{\top}\right] \\
&= \frac{1}{T-p}\sum_{t=p+1}^{T}\left[\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right) - \left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)\right]\left(\widehat{\boldsymbol{e}}_{t,p} - \widetilde{\boldsymbol{e}}_{t,p}\right)^{\top} \\
&\quad - \frac{1}{T-p}\sum_{t=p+1}^{T}\left[\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right) - \left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)\right]\left(\bar{\widehat{\boldsymbol{e}}}_{T,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)^{\top} \\
&\quad + \frac{1}{T-p}\sum_{t=p+1}^{T}\left[\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right)\left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)^{\top}\right] \\
&\quad + \frac{1}{T-p}\sum_{t=p+1}^{T}\left[\left(\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{T,p}\right)\left(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{T,p}\right)^{\top}\right].
\end{aligned}$$

Recall that $\bar{\widetilde{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widetilde{\boldsymbol{e}}_{t,p}$ and $\bar{\widehat{\boldsymbol{e}}}_{T,p} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\boldsymbol{e}}_{t,p}$, by triangular inequality, it suffices to study the leading term $\frac{1}{T-p}\sum_{t=p+1}^{T}\left[(\widehat{\boldsymbol{e}}_{t,p} - \bar{\widehat{\boldsymbol{e}}}_{t,p}) - (\widetilde{\boldsymbol{e}}_{t,p} - \bar{\widetilde{\boldsymbol{e}}}_{t,p})\right](\widehat{\boldsymbol{e}}_{t,p} - \widetilde{\boldsymbol{e}}_{t,p})^{\top}$. For this, it is sufficient to consider the order of $\left\|\frac{1}{T-p}\sum_{t=p+1}^{T}(\widehat{\boldsymbol{e}}_{t,p} - \widetilde{\boldsymbol{e}}_{t,p})(\widehat{\boldsymbol{e}}_{t,p} - \widetilde{\boldsymbol{e}}_{t,p})^{\top}\right\|_F$. We then have the bound

$$\begin{aligned}
\frac{1}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{e}}_{t,p} - \widetilde{\boldsymbol{e}}_{t,p}\right\|^2 &\leq 3\sum_{l=1}^{p}\left\|\widehat{\boldsymbol{A}}_{l,p} - \widetilde{\boldsymbol{A}}_{l,p}\right\|_F^2 \frac{1}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{f}}_{t-l}\right\|^2 \\
&\quad + \frac{3}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right\|^2 + 3\sum_{l=1}^{p}\left\|\widetilde{\boldsymbol{A}}_{l,p}\right\|_F^2 \frac{1}{T-p}\sum_{t=p+1}^{T}\left\|\widehat{\boldsymbol{f}}_{t-l} - \boldsymbol{f}_{t-l}\right\|^2 \\
&= O_P\left(\left\|\widehat{\boldsymbol{A}}_p - \widetilde{\boldsymbol{A}}_p\right\|_F^2\right) + O_P\left(p\left\|\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right\|^2\right),
\end{aligned}$$

which converges to 0 in probability by the results of Lemmas 2.4 and 2.5. Hence we can conclude that $\left\|\widehat{\boldsymbol{\Sigma}}_{e,p} - \widetilde{\boldsymbol{\Sigma}}_{e,p}\right\|_F \to 0$ in probability.

Lastly, $\left\|\widehat{\boldsymbol{\Sigma}}_{e,p} - \boldsymbol{\Sigma}_{e,p}\right\|_F = o_P(1)$ follows directly from $\left\|\widehat{\boldsymbol{\Sigma}}_{e,p} - \widetilde{\boldsymbol{\Sigma}}_{e,p}\right\|_F = o_P(1)$, $\left\|\widetilde{\boldsymbol{\Sigma}}_{e,p} - \boldsymbol{\Sigma}_{e,p}\right\|_F = o_P(1)$, and the triangular inequality. □

Chapter 3

Homogeneity and Sub-homogeneity Pursuit: Iterative Complement Clustering PCA

3.1 Introduction

Since its introduction, principal component analysis (PCA) (Jolliffe, 2002; Anderson, 2003) has become one of the most popular statistical tools for data analysis in a wide range of areas. Literature on PCA dates back to the early twentieth century (e.g., see Pearson (1901) and Hotelling (1933)). However, with an increasing dimension of data, PCA has now been reconsidered and widely discussed for the purpose of analyzing high-dimensional data. Jolliffe and Cadima (2016) and Fan et al. (2018) reviewed recent developments in PCA for statistical analysis on high-dimensional data, including its sparsity and robustness. However, are more data together really benefiting statistical analysis? Boivin and Ng (2006) provided a negative answer at the aspect of forecasting. This is due to an increased complexity when more data from different populations are grouped together, as a proportion of data can exhibit a different pattern compared with the rest. More specifically, increased complexity refers to heterogeneity when more data are collected from different populations.

In this study, we consider that information from a particular group of data collected from one population can be divided into two categories. One is shared with other groups and forms the homogeneity within the entire data,

while the other is group-specific and is the main source of heterogeneity. This type of heterogeneity can be treated as sub-homogeneity, which refers to the homogeneity for a particular group of data.

However, sub-homogeneity may not be identified using traditional estimation methods such as PCA. One reason for this is that the sub-homogeneity for a particular group of data can be relatively small compared with the homogeneity within the entire data because it usually contributes less than the homogeneity to the total variance of all of the data. In such a situation, traditional PCA may regard the sub-homogeneity as negligible compared with homogeneity and ignore this group-specific pattern. Moreover, from the interpretation perspective, principal components that are produced using traditional PCA on all of the data do not target a specific group (e.g., information on which component corresponds to which group of data is not known). In previous studies, the discussion of PCA has mainly focused on the large eigenvalues, which correspond to the homogeneity in this chapter. For example, Johnstone (2001) studied the asymptotic distribution of the largest eigenvalue of PCA. Recently, Cai et al. (2017) extended the discussion to the asymptotic distribution of the spiked eigenvalues, while Morales-Jimenez et al. (2018) studied the asymptotics for the leading eigenvalues and eigenvectors of the correlation matrix in the class of spiked models. None of them have considered the existence of sub-homogeneity within the data.

To the best of our knowledge, the sub-homogeneity has not been well discussed, but it can be very important in high-dimensional data analysis. The following example of analyzing stock returns from different industries is used to explain the importance of finding sub-homogeneity. As stated in Fama and French (1997), stock returns from various industries can have varying performance over time, although the homogeneity (e.g., market return) can be deemed to be driven by some common economic variables. In a situation in which a vast number of individual stock returns are collected together for statistical analysis, traditional dimension-reduction techniques such as PCA may be able to capture the market effect but can fail to identify the sub-homogeneity within

each industry (industry-specific pattern). This is because some industry-specific components may have much smaller variance than the market component and are highly likely to be omitted by PCA. However, these industry-specific components may be very important in capturing the movement of the stock returns within the industry. This loss of information may result in a very poor forecast of stock returns for some companies in which sub-homogeneity exists. In addition, although it would be interesting to study which industry has a larger industry-specific effect on stock returns, traditional PCA performed on the entire data does not allow us to draw such conclusions. Therefore, this study aims at identifying both homogeneity and sub-homogeneity in high-dimensional data analysis.

The sub-homogeneity in several parts of the data can be identified and estimated by dividing the whole data set into several groups. However, in most situations, the group structure is not known in advance. Therefore, a clustering method can be used to group the data at the first stage, and PCA can then be applied to identify the sub-homogeneity within each group. Similarly, Liu et al. (2002) performed PCA in different blocks of variables, while Tan et al. (2015) penalised the variables in each group differently when using the graphical lasso. Both studies used hierarchical clustering to group the variables to take into account the heterogeneity. However, when homogeneity exists across different groups, the sub-homogeneity in each group cannot be successfully identified. This is mainly because the homogeneity shared by most groups can dominate the sub-homogeneity so that the data from different groups all seem to be highly correlated. In this situation, the sub-homogeneity is masked by the homogeneity, and these clustering methods tend to group all individuals (variables) into one cluster. Therefore, the homogeneity must be correctly identified and removed before we can successfully group the individuals and discover the sub-homogeneity of each group.

On the other hand, the estimation of homogeneity can often be problematic in a high-dimensional setting. This is because of the inconsistency of PCA estimates when the number of variables p is comparable with or greater than the sample

size n, as discussed in Johnstone and Lu (2009). Motivated by the group structure of the data, we suggest first clustering the individuals into groups and then performing a traditional approach such as PCA on the level of groups, followed by a second layer of PCA that extracts common information shared by each group. This approach can improve the estimation accuracy of the homogeneity because the components are now extracted from groups in which the dimension has already been reduced (to the number of variables in a group), while the traditional dimension-reduction method (e.g., PCA) is performed on the full dimension p. This can be viewed as an effective way to alleviate the potential problem caused by the curse of dimensionality. Other studies have modified the traditional approaches to deal with the curse of dimensionality. For example, Johnstone and Lu (2009) suggest that some initial dimension-reduction work is necessary before using PCA, as long as a sparse basis exists. Further, Zou et al. (2006) introduced sparse PCA, which uses the lasso (elastic net) to modify the principal components so that sparse loadings can be achieved. In addition, a review of some sparse versions of PCA can be found in Tibshirani et al. (2015), while some general discussions about the blessing and curse of dimensionality can be found in Donoho et al. (2000) and Fan et al. (2014).

To conclude, homogeneity must be removed to correctly identify the cluster structure and sub-homogeneity, but the cluster structure of the data is used to accurately find the homogeneity. Therefore, we introduce a novel "iterative complement-clustering principal component analysis" (CPCA) to iteratively estimate the homogeneity and sub-homogeneity. Details of the CPCA are provided in Section 3.4.

The contributions of this study can be summarised as follows. First, we propose CPCA to identify homogeneity and sub-homogeneity and handle the interaction between them when the whole data set exhibits a group structure. Second, our proposed estimation method not only correctly captures the sub-homogeneity, but also provides very reliable cluster information (e.g., which part of the data is from the same group), which can be useful in understanding and explaining the data structure. Third, inspired by Chiou and Li (2007), we develop

a leave-one-out principal component regression (PCR) clustering method that can outperform the hierarchical clustering method used in previous studies. In addition, we theoretically illustrate that if the sub-homogeneity from different clusters is distinct, our proposed clustering procedure can effectively separate the variables from different clusters. This is also numerically confirmed by the simulation study and real data analysis. Details of the proposed clustering method are provided in Section 3.4.

The rest of this study is organised as follows. A low-rank representation of the data that captures both homogeneity and sub-homogeneity is introduced in Section 3.2, and some related PCA methods are discussed in Section 3.3. In Section 3.4, a novel estimation method called CPCA is proposed, followed by a discussion of more details of the algorithm. Section 3.5 demonstrates and explains the effectiveness of the proposed clustering method. Extensive simulations, along with two applications (PCR and covariance estimation) of our proposed method are provided in Sections 3.6 and 3.7. Section 3.8 analyses a stock return dataset using our method. Lastly, Section 3.9 concludes the study.

3.2 Homogeneity and sub-homogeneity

Considering the data $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})^\top \in \mathbb{R}^p$ from i^{th} observation with p variables, the singular value decomposition (SVD) of the data $\boldsymbol{x}_i$ can be written as:

$$\begin{aligned} \boldsymbol{x}_i &= \sum_{k=1}^{K} g_{ik}\boldsymbol{\phi}_k + \boldsymbol{u}_i, \\ \text{where } \boldsymbol{u}_i &= \sum_{k=K+1}^{\min(n,p)} g_{ik}\boldsymbol{\phi}_k \quad i = 1, \ldots, n, \end{aligned} \tag{3.1}$$

where g_{ik} is defined as k^{th} principal component score and $\boldsymbol{\phi}_k$ denotes a $p \times 1$ eigenvector for g_{ik}. Traditional PCA summaries the data using the first k principal components, and it treats $\boldsymbol{u}_i$ as noise because g_{ik} for $k = m+1, \ldots, \min(n, p)$ has lower variance. However, under certain conditions, some of the information

contained in $\boldsymbol{u}_i$ can be useful in prediction or forecasting problems, particularly for one or more specific groups of data. Therefore, when the data exhibit a group structure, we propose the following low-rank representation for the data to capture both homogeneity and sub-homogeneity:

$$\begin{aligned} \boldsymbol{x}_i &= \sum_{k=1}^{r_c} g_{ik}\boldsymbol{\phi}_k + \boldsymbol{u}_i, \\ \text{where } \boldsymbol{u}_i &= \sum_{j=1}^{J}\sum_{h=1}^{r_j} f_{ih}^{(j)}\boldsymbol{\gamma}_h^{(j)} + \sum_{j=1}^{J}\boldsymbol{I}^{(j)}\boldsymbol{\epsilon}_i^{(j)}, \quad i=1,\ldots,n, \quad j=1,\ldots,J. \end{aligned} \tag{3.2}$$

The first line in (3.2) measures the homogeneity among all variables from all groups, where $g_{ik}, k = 1,\ldots,r_c$ is k^{th} principal component, which we call the common component, and $\boldsymbol{\phi}_k$ is its corresponding eigenvector. The first part of $\boldsymbol{u}_i$ in (3.2) consists of J cluster-specific components from which the sub-homogeneity of the data is derived. Assuming p variables can be split into J clusters, $f_{ih}^{(j)}, h = 1,\ldots,r_j$ measures the within-cluster principal components for cluster j. The within-cluster eigenvector with dimension $p \times 1$ has the form of $\boldsymbol{\gamma}_h^{(j)} = (\mathbf{0}^{(1)\top},\ldots,\boldsymbol{\eta}_h^{(j)\top},\ldots,\mathbf{0}^{(J)\top})^\top$, where $\boldsymbol{\eta}_h^{(j)}$ defines a $p_j \times 1$ vector and p_j is the number of variables in cluster j so that $\sum_{j=1}^{J} p_j = p$. That is, the values of $\boldsymbol{\gamma}_h^{(j)}$ for variables that do not belong to cluster j are zero. This implies that after removing the effect of the common principal components, data from different clusters are uncorrelated (e.g., $\boldsymbol{u}_i$ has a block-diagonal covariance structure). In the second part of $\boldsymbol{u}_i$, $\boldsymbol{\epsilon}_i^{(j)}$ is simply a p_j-dimensional error that has variance ${\sigma^{(j)}}^2$, and $\boldsymbol{I}^{(j)}$ is a diagonal matrix, but the diagonals for variables that do not belong to cluster j are zero.

We further define $\boldsymbol{g}_i = (g_{i1},\ldots,g_{ir_c})^\top$ and $\boldsymbol{f}_i^{(j)} = (f_{i1}^{(j)},\ldots,f_{ir_j}^{(j)})^\top, j = 1,\ldots,J$ as the vector forms of the common components and cluster-specific components, and the eigenvector matrices as $\boldsymbol{\Phi}_{(p\times r_c)} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{r_c})$ and $\boldsymbol{\Gamma}_{(p\times r_j)}^{(j)} = (\boldsymbol{\gamma}_1^{(j)},\ldots,\boldsymbol{\gamma}_{r_j}^{(j)}), j = 1,\ldots,J$, respectively. Without loss of generality, $E(\boldsymbol{g}_i) =$

$E(\boldsymbol{f}_i^{(j)}) = E(\boldsymbol{\epsilon}_i^{(j)}) = \mathbf{0}$ and $\boldsymbol{g}_i, \boldsymbol{f}_i^{(j)}, \boldsymbol{\epsilon}_i^{(j)}, j = 1, \ldots, J$ are mutually uncorrelated and:

$$\begin{aligned} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} &= I_p, \quad \text{cov}(\boldsymbol{g}_i) \text{ is diagonal;} \\ {\boldsymbol{\Gamma}^{(j)}}^\top \boldsymbol{\Gamma}^{(j)} &= I_p, \quad \text{cov}(\boldsymbol{f}_i^{(j)}) \text{ is diagonal,} \quad j = 1, \ldots, J. \end{aligned} \tag{3.3}$$

Under the data structure given in (3.2), the population covariance is given by a low-rank plus block-diagonal representation:

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi} \text{cov}\left(\boldsymbol{g}_i\right) \boldsymbol{\Phi}^\top + \sum_{j=1}^{J} \boldsymbol{\Gamma}^{(j)} \text{cov}(\boldsymbol{f}_i^{(j)}) \boldsymbol{\Gamma}^{(j)\top} + \sum_{j=1}^{J} \boldsymbol{I}^{(j)} \sigma^{(j)2}, \quad j = 1, \ldots, J. \tag{3.4}$$

If we denote the data $\boldsymbol{X}_{(n\times p)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, the common components and cluster-specific components in a matrix form $\boldsymbol{G}_{(n\times r_c)} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)^\top$ and $\boldsymbol{F}^{(j)}_{(n\times r_j)} = (\boldsymbol{f}_1^{(j)}, \ldots, \boldsymbol{f}_n^{(j)})^\top$, respectively, data representation (3.2) can also be presented as a matrix form:

$$\begin{aligned} \boldsymbol{X} =& \boldsymbol{G}\boldsymbol{\Phi}^\top + \boldsymbol{U}, \\ \text{where } \boldsymbol{U} =& \sum_{j=1}^{J} \boldsymbol{F}^{(j)} \boldsymbol{\Gamma}^{(j)\top} + \boldsymbol{E}, \quad j = 1, \ldots, J. \end{aligned} \tag{3.5}$$

In general, there are two goals in using representation (3.2). In the presence of an unknown cluster structure, the first goal is to cluster the variables (i.e., determine which variables are in the same group). From the interpretation perspective, it is interesting to know how variables are clustered and which variables belong to the same cluster. The second goal is to correctly estimate both the common components $\boldsymbol{g}_i$ and the cluster-specific components $\boldsymbol{f}_i^{(j)}$. These components serve as a low-rank representation of the data and can therefore be used for further applications. One obvious application is to estimate $\boldsymbol{\Sigma}$ as in (3.4). The use of both $\boldsymbol{g}_i$ and $\boldsymbol{f}_i^{(j)}$ captures both the low rank and the block-diagonal representation of $\boldsymbol{\Sigma}$, which results in a more efficient estimation compared with using $\boldsymbol{g}_i$ only. Another important application is PCR. In some situations, it is a cluster-specific component $\boldsymbol{f}_i^{(j)}$ that contributes most to determining the

response in PCR rather than the common component g_i. It is important to identify each cluster-specific component $f_i^{(j)}$ to explain which cluster of variables has a greater effect in predicting the response. We will discuss more about these two applications using simulated data in Section 3.7.

3.3 Relationship with existing PCA methods

Our proposed data representation (3.2) should be used to perform dimension reduction of the data with a cluster structure because it captures both the common effect (homogeneity) and the cluster-specific effect (sub-homogeneity). Consequently, this method can be viewed as an extension of many other widely used dimension-reduction methods. Three of these methods are discussed below.

- Case 1: When there is no cluster-specific effect, for example, $f_i^{(j)} = \mathbf{0}, j = 1, \ldots, J$ and $\sigma^{(j)2} \equiv \sigma^2$, representation (3.2) simply reduces to the well-known PCA:

$$x_i = \sum_{k=1}^{r_c} g_{ik}\phi_k + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.6}$$

  where g_{ik} can be found by the principal component with k^{th} largest eigenvalue and ϕ_k is the corresponding eigenvector. However, in the presence of a small cluster-specific effect, traditional PCA generally only identifies the large common effect and treats the rest as noise, disregarding the sub-homogeneity within the data.

- Case 2: When the common components do not exist (e.g., $g_i = \mathbf{0}$), representation (3.2) demonstrates a well-studied block-diagonal covariance structure of the data, as discussed in Liu et al. (2002) and Tan et al. (2015). Thus, representation (3.2) can be reduced to:

$$x_i = \sum_{j=1}^{J}\sum_{h=1}^{r_j} f_{ih}^{(j)}\gamma_h^{(j)} + \sum_{j=1}^{J} I^{(j)}\epsilon_i^{(j)}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J. \tag{3.7}$$

In this case, hierarchical clustering is usually used to group the variables with high correlations, and PCA is then performed on each group of variables to estimate the cluster-specific components $\boldsymbol{f}_i^{(j)}$. In many situations, this cluster structure is masked by a dominant common effect. Ignoring this common effect will result in a non-identifiable cluster structure.

- Case 3: Fan et al. (2013a) and Li et al. (2018) consider the following low-rank representation, in which the error covariance matrix is assumed to be cross-sectional dependent after the common components have been taken out:

$$\boldsymbol{x}_i = \sum_{k=1}^{r_c} g_{ik}\boldsymbol{\phi}_k + \boldsymbol{u}_i, \quad i = 1, \ldots, n, \tag{3.8}$$

where they assume that the covariance of $\boldsymbol{u}_i$, Σ_u, is sparse and propose a method called Principal Orthogonal complEment Thresholding (POET) to explore such a high-dimensional structure with sparsity. Li et al. (2018) used weighted PCA to find a more efficient estimator of common components $\boldsymbol{g}_i$. Hong et al. (2018) and Deville and Malinvaud (1983) also applied weighted PCA to estimate the covariance matrix when heteroscedastic noise of samples or variables exists. However, in this chapter, we consider $\boldsymbol{u}_i$ following a cluster structure with a block-diagonal covariance and aim to find its low-rank presentation, which we call sub-homogeneity. In addition, we propose iteratively estimating the common components and cluster-specific components.

3.4 Estimation methods

Correctly estimating $\boldsymbol{g}_i$ and $\boldsymbol{f}_i^{(j)}$ poses many challenges. First, PCA, which is widely used to estimate the common component $\boldsymbol{g}_i$, often performs very poorly when p is much larger than n. The group structure of the variables motivates us to separate the data according to the clusters and then extract the common information from each cluster to determine the common components. This is less

influenced by the curse of dimensionality because each cluster of the data results in a lower dimension of variable p_j. However, to accurately identify the clusters, we must remove the common effect and then perform the clustering method based on the complement $(\boldsymbol{x}_i - \sum_{k=1}^{r_c} g_{ik}\boldsymbol{\phi}_k)$, but the common components and its eigenvectors are not known in advance. This inspires us to propose a new iterative method, CPCA, to cluster the variables and estimate the components simultaneously. A flowchart that summaries our estimation method is presented in Figure 3.1. Details of the method are described in Algorithm 1.

Algorithm 3.1 (CPCA)**.**

1. Initial Step:

   (a) Perform PCA directly on the entire data $\boldsymbol{X}$ and select the number of components according to the largest drop in eigenvalues. The resulting eigenvectors $\boldsymbol{\Phi}_0$ and principal components $\boldsymbol{G}_0$ are served as the initial estimates of $\boldsymbol{\Phi}$ and $\boldsymbol{G}$. Then, find the initial complement $\boldsymbol{X}_0^c = \boldsymbol{X} - \boldsymbol{G}_0\boldsymbol{\Phi}_0^\top$

   (b) Perform hierarchical clustering for $\boldsymbol{X}_0^c$ based on a similarity matrix given by the absolute value of the empirical correlation matrix $\boldsymbol{X}_0^c$. The obtained clusters $C_0^{(j)}, j = 1, \ldots, J_0$ are served as the initial clusters.

2. Iterative Step: for $s = 1, 2, \ldots$

   (a) Cluster the variables into J_{s-1} groups according to $C_{s-1}^{(j)}$ and define variables from j^{th} cluster as $\boldsymbol{X}_s^{(j)}$. Perform PCA on each cluster of variables $\boldsymbol{X}_s^{(j)}$ and denote the obtained principal components as $\boldsymbol{\Psi}_s^{(j)}, j = 1, \ldots, J_{s-1}$. Combine these principal components as $\boldsymbol{\Psi}_s = (\boldsymbol{\Psi}_s^{(1)}, \ldots, \boldsymbol{\Psi}_s^{(J_0)})$ and perform a further step of PCA on $\boldsymbol{\Psi}_s$. Define the principal components as $\boldsymbol{G}_s$ and their corresponding eigenvectors as $\boldsymbol{\Phi}_s$. Then, compute the updated complement $\boldsymbol{X}_s^c = \boldsymbol{X} - \boldsymbol{G}_s\boldsymbol{\Phi}_s^\top$. Details for finding the eigenvectors $\boldsymbol{\Phi}_1$ can be found in Appendix 3.A.

   (b) Perform leave-one-out clustering for variables in $\boldsymbol{X}_s^c$ using PCR (more details of this clustering method are discussed in Remark 3):

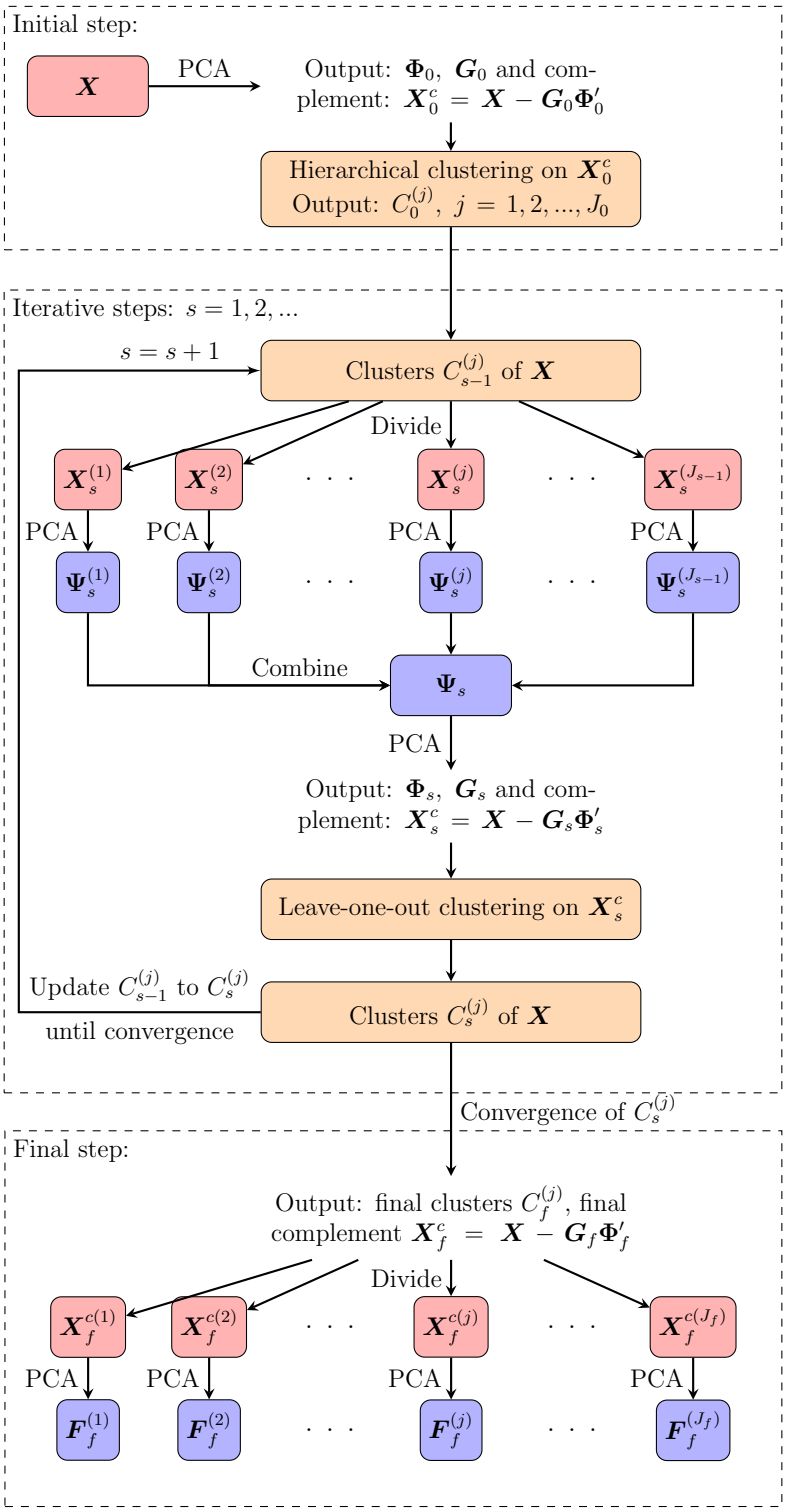


Figure 3.1: Flowchart of Algorithm 1.

(i) For $k, k = 1, \ldots, p$, leave k^{th} variable out of $\boldsymbol{X}_s^c$.

(ii) Group the rest of the variables of $\boldsymbol{X}_s^c$ based on $C_{s-1}^{(j)}$, perform PCA on each cluster again and denote the obtained components as $\boldsymbol{F}_{s-1}^{(j)}, j = 1, \ldots, J_{s-1}$.

(iii) Fit J_{s-1} PCRs by using k^{th} variable of $\boldsymbol{X}_s^c$ as the response and $\boldsymbol{F}_{s-1}^{(j)}$ as the predictor for each $j = 1, \ldots, J_{s-1}$, respectively.

(iv) Compute the sum of squared residuals (SSR) for each J_{s-1} PCR model. Assign k^{th} variable to the cluster with the minimum SSR. Update the cluster index for k^{th} variable in $C_{s-1}^{(j)}$.

(v) Repeat (i)$\sim$(iv) for each k and denote the updated clusters as $C_s^{(j)}$.

(c) Repeat (a) and (b) within this step until the clusters converge, and define this final converged cluster as $C_f^{(j)}$.

3. Final Step:

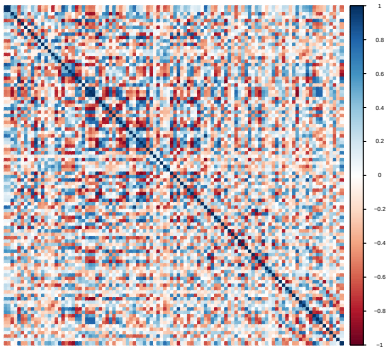
(a) Repeat (a) in the Iterative Step, but using cluster $C_f^{(j)}$. The final complement is denoted by $\boldsymbol{X}_f^c = \boldsymbol{X} - \boldsymbol{G}_f \boldsymbol{\Phi}_f^\top$.

(b) Cluster variables of $\boldsymbol{X}_f^c$ based on $C_f^{(j)}$, perform PCA on each cluster $\boldsymbol{X}_f^{c(j)}$ and denote the obtained cluster-specific components as $\boldsymbol{F}_f^{(j)}, j = 1, \ldots, J_f$.

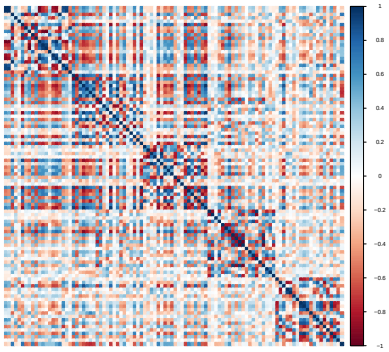
Therefore, CPCA produces these required outputs: the final clusters of p variables $C_f^{(j)}, j = 1, \ldots, J_f$, the final estimate of the common components $\boldsymbol{G}_f$ and the cluster-specific components $\boldsymbol{F}_f^{(j)}, j = 1, \ldots, J_f$, along with their eigenvectors. Some details and discussions of CPCA are provided in the following remarks.

Remark 3.1. [Iterative Step]:
One of the key contributions of our algorithm is iteratively estimating the common components and cluster-specific components. Directly using PCA on the entire data $\boldsymbol{X}$ generally leads to a very poor estimate of the common components because a large p may blur the spike structure of the sample covariance matrix. Motivated by the group structure of the data, we utilise Iterative Step (a) to first cluster variables and then perform PCA on the level of clusters, followed by a

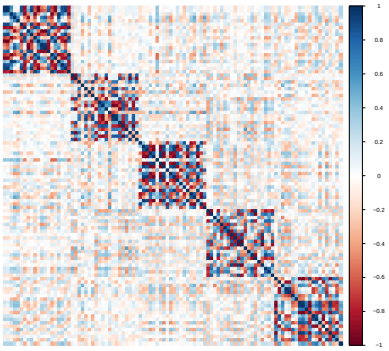
second layer of PCA that extracts common information shared by each group. Iterative Step (b) is then implemented to update and improve the clusters information. This is more effective in estimating the common components because PCA is performed in a smaller p case, which is also numerically shown in Figure 3.2. The data $\boldsymbol{X}$ used here are generated by simulation Example 2. Figure 3.2a presents the correlation plot for the original data $\boldsymbol{X}$. We observe that the cluster structure is masked by the common effect. Figure 3.2b demonstrates that after the common effects estimated in the Initial Step are removed, it is still difficult to perceive the cluster structure. However, if we remove the common effects estimated in the Final Step, the block-diagonal structure is clear, implying a prominent heterogeneity within the data. Therefore, iteratively estimating the common components is advantageous.



(a) Correlation for the original data $\boldsymbol{X}$



(b) Correlation for $\boldsymbol{X}_0^c = \boldsymbol{X} - \boldsymbol{G}_0\boldsymbol{\Phi}_0^\top$



(c) Correlation for $\boldsymbol{X}_f^c = \boldsymbol{X} - \boldsymbol{G}_f\boldsymbol{\Phi}_f^\top$

Figure 3.2: Correlation plot using data generated by simulation example 2

Remark 3.2. [Initial Step (a) and Iterative Step (a)]:

In Section 3.2, we have assumed that the common components have a larger

variance than the cluster-specific components. However, in practice, when the number of common components is greater than 1 (e.g., $r_c > 1$), the discrepancy of the eigenvalues (from the sample covariance) among the common components can be accidentally greater than the one between the common components and the cluster-specific component, especially when some groups have a relatively larger cluster-specific effect (e.g., variables in one group are on a large scale or highly correlated). In this situation, selecting the number of components using the largest drop in eigenvalues in Initial Step (a) and Iterative Step (a) may lead to underestimating the number of common components. One possible way to alleviate this problem is to remove the scale effect by scaling $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ before applying PCA in these two steps. After the number of common components is determined, we multiply the standard deviation back while calculating $\boldsymbol{G}\boldsymbol{\Phi}^\top$. This standardisation, along with the Iterative Step, can avoid underestimating the number of common components, which will be demonstrated in the simulation studies.

Remark 3.3. [Initial Step (b), Iterative Step (b) and (c)]:
Initial Step (b), an ordinary hierarchical clustering method using average-linkage as dissimilarity between clusters is performed. As discussed in Bühlmann et al. (2013), we choose the number of clusters using the largest increment in height between iterations proceeded in a agglomerative way, such that $J_0 = \text{argmax}_j(h_{j+1} - h_j)$.

In Iterative Step (b), we propose a new clustering method called "leave-one-out PCR clustering" (LOO-PCR clustering), which is another key contribution of the CPCA. The underlying idea of this new clustering method is that if one variable belongs to a cluster, it should be well-predicted by the principal components extracted from that cluster. This is more aligned with our model set-up than hierarchical clustering based on correlations. A similar idea of leaving one out is used for functional clustering, as discussed in (Chiou and Li, 2007). In addition, hierarchical clustering can perform poorly when the dimension of variables p is larger than n. In step (iv), we assign k^{th} variable to the cluster that achieves the minimum SSR. However, when the minimum SSR is larger than a

threshold τ, we treat the k^{th} variable itself as a cluster, because in such a situation, this variable cannot be well-predicted by any clusters of variables. Hence, the number of clusters is driven by the data and can vary in each iteration. This suggests that our proposed method is more flexible than hierarchical clustering. We set $\tau = 0.95$, and Example 4 in the simulation studies demonstrates that our proposed LOO-PCR clustering, served as a clustering method itself, outperforms hierarchical clustering in a large p small n case. Identifiability of LOO-PCR clustering is also demonstrated theoretically in Section 3.5.

In Iterative Step (c), we stop the iteration when the clusters $C_s^{(j)}$ converge. In our algorithm, we adopt the adjusted rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985) between the clusters $C_s^{(j)}$ and the one in the previous iteration $C_{s-1}^{(j)}$ as the stopping criterion. When the ARI is above a certain threshold η, we stop the iteration. Details of the ARI are discussed in Section 3.6. In this study, we use $\eta = 0.97$.

Remark 3.4. [Final Step (b)]:
In Final Step (b), after the variables are clustered based on $C_f^{(j)}$, we perform PCA on each cluster $\boldsymbol{X}_f^{c(j)}, j = 1, \ldots, J_f$ (complement) separately. To prevent losing too much information, the principal components selected in each cluster should explain a certain percentage α of the total variation within that cluster (Jolliffe, 2002). Therefore, we select the number of cluster-specific principal components r_j at which the largest drop in eigenvalues occurs, given that those principal components explain at least α percent of the total variation in the corresponding cluster, such that

$$\widehat{r}_j = \underset{r_j \in R_j}{\operatorname{argmax}} \{\widehat{\lambda}_{r_j}^{(j)} - \widehat{\lambda}_{r_j-1}^{(j)}\},$$
$$\text{where } R_j = \left\{ r_j : \sum_{h=1}^{r_j} \widehat{\lambda}_h^{(j)} \Big/ \sum_{h=1}^{\min(n,p)} \widehat{\lambda}_h^{(j)} \mathbb{I}\{\widehat{\lambda}_h^{(j)} > 0\} > \alpha \right\} \tag{3.9}$$

and $\widehat{\lambda}_h^{(j)}$ defines h^{th} largest eigenvalue of the sample covariance of $\boldsymbol{X}_f^{c(j)}$. In this study, $\alpha = 0.8$.

3.5 Identifiability of LOO-PCR clustering approach

Achieving an accurate sub-homogeneity pursuit relies on the effectiveness of the clustering procedure. In this section, we investigate the proposed LOO-PCR clustering method theoretically. More specifically, we examine the identifiability of cluster membership in our proposed method.

Consider one random variable X_{mi} that belongs to cluster l. Based on the variable's structure in cluster l, there exists $m \in \{1, 2, \ldots, p\}$, such that:

$$X_{mi} = \sum_{k=1}^{r_c} g_{ik}\phi_{mk} + u_{mi}^{(l)}, \quad u_{mi}^{(l)} = \sum_{h=1}^{r_l} f_{ih}^{(l)}\gamma_{mh}^{(l)} + \epsilon_{mi}^{(l)}. \tag{3.10}$$

We now consider another cluster d with components $g_{ik}, k = 1, \ldots, r_c$ and $f_{ih}^{(d)}, h = 1, \ldots, r_d$. When applying our proposed LOO-PCR clustering method, we regress $u_{mi}^{(l)}$ on $f_{ih}^{(d)}, h = 1, 2, \ldots, r_d$ and measure its goodness of fit to determine whether the random variable X_{mi} belongs to cluster d, such that:

$$u_{mi}^{(l)} = \sum_{h=1}^{r_d} \beta_h f_{ih}^{(d)} + \zeta_{mi}^{(d)}, \tag{3.11}$$

where $\beta_1, \ldots, \beta_{r_d}$ are coefficients and $\zeta_{mi}^{(d)}$ is the error. Naturally, we expect that features (e.g., principal components) from cluster d cannot explain $u_{mi}^{(l)}$ sufficiently. Based on (3.10) and (3.11), intuitively, a large discrepancy between principal components $f_{ih}^{(d)}$ from cluster d and principal components $f_{ih}^{(l)}$ from cluster l will result in a large residual $\zeta_{mi}^{(d)}$ in (3.11).

The following theorem will show the property of $\widehat{\zeta}_{mi}^{(d)}$, which is an estimator of the error $\zeta_{mi}^{(d)}$ from our proposed clustering approach.

Theorem 3.1. *For any cluster d and any random variable X_{mi} from another cluster l, we have the following evaluation:*

$$\|\widehat{\boldsymbol{\zeta}}_m^{(d)}\|_2 = \|\boldsymbol{M}_{\boldsymbol{F}^{(d)}}\boldsymbol{M}_{\boldsymbol{G}}\boldsymbol{x}_m\|_2 + O_p\left(\max(\alpha_{np_d}, \gamma_{nJ})\right)\|\boldsymbol{x}_m\|_2, \tag{3.12}$$

where $\widehat{\boldsymbol{\zeta}}_m^{(d)} = \left(\widehat{\zeta}_{m1}^{(d)}, \widehat{\zeta}_{m2}^{(d)}, \ldots, \widehat{\zeta}_{mn}^{(d)}\right)^\top$, $\|\boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} - \boldsymbol{M}_{\boldsymbol{F}^{(d)}}\|_2 := O_p(\alpha_{np_d})$, $\|\boldsymbol{M}_{\widehat{\boldsymbol{G}}} - \boldsymbol{M}_{\boldsymbol{G}}\|_2 :=$

$O_p(\gamma_{nJ})$, $\boldsymbol{x}_m = (X_{m1}, X_{m2}, \ldots, X_{mn})^\top$, $\boldsymbol{M}_{\boldsymbol{F}^{(d)}} = \boldsymbol{I}_n - \boldsymbol{F}^{(d)} \left(\boldsymbol{F}^{(d)\top} \boldsymbol{F}^{(d)} \right)^{-1} \boldsymbol{F}^{(d)\top}$, *and* $\boldsymbol{M}_{\boldsymbol{G}} = \boldsymbol{I}_n - \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{G})^{-1} \boldsymbol{G}^\top$; *here,* $\boldsymbol{F}^{(d)}$ *is an* $n \times r_d$ *matrix with* (i,h)*-th element being* $f_{ih}^{(d)}$, *and* $\boldsymbol{G}$ *is an* $n \times r_c$ *matrix with* (i,k)*-th element being* g_{ik}. *Here,* $\boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}}$ *and* $\boldsymbol{M}_{\widehat{G}}$ *are* $\boldsymbol{M}_{\boldsymbol{F}^{(d)}}$ *and* $\boldsymbol{M}_{\boldsymbol{G}}$, *but with* $\boldsymbol{F}^{(d)}$ *and* $\boldsymbol{G}$ *replaced by* $\widehat{\boldsymbol{F}}^{(d)}$ *and* $\widehat{\boldsymbol{G}}$, *respectively.*

A brief proof of Theorem 3.1 is provided in Appendix 3.B.

Remark 3.5. It is expected that $\widehat{\boldsymbol{\zeta}}_m^{(d)}$ will be large when cluster d and cluster l are different. When $l = d$, the first term on the right-hand side of (3.12) is equal to $\|\boldsymbol{M}_{\boldsymbol{F}^{(d)}} \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{\zeta}_m^{(d)}\|_2$. In contrast, when $l \neq d$ (e.g., $\|\boldsymbol{F}^{(l)}\|_2$ and $\|\boldsymbol{F}^{(d)}\|_2$ are distinct), the first term on the right-hand side of (3.12) is equal to $\|\boldsymbol{M}_{\boldsymbol{F}^{(d)}} \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{F}^{(l)} \boldsymbol{\gamma}^{(l)} + \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{\zeta}_m^{(l)}\|_2$, which is dominated by $\|\boldsymbol{F}^{(l)}\|_2$. Meanwhile, the second term on the right-hand side of (3.12) is determined by the estimation accuracy of $\boldsymbol{F}^{(d)}$ and $\boldsymbol{G}$. It is related to the dimension of cluster d (e.g., p_d), the sample size n and the total number of clusters J. Bai and Ng (2002a) provided the rate of convergence for the projection matrices of principal components. Fan et al. (2013a) and Fan et al. (2018) also studied the properties of principal components for high-dimensional data. In this study, we do not pursue the exact expression of α_{np_d} and γ_{nJ}. However, given that the homogeneity $\boldsymbol{M}_{\boldsymbol{G}}$ and sub-homogeneity $\boldsymbol{F}^{(d)}$ can be estimated accurately, the first term on the right-hand side of (3.12) is expected to dominate $\|\widehat{\boldsymbol{\zeta}}_m^{(d)}\|_2$. This implies that two different clusters d and l are identifiable, given that $\|\boldsymbol{F}^{(l)}\|_2$ has a higher order than $\|\boldsymbol{\zeta}_m^{(d)}\|_2$. As a general discussion, Theorem 3.1 shows that when the sub-homogeneity from the different clusters is distinct, our clustering procedure can effectively separate the variables from the different clusters.

3.6 Simulation studies

In this section, we conduct various simulation studies to investigate the performance of our proposed CPCA method compared with traditional PCA under different simulation settings.

3.6.1 Simulation settings

We generate the data from representation (3.2):

$$\boldsymbol{x}_i = \sum_{k=1}^{r_c} g_{ik}\boldsymbol{\phi}_k + \sum_{j=1}^{J}\sum_{h=1}^{r_j} f_{ih}^{(j)}\boldsymbol{\gamma}_h^{(j)} + \sum_{j=1}^{J} \boldsymbol{I}^{(j)}\boldsymbol{\epsilon}_i^{(j)}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,J.$$

First, we generate $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{r_c})$ as a $p \times r_c$ orthonormal matrix. Second, $p_j \times r_j$ orthonormal matrix $\boldsymbol{\Psi}^{(j)} = (\boldsymbol{\psi}_1^{(j)},\ldots,\boldsymbol{\psi}_{r_j}^{(j)})$ is generated randomly and independently across different clusters for $j = 1,\ldots,J$ and let $\boldsymbol{\Gamma}_h^{(j)} = (\mathbf{0}^{(1)},\ldots,\boldsymbol{\Psi}_h^{(j)},\ldots,\mathbf{0}^{(J)})$. We then generate $g_{ik}, i = 1,\ldots,n, k = 1,\ldots,r_c$ from $\sqrt{\delta_k}W_{ik}$, where W_{ik} are i.i.d. standard normal random variables and the eigenvalues δ_h are defined as $\delta_1 > \ldots > \delta_{r_c} > 0$. Similarly, we generate $f_{ih}^{(j)}, i = 1,\ldots,n, h = 1,\ldots,r_j, j = 1,\ldots,J$ from $\sqrt{\lambda_h^{(j)}}Z_{ih}^{(j)}$, where $Z_{hi}^{(j)}$ are again i.i.d. standard normal random variables and the eigenvalues $\lambda_h^{(j)}$ are defined as $\lambda_1^{(j)} > \ldots > \lambda_{r_j}^{(j)} > 0, j = 1,\ldots,J$. Lastly, $\boldsymbol{\epsilon}_i^{(j)}$ is a $p_j \times 1$ vector consisting of p_j i.i.d. normal random variables with mean 0 and variance ${\sigma^{(j)}}^2$. Data from each cluster are generated according to the above setting and then combined to obtain $\boldsymbol{X}$.

In our simulation study, we consider four different settings, which are outlined below.

- Example 1 ($n = 50, p = 100$). In this example, we consider the number of common components $r_c = 3$ and generate δ_k, $k = 1,\ldots,r_c$, from $\mathcal{N}(75,5)$. We further simulate W_{ik}, $k = 1,2,\ldots,r_c$, $i = 1,2,\ldots,2n$, from a standard normal distribution independently so that $\boldsymbol{g}_i$ can be constructed.

  In terms of the cluster-specific components, we set $J = 5$, and for each of the five clusters, we consider $r_j \equiv 2$ and generate $\lambda_h^{(j)} \sim \mathcal{N}(5,1)$ for $h = 1,2, j = 1,2,3$, but $\lambda_h^{(j)} \sim \mathcal{N}(25,1)$ for $h = 1,2, j = 4,5$. That is, each of the five clusters has the same number of cluster-specific components, while the components from the last two clusters have higher variance than those from the first three clusters. Then, setting $p_j \equiv 20$ for each cluster results in a total number of variables $p = 100$. According to (3.2), $Z_{ih}^{(j)}$ for $h = 1,2$

and $\boldsymbol{\epsilon}_i^{(j)}$ are simulated independently for each cluster j, with $\sigma^{(j)} = 0.1$ for $j = 1, 2, 3$ and $\sigma^{(j)} = 0.5$ for $j = 4, 5$ so that $2n$ observations are constructed based on (3.2). Then, the first n observations are served as the training sample $\boldsymbol{X}_{train} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and the last n observations are designated as the testing sample $\boldsymbol{X}_{test} = (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{2n})$. In this example, $n = 50$.

- Example 2 ($n = 30, p = 100$). This example is identical to Example 1, except that the sample size is decreased from 50 to 30. We will show that our CPCA is more reliable than other competing methods when the sample size is relatively small.

- Example 3 ($n = 50, p = 200$). In this example, we consider data that consist of more clusters, that is, we consider $J = 10$ clusters compared with five in Example 1, leading to a number of variables $p = 200$ given $p_j \equiv 20$. We maintain $r_j \equiv 2$ but generate $\lambda_h^{(j)} \sim \mathcal{N}(5, 1)$ for $j = 1, \ldots, 6$ and $\lambda_h^{(j)} \sim \mathcal{N}(25, 1)$ for $j = 7, \ldots, 10$. Alternatively, the first six clusters have smaller cluster-specific effects than the last four clusters. Accordingly, $\boldsymbol{\epsilon}_i^{(j)}$ is simulated independently for each of the 10 clusters, with $\sigma^{(j)} = 0.1$ for $j = 1, \ldots, 6$ and $\sigma^{(j)} = 0.5$ for $j = 7, \ldots, 10$. The settings for the common components remain the same as those in Example 1.

- Example 4 ($n = 30, p = 100$). This example is identical to Example 2, but without the common effect, which is a special case of our proposed data representation as in (3.7). Correspondingly, the response is generated as:

$$y_i = \sum_{j=1}^{J} \boldsymbol{f}_i^{(j)\top} \boldsymbol{\beta}_j + e_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J. \tag{3.13}$$

  In this example, we do not estimate the common effect, and the clustering method is directly applied to the data. Therefore, the purpose of considering this example is to demonstrate that our proposed LOO-PCR clustering (Iterative Step (b) and (c) in CPCA), served as a clustering method itself, can outperform the hierarchical clustering.

For all four examples, we repeat the simulation procedure 100 times and investigate the clustering and recovering accuracy of the following methods:

- PCA: The classical PCA method is performed directly on the whole data set without clustering variables. In addition, clustering accuracy is not considered for this method. The number of principal components is determined according to (3.9) as discussed in Remark 3.4.

- CPCA_I_ns: This procedure is simply the CPCA, but without the Iterative Step. First, the complement $\boldsymbol{X}_0^c = \boldsymbol{X} - \boldsymbol{G}_0\boldsymbol{\Phi}_0^\top$ is found, and then clusters $C_0^{(j)}$ are obtained using hierarchical clustering for $\boldsymbol{X}_0^c$. Variables of $\boldsymbol{X}_0^c$ are clustered, and the cluster-specific components $\boldsymbol{F}_0^{(j)}$ are finally obtained using PCA on each cluster. In this method, $\boldsymbol{X}$ is not scaled. Note that this method is not considered in Example 4.

- CPCA_I: This is identical to CPCA_I_ns, but $\boldsymbol{X}$ is scaled when we find the number of common components.

- CPCA_F_ns: This is exactly the estimation procedure described in CPCA. Different to CPCA_I_ns, the clusters and components are estimated iteratively. In this method, $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ are not scaled. Note that this method is not considered in Example 4.

- CPCA_F: This is identical to CPCA_F_ns, but $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ are scaled when we find the number of common components, as discussed in Remark 3.2.

The main purposes of this comparison are to demonstrate that: 1) traditional PCA fails to capture the cluster-specific components; 2) our proposed iterative estimation of the clusters and components significantly outperforms the initial estimation; 3) scaling $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ can improve the performance of CPCA; 4) our proposed LOO-PCR clustering outperforms the traditional hierarchical clustering method.

In this study, clustering accuracy is measured by the ARI (Rand, 1971; Hubert and Arabie, 1985). The ARI is a corrected version of the rand index that measures

the similarity between two clusterings of the same data using the proportions of agreements between these two partitions. The correction is achieved by subtracting the rand index by its expected value. The ARI can yield a maximum value of 1, and a high value implies high similarity. In this simulation study, we use the function '*adj.rand.index*' in R package '*pdfCluster*' to compute the ARI between the partitions of the variables produced by these methods and the true partitions as in the data generation process.

The recovering accuracy is measured by the mean squared recovering error (MSRE). In this study, the MSRE is defined as:

$$MSRE = \frac{1}{np} \left\| \widehat{\boldsymbol{X}}_{test} - \boldsymbol{X}_{test} \right\|_F^2, \tag{3.14}$$

where $\widehat{\boldsymbol{X}}_{test}$ is the recovered testing sample computed using the testing sample and the eigenvectors $(\boldsymbol{\Phi}, \boldsymbol{\Lambda}^{(j)})$ estimated from the training sample $\boldsymbol{X}_{train}$. We prefer a lower MSRE because it indicates a better recovering of the data.

3.6.2 Simulation results

The boxplots of the ARI, total number of principal components selected (No.PCs), and MSRE for Example 1 are presented in the first row of Figure 3.3. As shown, CPCA_F and CPCA_F_ns (blue boxes) achieve a consistently higher ARI than their non-iterative counterparts (red boxes), which implies the advantages of using the iterative estimation. In the initial clustering, the common effects are estimated without partitioning the variables, while the iterative clustering partitions the variables in the previous step to estimate the common components, resulting in a more accurate estimation. This is also stated in Remark 3.1. Further, the first column of Table 3.1 shows that CPCA_F yields the most desirable average ARI, which is 15% higher than that of CPCA_I. Compared with CPCA_F_ns, CPCA_F achieves a higher average ARI but lower standard errors, which numerically confirms that estimating the common components using scaled $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ leads to more accurate and stable results.

From the recovering accuracy perspective, it is worth noting from the first

row of Figure 3.3 and the first column of Table 3.1 that PCA always performs poorly under this cluster structure of the data, while all CPCA-based methods demonstrate superior performance in recovering because they achieve lower MSRE than PCA. Of these methods, CPCA_F generally obtains the lowest MSRE. This is to be expected because CPCA_F results in more accurate partitions and produces more reliable estimations of the cluster-specific components, especially those with small variations, which contributes the most to predicting the response according to our simulation settings. This further confirms the advantages of using our proposed method. The outperformance of CPCA_F can also be explained by the number of principal components selected in each method. As shown in Figure 3.3 1(b), classical PCA always underestimates the true number of components (the dashed line), while CPCA_I and CPCA_I_ns tend to overestimate it. Of these methods, CPCA_F estimates the number of components most accurately and stably, thus obtaining the lowest MSRE with the smallest standard errors.

The second row of Figure 3.3 and the second column of Table 3.1 summarize the simulation results for Example 2. As shown, when the sample size is small ($n = 30$), the ARI for all methods declines as expected, while CPCA_F still achieves a satisfactory ARI of 82%. In terms of recovery accuracy, it is evident that CPCA_F outperforms all other methods, because the discrepancies in MSRE between CPCA_F and other methods are more significant compared with those in Example 1. This indicates that the iterative estimation is less sensitive to a small n large p situation.

We now investigate the simulation results for Example 3, which are displayed in the third row of Figure 3.3 and the third column of Table 3.1. As demonstrated in the last panel of Table 3.1, in terms of the recovering, we observe similar findings as in Example 2. CPCA_I and CPCA_I_ns are very unstable in recovering the test data; however, CPCA_F still achieves a satisfactory and stable MSRE regardless of whether n decreases or p increases.

Lastly, we investigate Example 4, which does not take into account the common effect. In this example, CPCA_I_ns and CPCA_F_ns are not considered,

and CPCA_I and CPCA_F simply reduce to performing PCA on each cluster of variables based on hierarchical clustering and LOO-PCR clustering, respectively. From the last row of Figure 3.3 and the last column of Table 3.1, we observe that applying PCA after clustering the variables demonstrates better performance than applying PCA directly to the data. Compared with CPCA_I, CPCA_F achieves a higher average ARI and lower average MSRE with smaller standard deviations, implying that our proposed LOO-PCR clustering outperforms hierarchical clustering when p is large but n is small.

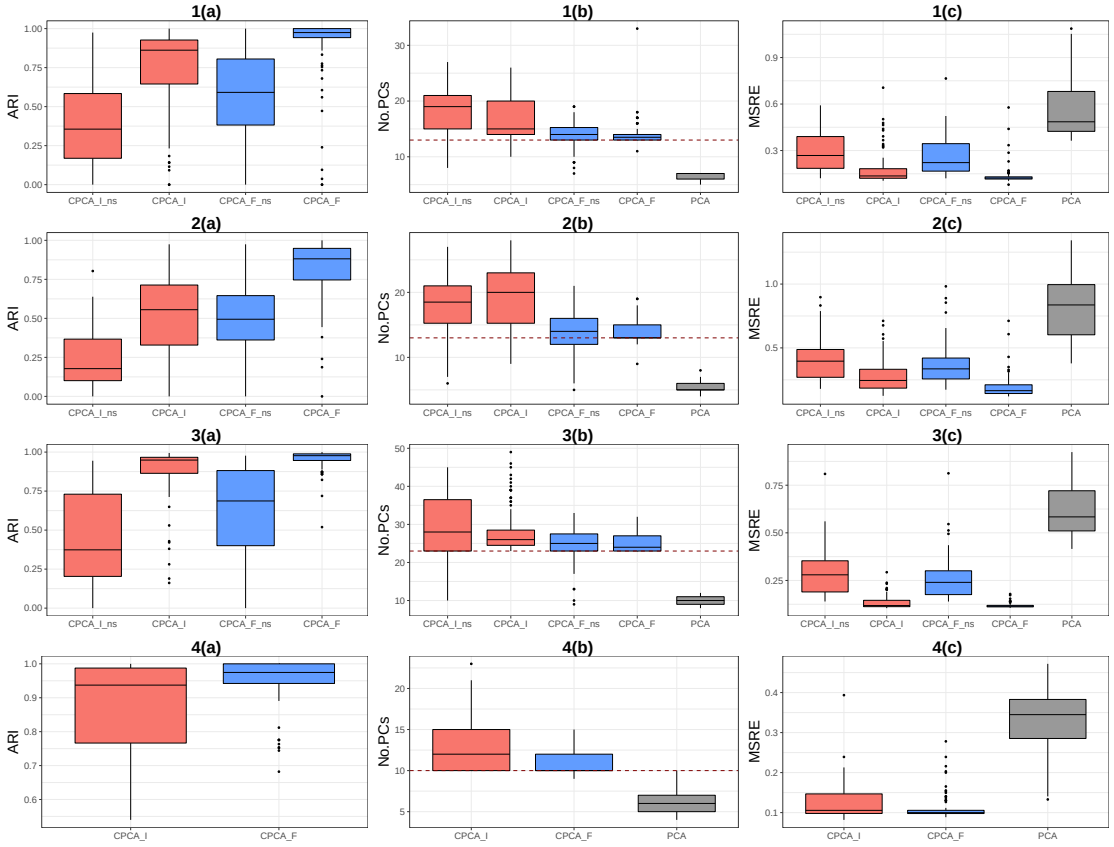


Figure 3.3: Boxplots of the following three measurements based on 100 simulations from Example 1 to 4: (a) ARI, (b) No.PCs, and (c) MSRE.

3.7 Applications of CPCA

Table 3.1: Averages (standard errors) of ARI, total number of principal components selected (No.PCs), and MSPE for Example 1, 2, 3, and 4.

	Method	Example 1	Example 2	Example 3	Example 4
ARI	CPCA_I_ns	0.39	0.23	0.45	-
		(0.26)	(0.18)	(0.29)	-
	CPCA_I	0.75	0.52	0.88	0.88
		(0.26)	(0.24)	(0.17)	(0.12)
	CPCA_F_ns	0.57	0.48	0.63	-
		(0.27)	(0.23)	(0.27)	-
	CPCA_F	**0.90**	**0.82**	**0.95**	**0.95**
		(0.21)	(0.19)	(0.06)	(0.07)
No.PCs	CPCA_I_ns	17.90	17.54	29.08	-
		(3.54)	(4.76)	(8.41)	-
	CPCA_I	16.66	19.31	28.61	12.83
		(3.64)	(4.36)	(6.14)	(2.96)
	CPCA_F_ns	13.77	13.63	25.04	-
		(2.28)	(3.19)	(4.34)	-
	CPCA_F	**14.08**	**14.11**	**25.12**	**10.99**
		(2.28)	(1.73)	(2.25)	(1.31)
	PCA	6.43	5.61	9.91	6.10
		(0.77)	(1.00)	(1.09)	(1.45)
MSRE	CPCA_I_ns	0.29	0.41	0.29	-
		(0.12)	(0.16)	(0.12)	-
	CPCA_I	0.18	0.28	0.13	0.13
		(0.10)	(0.13)	(0.03)	(0.04)
	CPCA_F_ns	0.26	0.37	0.26	-
		(0.12)	(0.16)	(0.10)	-
	CPCA_F	**0.14**	**0.19**	**0.12**	**0.11**
		(0.06)	(0.09)	(0.01)	(0.03)
	PCA	0.58	0.82	0.62	0.33
		(0.20)	(0.25)	(0.13)	(0.08)

3.7.1 Principal component regression

As aforementioned, one of the most important uses of the principal components is PCR. We consider a PCR model with an univariate response y_i for i^{th} observation as:

$$y_i = \boldsymbol{g}_i^\top \boldsymbol{\alpha} + \sum_{j=1}^{J} \boldsymbol{f}^{(j)\top}{}_i \boldsymbol{\beta}_j + e_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J. \tag{3.15}$$

where $\boldsymbol{\alpha}$ is a $r_c \times 1$ vector representing the regression coefficients of the common components $\boldsymbol{g}_i$, $\boldsymbol{\beta}_j$ is a $r_j \times 1$ vector denoting the regression coefficients of j^{th} cluster-specific components $\boldsymbol{f}_i^{(j)}$ and e_i is simply the error term with mean 0 and variance θ^2.

Therefore, after the the common components $\boldsymbol{g}_i$ and cluster-specific components $\boldsymbol{f}_i^{(j)}$ are estimated via CPCA, we can fit a PCR to predict the response. However, in many situations, only a few groups are useful in predicting the response. To investigate which clusters of variable have an impact on predicting the response, we utilise the group lasso (Yuan and Lin, 2006) using the components produced by CPCA as the coviarates to estimate the regression coefficients as in (3.15),

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{g}_i^\top \boldsymbol{\alpha} - \sum_{j=1}^{J} \boldsymbol{f}^{(j)\top}{}_i \boldsymbol{\beta}_j \right)^2 + \lambda \left(\|\boldsymbol{\alpha}\|_2 + \sum_{j=1}^{J} \|\boldsymbol{\beta}^{(j)}\|_2 \right), \tag{3.16}$$

where λ is the tuning parameter that controls the sparsity of the regression coefficients. Using this group lasso penalty, $\boldsymbol{\beta}_j$ for some j will be shrunk to zero exactly so that we can identify which clusters of variables are important in predicting the response.

To better demonstrate the performance of our method, we conduct similar simulation studies as in Section 3.6. We consider the following three examples:

- Example 1 ($n = 50, p = 100$). We generate $\boldsymbol{X}$ using the same settings as Example 1 in Section 3.6.1. Then, we simulate the response y_i according

to (3.15) by setting regression coefficients $\boldsymbol{\alpha} = (1, 1, 1)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5) = (10, 10, 0, \ldots, 0)$ and standard deviation $\theta = 1$. That is, only the common effect and the cluster-specific effects for the first cluster are important in predicting the response. This regression setting is interesting because the cluster-specific components with smaller variance (e.g. 1^{st} cluster) are highly likely to be omitted or estimated poorly in traditional PCA, but they can sometimes be very important in predicting the response.

- Example 2 ($n = 30, p = 100$). This example is identical to Example 1 above, except that the sample size is decreased from 50 to 30.

- Example 3 ($n = 50, p = 200$). We generate $\boldsymbol{X}$ using the same settings as Example 3 in Section 3.6.1. That is, we consider $J = 10$ clusters in this example. In regard to the PCR, we set regression coefficients $\boldsymbol{\alpha} = (1, 1, 1)$ again and consider $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{10}) = (10, 10, 0, \ldots, 0)$ such that only the common effect and the cluster-specific effects for the first cluster are useful and the rest of clusters are noise.

For all three examples, we repeat the simulation procedure 100 times and investigate prediction accuracy of the methods mentioned in Section 3.6.1. The MSPE is used to determine whether the selected components can accurately predict the response:

$$MSPE = \frac{1}{n} \sum_{i=1}^{n} (\widehat{y}_{i,test} - y_{i,test})^2 , \tag{3.17}$$

where $\widehat{y}_{i,test}$ is computed using (3.15) with the regression coefficients $(\gamma, \boldsymbol{\beta})$ and the eigenvectors $(\boldsymbol{\Phi}, \boldsymbol{\Lambda}^{(j)})$ estimated from the training sample.

The first panel of Table 3.2 and the first row of Figure 3.4 display mean/standard errors and the boxplot of MSPE for all three examples. Note that in this section, CPCA_F and CPCA_F_g represent the ordinary least squares (OLS) and the group lasso regressed on the principal components produced by CPCA_F, respectively. Similar applies to the rest of CPCA-based methods. When the traditional PCR (OLS) is utilised, we clearly see that CPCA_F achieves much lower average MSPE

Table 3.2: Averages (standard errors) of MSPE and $\|\widehat{\Sigma} - \Sigma\|_F^2$ for Example 1, 2, and 3.

	Method	Example 1	Example 2	Example 3
MSPE	CPCA_I_ns	2.06	4.51	4.02
		(1.09)	(3.60)	(2.65)
	CPCA_I	1.86	5.48	3.43
		(0.87)	(4.61)	(2.91)
	CPCA_F_ns	2.68	4.00	2.98
		(2.11)	(2.38)	(1.71)
	CPCA_F	**1.64**	**2.56**	**2.31**
		(0.59)	(1.44)	(0.64)
	PCA	6.76	6.98	7.22
		(1.55)	(1.96)	(2.14)
	CPCA_I_ns_g	1.80	3.24	2.52
		(1.08)	(1.75)	(1.50)
	CPCA_I_g	1.56	2.37	1.79
		(0.77)	(1.11)	(0.53)
	CPCA_F_ns_g	2.62	3.70	2.43
		(2.07)	(2.21)	(1.38)
	CPCA_F_g	**1.50**	**2.18**	**1.77**
		(0.53)	(0.97)	(0.46)
$\|\widehat{\Sigma} - \Sigma\|_F^2$	CPCA_I_ns	63.09	79.08	85.49
		(9.70)	(12.99)	(9.84)
	CPCA_I	56.56	74.73	73.62
		(9.03)	(13.85)	(8.72)
	CPCA_F_ns	59.83	77.44	84.50
		(8.62)	(13.34)	(9.48)
	CPCA_F	**54.87**	**72.37**	**72.48**
		(7.73)	(13.52)	(7.94)
	PCA	58.33	77.26	78.00
		(7.71)	(13.63)	(7.69)
	POET	57.18	74.86	76.41
		(7.27)	(12.49)	(7.28)

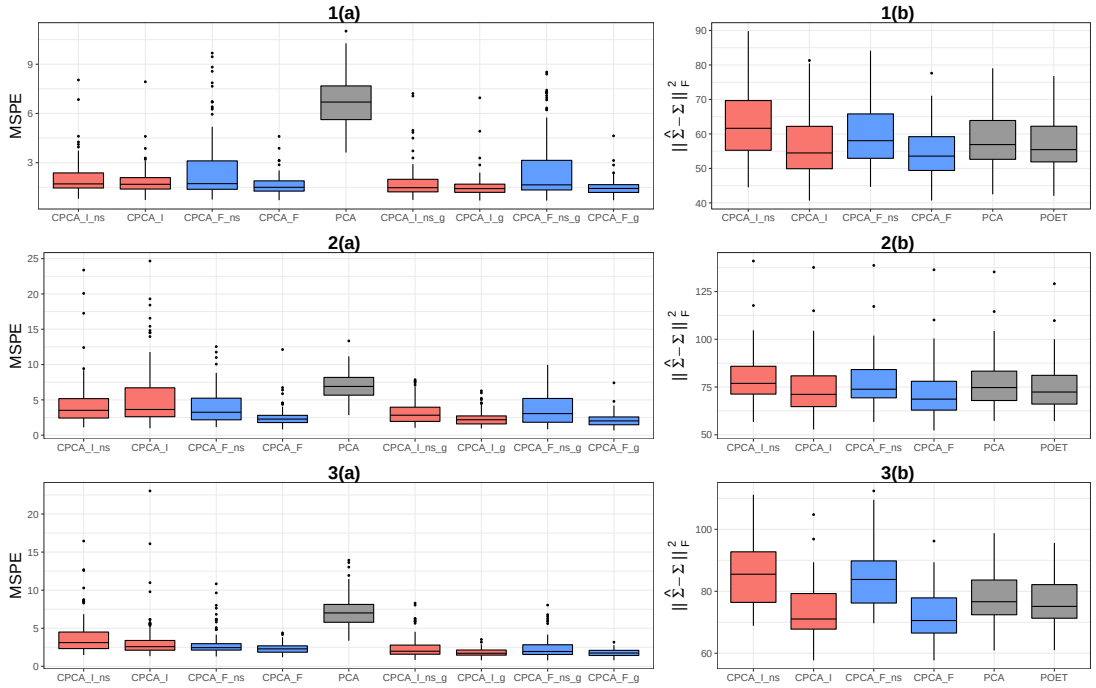


Figure 3.4: Boxplots of the following two measurements based on 100 simulations from Example 1 to 3: (a) MSPE and (b) $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F^2$.

and standard deviation than other methods, especially when n is small or p is large. Not surprisingly, the traditional PCA performs the worst because it fails to capture any of the sub-homogeneity, which is important in predicting the response under our setting. Moreover, we observe that PCR using the group lasso significantly outperforms those OLS counterparts, especially for CPCA_I_ns and CPCA_I when n is small or p is large. The group lasso can provide substantial reductions in MSPE for these methods because they tend to select a large number of principal components, only some of which are useful. Overall, CPCA_F_g achieves the lowest mean and standard error of MSPE among all the methods in all three examples, because it not only accurately extracts both common components and group-specific components, but also shrinks the coefficients of those unimportant groups down to zero.

3.7.2 Covariance estimation

Another useful application of CPCA is to estimate $\mathbf{\Sigma}$ as in (3.4). Using $\boldsymbol{g}_i$ and $\boldsymbol{f}_i^{(j)}$ leads to a better estimation of $\mathbf{\Sigma}$ compared with using $\boldsymbol{g}_i$ only, because $\boldsymbol{g}_i$

and $\boldsymbol{f}_i^{(j)}$ can capture both the low rank and the block-diagonal representation of $\boldsymbol{\Sigma}$. Therefore, we estimate $\boldsymbol{\Sigma}$ using clusters, the common components, and the cluster specific components, along with their associated eigenvectors estimated via CPCA. In this section, we numerically illustrate the advantage of using CPCA to estimate the covariance matrix in comparison with other traditional PCA methods. Again, we conduct three simulation examples mentioned in Section 3.7.1 and utilise the Euclidean distance (ED) between the estimated covariance and population covariance, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2$, to measure the performances of different methods. Recall $\boldsymbol{\Sigma}$ is generated from (3.4):

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}\text{cov}\left(\boldsymbol{g}_i\right)\boldsymbol{\Phi}^\top + \sum_{j=1}^{J}\boldsymbol{\Gamma}^{(j)}\text{cov}(\boldsymbol{f}_i^{(j)})\boldsymbol{\Gamma}^{(j)\top} + \sum_{j=1}^{J}\boldsymbol{I}^{(j)}\sigma^{(j)2}.$$

In this section, we add another prevalent covariance estimation method POET (Fan et al., 2013a) as aforementioned into our comparison.

The second panel of Table 3.2 and the second column of Figure 3.4 demonstrate the mean/standard errors and boxplot of $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2$ for all three examples, computed using methods discussed before. From these results, we see that POET performs better than PCA but worse than CPCA_I and CPCA_F, indicating that the sparsity structure implemented in POET can partly capture the sub-homogeneity, while not in a very efficient way. Among these methods, CPCA_F achieves the lowest mean and standard error of $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F^2$ because it can best identify both homogeneity and sub-homogeneity to accurately estimate $\boldsymbol{\Sigma}$.

3.8 Real data analysis

For further illustration, we analyze a stock return data set using the proposed CPCA and compare it with traditional PCA. As a result of the poor performance of CPCA_I_ns and CPCA_F_ns in our simulation studies, as well as the real analysis below, we have removed these two methods from the discussion in this section.

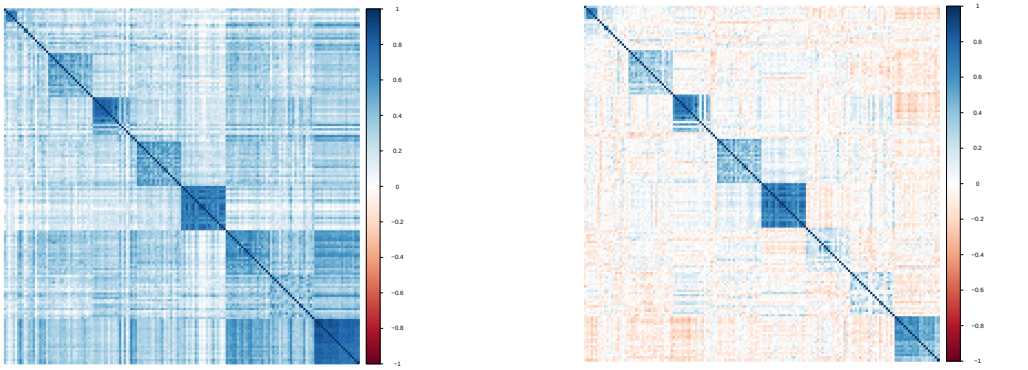
The data are collected from the Center for Research in Security Prices and

include the daily stock returns of 160 companies from 1st January, 2014 to 31st December, 2014, with 252 trading days. The 160 stocks are selected from eight different industries according to Fama and French's 48-industry classification (Fama and French, 1997), namely, Candy and Soda, Tobacco Products, Apparel, Aircraft, Shipbuilding and Railroad Equipment, Petroleum and Natural Gas, Measuring and Control Equipment, and Shipping Containers, with 20 stocks from each industry. The data for the first 126 trading days are treated as the training sample, and the rest are the testing sample. Thus, the training data have the dimensions $n = 126$ and $p = 160$. In this example, after removing the common effect from the data, we aim to identify the clusters that consist of companies from the same industry. This cluster structure of stock returns is also discussed in Fan et al. (2013a).

Figure 3.5a shows the correlation plot for the original stock data. We can observe a vague cluster structure, but the off-diagonals are clearly non-zero, implying that the stocks from the different industries selected in our study tend to be positively correlated. We apply hierarchical clustering directly to the original data and find that most stocks from different industries are clustered in the same group, as displayed in Figure 3.6b. Only stocks from 5^{th} industry stand out as another cluster, because they have a relatively lower correlation with other stocks. If we consider the industries as true clusters, the hierarchical clustering of the original data results in an ARI of only 0.09. This indicates that common components may exist and conceal the cluster structure within the data.

Next, we apply our proposed CPCA method to the stock data, and one common component is determined. This is not surprising because the largest eigenvalue of the sample covariance (0.020) is much larger than the second-largest eigenvalue (0.005). Figure 3.5b presents the correlation plot of the data after the common effect is removed. Comparatively, the cluster structure is more apparent. One can interpret the common effect as the market effect and the rest of the components as industry-specific effects. Further, the final clusters produced by the CPCA obtain an ARI of 0.67. As demonstrated in Figures 3.6, the final clusters mainly capture the industry information, but a few stocks are not

clustered into their own industries. This is to be expected because some stocks from the same industry are not highly correlated, as shown in Figures 3.5 (e.g., some stocks from the first, sixth and seventh industries), which can be common in reality. From the prediction perspective, we compute the MSRE for PCA, CPCA_I, and CPCA_F as 1.87, 1.18, and 0.95 (in unit of 10^{-4}), respectively. All of these findings firmly support that our proposed CPCA method is appropriate for analyzing these stock return data.

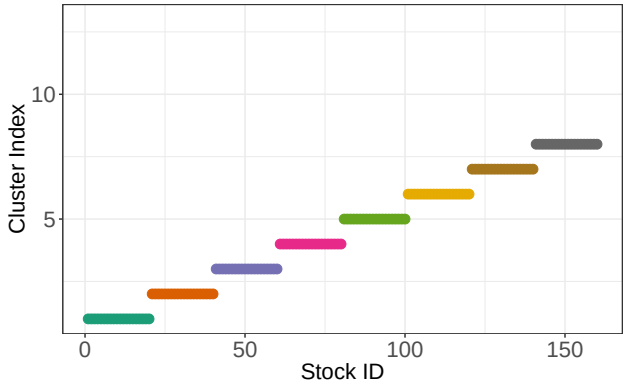


(a) The original stock return $\boldsymbol{X}$ (b) The complement $\boldsymbol{X}_f^c = \boldsymbol{X} - \boldsymbol{G}_f \boldsymbol{\Phi}_f^\top$

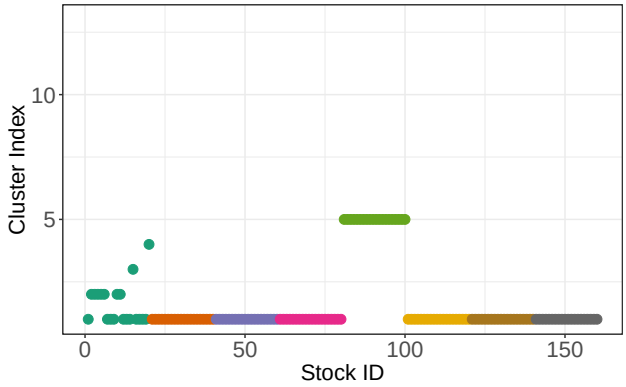
Figure 3.5: Correlation plot for stock return data

3.9 Conclusions and discussions

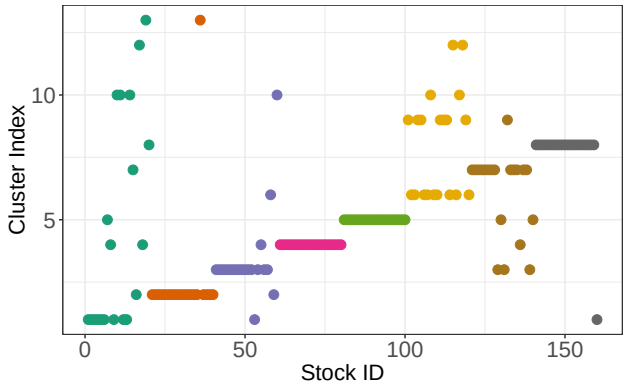
To conclude, we introduced a novel CPCA method to study the homogeneity and sub-homogeneity of high-dimensional data collected from different populations, where the sub-homogeneity refers to a group-specific feature from a particular population. Our numerical simulations confirmed that traditional PCA can only extract the homogeneity from the data, whereas CPCA not only provides a more accurate estimate of the common components, but also identifies the group-specific features, even for the group with small variations. The features extracted using CPCA can significantly outperform those selected by classical PCA in terms of prediction and covariance estimation, especially when n is small but p is large. Our real analysis of the stock return data also demonstrated that, when using the CPCA method, we can capture the industry information after the



(a) Cluster membership based on industry



(b) Hierarchical clustering of the original data



(c) Final clusters obtained from the CPCA

Figure 3.6: Cluster membership for 160 stocks. Stocks with the same color are from the same industry.

common component (i.e., market effect) is removed. All of these findings support the use of our proposed CPCA method in dimension-reduction problems.

The applications of CPCA are not limited to producing principal components in PCR and revealing the industry structure from the stock return data. CPCA can also be applied to any other data sets in which a group structure exists but hides in the homogeneity. Further, it can be used to estimate a covariance matrix and its inverse of a large data set that exhibits a group structure. More applications of CPCA will be explored in our future work.

3.A Appendix A: Estimations in CPCA Iterative Step (a)

In this part, we describe the estimation of the principal components, along with their eigenvectors, combined from each cluster in Iterative Step (a) of Algorithm 1.

First, we perform PCA in cluster j by:

$$\boldsymbol{X}^{(j)} = \boldsymbol{\Psi}^{(j)}\boldsymbol{\Pi}^{(j)\prime} + \boldsymbol{U}^{(j)},$$

where $\boldsymbol{\Psi}^{(j)}$ is an $n \times r_j$ matrix of principal components for variables in cluster j and $\boldsymbol{\Pi}^{(j)}$ is a $p_j \times r_j$ matrix in which each column represents an eigenvector of $\boldsymbol{X}^{(j)\prime}\boldsymbol{X}^{(j)}$. Here, recall that p_j denotes the number of variables in cluster j, and r_j is the number of principal components in cluster j. Then, we can combine the principal components $\boldsymbol{\Psi}^{(j)}$ from each cluster as $\boldsymbol{\Psi} = (\boldsymbol{\Psi}^{(1)}, ..., \boldsymbol{\Psi}^{(J_0)})$ and perform a further step of PCA on $\boldsymbol{\Psi}$ to obtain:

$$\boldsymbol{\Psi} = \boldsymbol{G}_1\boldsymbol{H}' + \boldsymbol{V},$$

where we assume the first r_c principal components can be used to summarize the common effects among all clusters. Then, $\boldsymbol{G}_1$ is a $n \times r$ matrix of principal components of $\boldsymbol{\Psi}$, and $\boldsymbol{H}$ is a $(\sum_{j=1}^{J_0} r_j) \times r_c$ matrix in which each column represents

an eigenvector of $\mathbf{\Psi}'\mathbf{\Psi}$. Lastly, we can find the complement for cluster j as:

$$\boldsymbol{X}_1^{(j)c} = \boldsymbol{X}^{(j)} - \boldsymbol{G}_1\boldsymbol{H}'\mathbf{\Pi}^{(j)\star\prime},$$

where $\mathbf{\Pi}^{(j)\star} = (\mathbf{0}^{(1)}, \mathbf{0}^{(2)}, ..., \mathbf{\Pi}^{(j)}, ..., \mathbf{0}^{(J_0)})$ is a $p_j \times (\sum_{j=1}^{J_0} r_j)$ matrix in which $\mathbf{\Pi}^{(j)}$ is a $p_j \times r_j$ matrix, as we described earlier, and $\mathbf{0}^{(l)}$ is a $p_j \times r_l$ zero matrix for $l = 1, 2, ..., J_0;\ l \neq j$. Once we obtain the complement from each cluster, we can combine them and present the total complement using:

$$\boldsymbol{X}_1^{c} = \boldsymbol{X} - \boldsymbol{G}_1\boldsymbol{H}'\mathbf{\Pi}',$$

where $\mathbf{\Pi} = (\mathbf{\Pi}^{(1)\star\prime}, \mathbf{\Pi}^{(2)\star\prime}, ..., \mathbf{\Pi}^{(j)\star\prime}, ..., \mathbf{\Pi}^{(J_0)\star\prime})'$ is a $p \times (\sum_{j=1}^{J_0} r_j)$ block-diagonal matrix. Hence, we can finally define $\mathbf{\Phi_1} = \mathbf{\Pi}\boldsymbol{H}$ as the corresponding eigenvectors for $\boldsymbol{G}_1$. It is also easy to show that $\mathbf{\Phi}_\mathbf{1}^{'}\mathbf{\Phi_1} = I_p$.

3.B Appendix B: Proof of Theorem 3.1

The matrix form of (3.11) is:

$$\boldsymbol{u}_m = \boldsymbol{F}^{(d)}\boldsymbol{\beta} + \boldsymbol{\zeta}_m^{(d)}, \tag{3.18}$$

where $\boldsymbol{u}_m = \left(u_{m1}^{(l)}, u_{m2}^{(l)}, \ldots, u_{mn}^{(l)}\right)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{r_d})$.
Hence, the least-squares estimation of the residual $\boldsymbol{\zeta}_m^{(d)}$ is:

$$\widetilde{\boldsymbol{\zeta}}_m^{(d)} = \boldsymbol{u}_m - \boldsymbol{F}^{(d)}\left(\boldsymbol{F}^{(d)'}\boldsymbol{F}^{(d)}\right)^{-1}\boldsymbol{F}^{(d)'}\boldsymbol{u}_m =: \boldsymbol{M}_{\boldsymbol{F}^{(d)}}\boldsymbol{u}_m. \tag{3.19}$$

It should be noted that $\boldsymbol{u}_m$ and $\boldsymbol{F}^{(d)}$ are not observed. In terms of our proposed clustering method, we estimate $\boldsymbol{u}_m$ via PCA, such that:

$$\widehat{\boldsymbol{u}}_m = \boldsymbol{M}_{\widehat{\boldsymbol{G}}}\boldsymbol{x}_m, \tag{3.20}$$

where $\boldsymbol{M}_{\widehat{\boldsymbol{G}}} = \boldsymbol{I}_n - \widehat{\boldsymbol{G}}(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{G}})^{-1}\widehat{\boldsymbol{G}}'$, with $\widehat{\boldsymbol{G}}$ being an estimator for $\boldsymbol{G}$, as described in CPCA.

For $\boldsymbol{F}^{(d)}$, we also perform PCA on cluster d and denote the estimator by $\widehat{\boldsymbol{F}}^{(d)}$. Based on (3.19) and (3.20), our estimator of $\boldsymbol{\zeta}_m^{(d)}$ can be written as:

$$\begin{aligned}
\widehat{\boldsymbol{\zeta}}_m^{(d)} &= \boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} \widehat{\boldsymbol{u}}_m = \boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} \boldsymbol{M}_{\widehat{\boldsymbol{G}}} \boldsymbol{x}_m \\
&= \boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} \left(\boldsymbol{M}_{\widehat{\boldsymbol{G}}} - \boldsymbol{M}_{\boldsymbol{G}} \right) \boldsymbol{x}_m + \boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{x}_m \\
&= \left(\boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} - \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \right) \left(\boldsymbol{M}_{\widehat{\boldsymbol{G}}} - \boldsymbol{M}_{\boldsymbol{G}} \right) \boldsymbol{x}_m + \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \left(\boldsymbol{M}_{\widehat{\boldsymbol{G}}} - \boldsymbol{M}_{\boldsymbol{G}} \right) \boldsymbol{x}_m \\
&\quad + \left(\boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} - \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \right) \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{x}_m + \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \boldsymbol{M}_{\boldsymbol{G}} \boldsymbol{x}_m. \qquad (3.21)
\end{aligned}$$

To define $\| \boldsymbol{M}_{\widehat{\boldsymbol{F}}^{(d)}} - \boldsymbol{M}_{\boldsymbol{F}^{(d)}} \|_2 = O_p(\alpha_{np_d})$, $\| \boldsymbol{M}_{\widehat{\boldsymbol{G}}} - \boldsymbol{M}_{\boldsymbol{G}} \|_2 = O_p(\gamma_{nJ})$, we obtain the result in this theorem.

Chapter 4

Autocovariance Test for High-dimensional Time Series

4.1 Introduction

In this chapter, we propose a novel test statistic named the autocovariance test to compare the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series. This autocovariance test is built based on a factor model approach where the temporal dependence of the original high-dimensional data are captured by the factors. The test statistic is computed based on a central limit theorem (CLT) on spiked eigenvalues of the symmetrized sample autocovariance matrix for high-dimensional time series, which is derived my joint work (Bi et al., 2020).

With recent developments in information technologies, more and more data are now available for statistical analysis, including hypothesis testing. Among various types of data, high-dimensional time series have been widely seen in many disciplines, including economics, finance, meteorology, and biology. Equivalent test for two high-dimensional time series is becoming increasingly important since multiple data-sets may be aggregated for further statistical analysis. As a traditional and fundamental statistical inference for univariate and multivariate data, hypothesis testing for comparing two samples has been widely discussed; for example, the well-known Z-test and Student's T-test have been considered for the test on the equivalence of means, and the F-test can be performed for the equivalence of variances. For multivariate data, the mean-

based Hotelling's T^2 test is a generalisation of the Student's T-test and has also received developments for growing dimensions (see, e.g. Pan and Zhou, 2011). Besides, for testing the equivalence of population means of high-dimensional data, Bai and Saranadasa (1996) proposes a test based on the squared norm of the difference between sample means, where the dimensions and sample sizes are required to be of the same order, while Chen and Qin (2010) proposed a two-sample test by restricting the trace of the common sample covariance matrix. However, for high-dimensional time series, there are both cross-sectional and temporal dependences within the data-sets, while the above tests for the equivalence of population means do not consider the temporal dependence of the data. In this chapter, we propose a test that incorporates both cross-sectional and temporal dependence within the data-set for comparing two high-dimensional time series.

Besides, statistical analysis for high-dimensional data, especially time series, also suffers from the 'curse of dimensionality', where sample estimates are not asymptotically consistent to their population counterparts. Hence, dimension-reduction methods can be utilised to project the original high-dimensional time series into a low-dimensional space spanned by the eigenvectors corresponding to the spiked eigenvalues of the autocovariance matrix. It is then convenient to study the projected data in the low-dimensional space. In the literature of dimension-reduction methods, principal component analysis (Jolliffe, 2002) and factor models (Bai and Ng, 2002b; Bai, 2003b) are popular and have received most attention and developments. For high-dimensional dependent data, to capture both cross-sectional and temporal dependence of the original data, dynamic principal components (Ku et al., 1995) can be considered. Alternatively, Lam et al. (2011) suggest using approximate factor models to study high-dimensional time series, where the temporal dependence of the original high-dimensional time series are captured by low-dimensional factor time series. In summary, to simultaneously study both the cross-sectional and temporal dependence between two high-dimensional time series, we can reduce the dimensions of the original high-dimensional time series by projecting them into the low-dimensional eigenspace

and study the autocovariance of the low-dimensional factors. In other words, we mainly contribute to checking if the factors of two high-dimensional time series, after projection onto the same eigenspace, have the same autocovariance matrix. Besides, it is worth to be noticed that we are testing if the factors have the same autocovariance, which is equivalent to test whether the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series are the same.

In summary we propose a novel autocovariance test based on the spiked eigenvalues of the symmetrized sample autocovariance matrices to compare two high-dimensional time series. Consequently, for two high-dimensional time series sharing the same eigenstructure of the autocovariance matrix, aggregated and simultaneous studies may also be considered.

The rest of chapter is organised as follows. Section 4.2 proposes the model of the autocovariance test. In specific, the hypothesis we are interested in and the test statistic based on the spiked eigenvalues of the symmetrized autocovariance matrix for two high-dimensional time series are proposed, where we also introduce the identification and regularisation conditions on the factor model and study the asymptotic power of the autocovariance test. Section 4.3 illustrates the implementation of the testing procedure, including the estimations for unknown parameters that determine the test statistic. In Section 4.4, we use numerical simulations to explore the empirical sizes and powers of the proposed test under various settings. Section 4.5 provides an example of applying our autocovariance test on age-specific mortality rates for multiple countries. Conclusions and discussions are presented in Section 4.6, whereas technical proofs are in Appendix 4.A.

4.2 Model

4.2.1 Hypotheses and test statistic

When the data dimension N increases, direct comparison on the sample autocovariance matrices for two high-dimensional time series is not feasible due to

the 'curse of dimensionality'. With dimension-reduction methods such as factor models proposed in Lam et al. (2011) and Lam and Yao (2012), high-dimensional time series can be projected into a low-dimensional space spanned by the eigenvectors (factor loadings) obtained from applying eigendecomposition on the autocovariance matrix of the original high-dimensional time series data. It is then feasible to compare two time series in the same low-dimensional space spanned by the eigenvectors (factor loadings) of both high-dimensional time series. To explain the idea of the autocovariance test, we will firstly introduce the factor models for high-dimensional time series.

Consider now two independent high-dimensional stationary time series $\left\{\boldsymbol{y}_t^{(1)} \in \mathbb{R}^N,\ t = 1, 2, ..., T\right\}$ and $\left\{\boldsymbol{y}_t^{(2)} \in \mathbb{R}^N,\ t = 1, 2, ..., T\right\}$ following factor models

$$\boldsymbol{y}_t^{(k)} = \boldsymbol{Q}^{(k)} \boldsymbol{f}_t^{(k)} + \boldsymbol{u}_t^{(k)},\ k = 1, 2, \tag{4.1}$$

where $\left\{\boldsymbol{f}_t^{(k)} \in \mathbb{R}^{r_k},\ t = 1, 2, ..., T\right\}$ are stationary factor time series with $r_k \ll N$, and $\boldsymbol{Q}^{(k)}$ is a $N \times r_k$ factor loading matrix with a normalisation condition $\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)} = \boldsymbol{I}_{r_k}$.

For high-dimensional time series $\left\{\boldsymbol{y}_t^{(k)}\right\}$ following factor models such as (4.1), $\boldsymbol{Q}^{(k)}$ is the time invariant factor loading matrix. Without altering the idea in Lam et al. (2011) and Lam and Yao (2012), we consider i.i.d. $\left\{\boldsymbol{u}_t^{(k)}\right\}$ which is also independent of $\left\{\boldsymbol{f}_t^{(k)}\right\}$ to illustrate the idea of our test. With this set-up, the autocovariance matrices of $\left\{\boldsymbol{y}_t^{(k)}\right\}$ can be represented by the factor loading matrix $\boldsymbol{Q}^{(k)}$ and the autocovariance matrices of $\left\{\boldsymbol{f}_t^{(k)}\right\}$ as

$$\boldsymbol{\Gamma}_y^{(k)}(\tau) = \boldsymbol{Q}^{(k)} \boldsymbol{\Gamma}_f^{(k)}(\tau) \boldsymbol{Q}^{(k)\top}$$

for $\tau \geq 1$, where $\boldsymbol{\Gamma}_y^{(k)}(\tau) = \text{Cov}\left(\boldsymbol{y}_t^{(k)}, \boldsymbol{y}_{t+\tau}^{(k)}\right)$ is the lag-τ autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$, and $\boldsymbol{\Gamma}_f^{(k)}(\tau) = \text{Cov}\left(\boldsymbol{f}_t^{(k)}, \boldsymbol{f}_{t+\tau}^{(k)}\right)$ is the lag-τ autocovariance matrix of $\left\{\boldsymbol{f}_t^{(k)}\right\}$. As a result, the temporal dependence, in the form of autocovariance matrix, of $\left\{\boldsymbol{y}_t^{(k)}\right\}$ is fully captured by $\left\{\boldsymbol{f}_t^{(k)}\right\}$.

Next, we will study the theoretical property of the factor loading matrix $\boldsymbol{Q}^{(k)}$. As discussed in Lam et al. (2011), we can decompose the symmetrized lag-τ

autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$ as

$$\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top = \boldsymbol{Q}^{(k)}\boldsymbol{\Gamma}_f^{(k)}(\tau)\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)}\boldsymbol{\Gamma}_f^{(k)}(\tau)^\top\boldsymbol{Q}^{(k)\top} = \boldsymbol{Q}^{(k)}\boldsymbol{\Gamma}_f^{(k)}(\tau)\boldsymbol{\Gamma}_f^{(k)}(\tau)^\top\boldsymbol{Q}^{(k)\top}. \tag{4.2}$$

Denote by $\mu_{i,\tau}^{(k)}$ the eigenvalues of $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top$. We then consider in this chapter the setting where there are r_k spiked eigenvalues in $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top$ and $\mu_{1,\tau}^{(k)} > \mu_{2,\tau}^{(k)} > ... > \mu_{r_k,\tau}^{(k)}$ tend to infinity with N, while $\mu_{r_{k+1},\tau}^{(k)} = \mu_{r_{k+2},\tau}^{(k)} = ... = \mu_{r_N,\tau}^{(k)} = 0$ for some $\tau \geq 1$. With this definition of spiked eigenvalues, we can show that the columns of $\boldsymbol{Q}^{(k)}$ are the eigenvectors of $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top$ corresponding to the spiked eigenvalues, as follow.

Write $\boldsymbol{W}^{(k)}$ for an $N \times (N - r_k)$ matrix where $\left(\boldsymbol{Q}^{(k)}, \boldsymbol{W}^{(k)}\right)$ forms a $N \times N$ orthogonal matrix so that $\boldsymbol{Q}^{(k)\top}\boldsymbol{W}^{(k)} = \boldsymbol{0}$ and $\boldsymbol{W}^{(k)\top}\boldsymbol{W}^{(k)} = \boldsymbol{I}_{N-r_k}$. It follows from (4.2) that $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top\boldsymbol{W}^{(k)} = \boldsymbol{0}$, which means the columns of $\boldsymbol{W}^{(k)}$ are precisely the eigenvectors associated with zero-eigenvalues. In other words, the columns of $\boldsymbol{Q}^{(k)}$ are the r_k eigenvectors of $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top$ corresponding to those non-zero eigenvalues, and those non-zero eigenvalues of $\boldsymbol{\Gamma}_y^{(k)}(\tau)\boldsymbol{\Gamma}_y^{(k)}(\tau)^\top$ are precisely the eigenvalues of $\boldsymbol{\Gamma}_f^{(k)}(\tau)\boldsymbol{\Gamma}_f^{(k)}(\tau)^\top$. Besides, the condition $\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)} = \boldsymbol{I}_{r_k}$ is not sufficient for $\boldsymbol{Q}^{(k)}$ to be uniquely defined, but only defines a so-called eigenspace as $\mathcal{M}\left(\boldsymbol{Q}^{(k)}\right)$.

Consequently, on one hand, $\mathcal{M}\left(\boldsymbol{Q}^{(k)}\right)$ is the eigenspace spanned by the columns of $\boldsymbol{Q}^{(k)}$, which is also the eigenvectors corresponding to the spiked eigenvalues of the symmetrized autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$. On the other hand, the eigenvalues of the symmetrized autocovariance matrix of $\left\{\boldsymbol{f}_t^{(k)}\right\}$, which summarise the information contained in the autocovariance matrix of $\left\{\boldsymbol{f}_t^{(k)}\right\}$, are precisely the spkied eigenvalues of the symmetrized autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$. Therefore, by assuming $\mathcal{M}\left(\boldsymbol{Q}^{(1)}\right) = \mathcal{M}\left(\boldsymbol{Q}^{(2)}\right)$, we can build a test statistic based on the difference between spiked eigenvalues of the symmetrized lag-τ sample autocovariance matrices of two high-dimensional time series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$.

In this chapter, it is worth noting that we typically focus on testing the

equivalence of spiked eigenvalues, but not the eigenspace of autocovariance matrices for two high-dimensional time series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$. Consequently, for a finite τ, the null and alternative hypothesis of the autocovariance test for two high-dimensional time series can be summarised as

Test 4.1. *(Autocovariance test for two high-dimensional time series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$)*

$$H_0\text{: } \mu_{i,\tau}^{(1)} = \mu_{i,\tau}^{(2)} \text{ for all } i = 1, 2, ..., r_k$$
$$H_1\text{: } \mu_{i,\tau}^{(1)} \neq \mu_{i,\tau}^{(2)} \text{ for at least one } i,\ i = 1, 2, ..., r_k$$

For factor models in canonical form which will be specified in Section 4.2.2, write $\gamma_{i,\tau}^{(k)} := \mathbb{E}\left(f_{i,1}^{(k)} f_{i,\tau+1}^{(k)}\right)$ and $\left(v_{i,\tau}^{(k)}\right)^2 := \frac{1}{T-\tau} Var\left(\sum_{t=1}^{T-\tau} f_{i,t}^{(k)} f_{i,t+\tau}^{(k)}\right)$ for a finite time lag τ, $i = 1, 2, ..., r_k$ and $k = 1, 2$. Denote by $\lambda_{i,\tau}^{(k)}$ the i-th largest spiked eigenvalue of the symmetrized lag-τ sample autocovariance matrix $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)^\top$, where $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau) = \frac{1}{T-\tau-1}\sum_{t=1}^{T-\tau}(\boldsymbol{y}_t^{(k)} - \overline{\boldsymbol{y}}_T^{(k)})(\boldsymbol{y}_{t+\tau}^{(k)} - \overline{\boldsymbol{y}}_T^{(k)})^\top$, for $k = 1, 2$. Then, for $i = 1, 2, ..., r_k$ and some finite τ, the test statistic is given by

$$Z_{i,\tau} = \sqrt{T}\frac{\gamma_{i,\tau}}{2\sqrt{2}v_{i,\tau}}\frac{\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}}, \tag{4.3}$$

where

$$\theta_{i,\tau} = \frac{\theta_{i,\tau}^{(1)} + \theta_{i,\tau}^{(2)}}{2},\ v_{i,\tau} = \frac{v_{i,\tau}^{(1)} + v_{i,\tau}^{(2)}}{2},\ \text{and } \gamma_{i,\tau} = \frac{\gamma_{i,\tau}^{(1)} + \gamma_{i,\tau}^{(2)}}{2}, \tag{4.4}$$

and $\theta_{i,\tau}^{(k)}$ is the asymptotic centring of $\lambda_{i,\tau}^{(k)}$. It is worth noting that, the exact definition of $\theta_{i,\tau}^{(k)}$ is in Proposition 1.3 of Bi et al. (2020), which is rather involved and requires technical details that are beyond the scope of this chapter. It is then clearly that $|Z_{i,\tau}|$ will be generally large if $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ follow different factor models where the i-th largest eigenvalues of the symmetrized lag-τ sample autocovariance matrix for two factor models are different. We name this test by autocovariance test since the idea behind is testing whether two independent high-dimensional time series observations share the same spiked eigenvalues of the autocovariance matrices.

4.2.2 Factor model and regularisation conditions

As discussed in Chapter 2, factor models have been widely discussed in the literature, and there are various identification conditions on $\boldsymbol{Q}^{(k)}$, $\boldsymbol{f}_t^{(k)}$, and $\boldsymbol{u}_t^{(k)}$ in factor models (4.1). In this work, we adopt the idea in Lam et al. (2011) again and assume the temporal dependence of $\left\{\boldsymbol{y}_t^{(k)}\right\}$ can be fully captured by the factors $\left\{\boldsymbol{f}_t^{(k)}\right\}$ with a time invariant factor loading matrix $\boldsymbol{Q}^{(k)}$. In other words, we still work in the scheme where a static relationship between $\left\{\boldsymbol{y}_t^{(k)}\right\}$ and $\left\{\boldsymbol{f}_t^{(k)}\right\}$ is maintained.

In a general factor model set-up that has been well discussed in Bai and Ng (2002b) and Bai (2003b), the idiosyncratic components $\left\{\boldsymbol{u}_t^{(k)} \in \mathbb{R}^N, t = 1, 2, ..., T\right\}$ are assumed to be independent of the factors $\left\{\boldsymbol{f}_t^{(k)}\right\}$, with $\mathbb{E}\left(u_{j,t}^{(k)}\right) = 0$ and $\mathbb{E}\left(u_{j,t}^{(k)}\right)^2 =: \left(\sigma_u^{(k)}\right)^2 < \infty$ for $j = 1, 2, ..., N;\ t = 1, 2, ..., T$. Without loss of generality, we can work on standardised factor models where the variance of noise component is normalised to one. For factor models with $\left(\sigma_u^{(k)}\right)^2 \neq 1$, we can standardise it by dividing $\sigma_u^{(k)}$ on both sides. This standardisation on $\left\{\boldsymbol{y}_t^{(k)}\right\}$, facilitates the comparison between two high-dimensional time series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ by restricting them to have the same scale of noises. In this work, we follow the set-up in Li et al. (2017) and assume $u_{j,t}^{(k)} \sim \mathcal{N}(0, 1)$ for all j, t and k.

In addition to the assumptions made on $\left\{\boldsymbol{u}_t^{(k)}\right\}$, we assume the factors $\left\{f_{i,t}^{(k)}, i = 1, 2, ..., r_k;\ t = 1, 2, ..., T;\ k = 1, 2\right\}$ in (4.1) are given by a stationary time series

$$f_{i,t}^{(k)} = \sum_{l=0}^{\infty} \psi_{i,l}^{(k)} z_{i,t-l}^{(k)},\ i = 1, 2, ..., r_k,\ t = 1, 2, ..., T, \tag{4.5}$$

where the random variables $\left\{z_{i,t}^{(k)}\right\}$ are i.i.d. with mean zero, variance one and finite fourth moments. Without loss of generality, we have assumed $\mathbb{E}\left(f_{i,t}^{(k)}\right) = 0$ for $i = 1, 2, ..., r_k,\ t = 1, 2, ..., T$, and $k = 1, 2$. Besides, to compare two factor models, we can also impose a condition on $\left\{f_{i,t}^{(k)}\right\}$ such that the variance is normalised to one, i.e., $\mathbb{E}\left(f_{i,t}^{(k)}\right)^2 = 1$. In other words, we require $\left\|\boldsymbol{\psi}_i^{(k)}\right\|_2 = 1$ for $i = 1, 2, ..., r_k$, and $k = 1, 2$, where $\boldsymbol{\psi}_i^{(k)} := \left(\psi_{i,1}^{(k)}, \psi_{i,2}^{(k)}, ...\right)$ is the vector of

coefficients for the i-th factor $f_{i,t}^{(k)}$. In general, this normalisation condition can be considered as an identification condition made on $\boldsymbol{Q}^{(k)}$ and $\boldsymbol{f}_t^{(k)}$, where $\boldsymbol{Q}^{(k)}$ is no longer orthonormal and $\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)} = diag\left(\left(\sigma_1^{(k)}\right)^2, \left(\sigma_2^{(k)}\right)^2, ,..., \left(\sigma_{r_k}^{(k)}\right)^2\right)$ with $\left(\sigma_i^{(k)}\right)^2$ the un-normalised variance of $f_{i,t}^{(k)}$.

Lastly, we impose an additional identification condition on the loading matrix $\boldsymbol{Q}^{(k)}$ in (4.1) for technical convenience. To simplify factor models (4.1), we firstly introduce the following normalisation process discussed in Li et al. (2017), where we can consider a factor model with the loading matrix $\boldsymbol{Q}^{(k)}$ in the following canonical form

$$\boldsymbol{Q}^{(k)} = \begin{pmatrix} \boldsymbol{I}_{r_k} \\ \boldsymbol{0}_{N-r_k} \end{pmatrix}. \tag{4.6}$$

This is because for any factor loading matrix $\boldsymbol{Q}^{(k)}$ in (4.1) that is not in the above canonical form, we can find an orthonormal matrix $\boldsymbol{P}^{(k)} = \left(\boldsymbol{Q}^{(k)}, \boldsymbol{W}^{(k)}\right)$ such that $\boldsymbol{W}^{(k)\top}\boldsymbol{Q}^{(k)} = \boldsymbol{0}$, and normalise the factor models by left multiplying the transpose of the orthonormal matrix $\boldsymbol{P}^{(k)}$. Note that by condition $\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)} = \boldsymbol{I}_r$, we have

$$\boldsymbol{P}^{(k)\top}\boldsymbol{y}_t^{(k)} = \boldsymbol{P}^{(k)\top}\boldsymbol{Q}^{(k)}\boldsymbol{f}_t^{(k)} + \boldsymbol{P}^{(k)\top}\boldsymbol{u}_t^{(k)} = \begin{pmatrix} \boldsymbol{I}_{r_k} \\ \boldsymbol{0}_{N-r_k} \end{pmatrix} \boldsymbol{f}_t^{(k)} + \boldsymbol{P}^{(k)\top}\boldsymbol{u}_t^{(k)},$$

where $\boldsymbol{P}^{(k)\top}\boldsymbol{u}_t^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}_N)$ since $\boldsymbol{u}_t^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1}_N)$ and $\boldsymbol{P}^{(k)}$ is orthonormal. Therefore, the transferred data $\boldsymbol{P}^{(k)\top}\boldsymbol{y}_t^{(k)}$ fulfils a factor model with canonical loading matrix. Since $\boldsymbol{P}^{(k)}$ is orthonormal, this transfer is nothing but a rotation made on the original data, which does not alter the eigenspace of the autocovariance matrix. Instead, we can assume the variances of factors are absorbed into the loading matrix $\boldsymbol{Q}^{(k)}$ and impose a different canonical form condition on $\boldsymbol{Q}^{(k)}$

as

$$\boldsymbol{Q}^{(k)} = \begin{pmatrix} \sigma_1^{(k)} & & \\ & \ddots & \\ & & \sigma_{r_k}^{(k)} \\ & \mathbf{0}_{N-r_k} & \end{pmatrix}, \tag{4.7}$$

where we stress the fact that $\boldsymbol{Q}^{(k)\top}\boldsymbol{Q}^{(k)} = diag\left(\left(\sigma_1^{(k)}\right)^2, \left(\sigma_2^{(k)}\right)^2, , ..., \left(\sigma_{r_k}^{(k)}\right)^2\right)$ with the normalisation on the variances of factors. It is worth noting that since we have required the variances of factors $\left\{f_{i,t}^{(k)}\right\}$ to be normalised to 1, the canonical loading matrix $\boldsymbol{Q}^{(k)}$ defined by (4.7) is different from that in Li et al. (2017).

Consequently, the factor models are now simplified to a canonical form as

$$\boldsymbol{y}_t^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{f}_t^{(k)} + \boldsymbol{u}_t^{(k)} = \begin{pmatrix} \sigma_1^{(k)} f_{1,t}^{(k)} \\ \vdots \\ \sigma_{r_k}^{(k)} f_{r_k,t}^{(k)} \\ \mathbf{0}_{N-r_k} \end{pmatrix} + \boldsymbol{u}_t^{(k)}, \tag{4.8}$$

where $\left\{\sigma_i^{(k)}, i = 1, 2, ..., r_k;\ k = 1, 2\right\}$ are positive real numbers representing (cross-sectional) factor strengths, where we refer to Lam et al. (2011) for the definition of factor strengths.

In summary, we consider factor models in canonical form (4.8), where the loading matrix $\boldsymbol{Q}^{(k)}$ is defined by (4.7) and the variances of $\left\{f_{i,t}^{(k)}\right\}$ and $\left\{u_{j,t}^{(k)}\right\}$ are normalised to 1. In addition, we assume the data $\left\{\boldsymbol{y}_t^{(k)}\right\}$ comes from strong factor models where $\sigma_i^{(k)}$ is divergent as $N \to \infty$ for $i = 1, 2, ..., r_k$ and $k = 1, 2$. Besides, for a general strong factor model that is not in the canonical form 4.7, it can be normalised by standardising the variance of $\left\{u_{j,t}^{(k)}\right\}$ to one first and then rotating the original data such that the loading matrix $\boldsymbol{Q}^{(k)}$ is in the canonical form (4.7).

Moreover, recall that for a finite time lag τ, $\gamma_{i,\tau}^{(k)} := \mathbb{E}\left(f_{i,1}^{(k)} f_{i,\tau+1}^{(k)}\right)$ is the population lag-τ autocovariance (autocorrelation) of the i-th factor time series $\left\{f_{i,t}^{(k)}\right\}$. Since each factor in (4.1) is assumed to be stationary following (4.5), $\gamma_{i,\tau}^{(k)}$

can also be written as

$$\gamma_{i,\tau}^{(k)} = \mathbb{E}\left(f_{i,1}^{(k)} f_{i,\tau+1}^{(k)}\right) = \sum_{l=0}^{\infty} \psi_{i,l}^{(k)} \psi_{i,l+\tau}^{(k)}.$$

Therefore, for high-dimensional time series $\left\{y_{i,t}^{(k)}\right\}$ fulfilling factor models (4.8), the population lag-τ autocovariance can be defined as

$$\mu_{i,\tau}^{(k)} := \mathbb{E}\left(y_{i,t}^{(k)} y_{i,t+\tau}^{(k)}\right) = \left(\sigma_i^{(k)}\right)^2 \gamma_{i,\tau}^{(k)}, \text{ for } i = 1, 2, ..., r_k;\ k = 1, 2,$$

where the time lag τ is a prescribed positive integer. Besides, to compare two high-dimensional time series in the same eigenspace, we require the eigenvalues of the symmetrized lag-τ sample autocovariance matrix to be distinct and well-separated asymptotically. For technical convenience, we have also assumed without loss of generality that $\mu_{1,\tau}^{(k)} > \mu_{2,\tau}^{(k)} > \cdots > \mu_{r_k,\tau}^{(k)} > 0$ in Section 4.2.1, therefore each population eigenvector can be uniquely identified and recovered (Lam and Yao, 2012).

Consequently, the assumptions for high-dimensional time series fulfilling factor models in canonical form (4.8) are summarised below.

Assumptions 4.1.

(i) $N \to \infty$ *and* $N/T \to c > 0$ *as* $T \to \infty$.

(ii) $r_k = o\left(T^{1/16}\right)$ *and* $r_k = o\left(\left(\sigma_i^{(k)}\right)^2\right)$ *as* $T \to \infty$ *for* $i = 1, 2, ..., r_k$.

(iii) $\sigma_i^{(k)} \to \infty$ *as* $T \to \infty$ *for* $i = 1, 2, ..., r_k$, *and* $\sigma_i^{(k)}/\sigma_j^{(k)} = \mathcal{O}(1)$ *for* $i, j = 1, 2, ..., r_k$.

(iv) τ *is a fixed non-negative integer and* $\mu_{1,\tau}^{(k)} > \mu_{2,\tau}^{(k)} > \cdots > \mu_{r_k,\tau}^{(k)} > 0$.

(v) $\left\{u_{j,t}^{(k)}\right\}$ *are i.i.d.* $\mathcal{N}(0,1)$.

(vi) The coefficients in the linear process of $\left\{f_{i,t}^{(k)}\right\}$ *fulfils* $\sup_i \left\|\boldsymbol{\psi}_i^{(k)}\right\|_1 < \infty$ *and* $\left\|\boldsymbol{\psi}_i^{(k)}\right\|_2 = 1$; $\left\{z_{i,t}^{(k)}\right\}$ *are i.i.d. with* $\mathbb{E}\left(z_{i,t}^{(k)}\right) = 0$, $\mathbb{E}\left(z_{i,t}^{(k)}\right)^2 = 1$ *and uniformly bounded* $4 + \epsilon$ *moment for some* $\epsilon > 0$.

Remark 4.1. We note that (i) and (iii) of Assumption 4.1 capture our asymptotic regime where we allow N diverging at the same rate as T, and the strength of all factors diverge at comparable rates. The condition in (ii) allows the number of factors to diverge but at a relatively slow rate compared with T, and it is trivially satisfied when the number of factors r_k is finite. The condition in (iv) ensures that the empirical eigenvalues, as well as the eigenvectors, are asymptotically separable. Besides, as discussed in the previous section, the normality assumption in (v) is for the purpose of reducing the model to a canonical form. Lastly, the moments' conditions in (vi) are standard for time series studies (Anderson, 1971), while the conditions on $\boldsymbol{\psi}_i^{(k)}$ are for normalising purpose and can be satisfied by any causal autoregressive moving average processes.

4.2.3 Asymptotic results for the autocovariance test

To study the asymptotic properties of the autocovariance test for comparing two high-dimensional time series, we firstly present a result in my joint work (Bi et al., 2020), where a central limit theorem (CLT) on spiked eigenvalues of the symmetrized sample autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$ is developed. For factor models in canonical form (4.8), recall that $\gamma_{i,\tau}^{(k)} = \mathbb{E}\left(f_{i,1}^{(k)} f_{i,\tau+1}^{(k)}\right)$ and $\left(v_{i,\tau}^{(k)}\right)^2 = \frac{1}{T-\tau} Var\left(\sum_{t=1}^{T-\tau} f_{i,t}^{(k)} f_{i,t+\tau}^{(k)}\right)$ for a finite time lag τ, $i = 1, 2, ..., r_k$, and $k = 1, 2$. Then for the i-th largest spiked eigenvalue $\lambda_{i,\tau}^{(k)}$ of the symmetrized lag-τ sample autocovariance matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$, we have the following CLT.

Lemma 4.1 (Theorem 1.5 in Bi et al. (2020))**.** *Suppose that Assumption 4.1 hold, for $i = 1, 2, ..., r_k$ and some finite τ, it holds that*

$$\sqrt{T} \frac{\gamma_{i,\tau}^{(k)}}{2v_{i,\tau}^{(k)}} \frac{\lambda_{i,\tau}^{(k)} - \theta_{i,\tau}^{(k)}}{\theta_{i,\tau}^{(k)}} \Rightarrow \mathcal{N}(0,1), \tag{4.9}$$

as $T, N \to \infty$, where $\theta_{i,\tau}^{(k)}$ is the asymptotic centring of $\lambda_{i,\tau}^{(k)}$ and the exact definition can be found in Proposition 1.3 of Bi et al. (2020).

Lemma 4.1 provides the asymptotic properties of $\left\{\lambda_{i,\tau}^{(k)},\ i=1,2,...,r_k\right\}$ which are associated with the i-th factors in (4.8). Consider again for the two independent N-dimensional time series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ with the sample size T. Recall that $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$ are the i-th spiked eigenvalues of the symmetrized lag-τ sample autocovariance matrices of $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$, respectively. If $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ are assumed following the same factor model under Assumptions 4.1, independently, $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$ will also share the same asymptotic distribution as shown in Lemma 4.1, independently. Therefore, to test whether $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ share the same spiked eigenvalues of the autocovariance matrices, it is natural to create the test statistic (4.3) base on the difference between $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$. When $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ follow the same factor model in the canonical form (4.8), we have the following CLT on the difference between $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$.

Theorem 4.1. *Under the same assumptions of Lemma 4.1, for two independent high-dimensional time series* $\left\{\boldsymbol{y}_t^{(1)}\right\}$ *and* $\left\{\boldsymbol{y}_t^{(2)}\right\}$ *following the same factors in canonical form (4.8), we have*

$$Z_{i,\tau} = \sqrt{T}\frac{\gamma_{i,\tau}}{2\sqrt{2}v_{i,\tau}}\frac{\lambda_{i,\tau}^{(1)}-\lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}} \Rightarrow \mathcal{N}(0,1), \tag{4.10}$$

as $T, N \to \infty$, *where* $\theta_{i,\tau}$, $v_{i,\tau}$ *and* $\gamma_{i,\tau}$ *are defined in (4.4).*

Theorem 4.1 is a direct result of Lemma 4.1, since an asymptotic distribution of $\frac{\lambda_{i,\tau}^{(1)}-\lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}}$ can be derived using the independence between $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$. According to Theorem 4.1, under the null hypothesis, the test statistic $Z_{i,\tau}$ converges weakly to a standard normal random variable when $T, N \to \infty$.

On the other hand, under certain alternative hypotheses such as $r_1 = r_2$, $\gamma_{i,\tau}^{(1)} = \gamma_{i,\tau}^{(2)}$, $v_{i,\tau}^{(1)} = v_{i,\tau}^{(2)}$, but $\left(\sigma_i^{(1)}\right)^2 \neq \left(\sigma_i^{(2)}\right)^2$ and $\theta_{i,\tau}^{(1)} \neq \theta_{i,\tau}^{(2)}$, it can be shown in the next theorem that for some significant levels α, the power of the autocovariance test converges to 1 as $T, N \to \infty$.

Theorem 4.2. *Under the same assumptions of Lemma 4.1, if we assume two independent high-dimensional time series* $\left\{\boldsymbol{y}_t^{(1)}\right\}$ *and* $\left\{\boldsymbol{y}_t^{(2)}\right\}$ *follow different factor models in*

canonical form (4.8) with

$$r_1 = r_2 = r,\ \gamma_{i,\tau}^{(1)} = \gamma_{i,\tau}^{(2)} = \gamma_{i,\tau},\ v_{i,\tau}^{(1)} = v_{i,\tau}^{(2)} = v_{i,\tau},\ \textit{and}\ \theta_{i,\tau}^{(1)} = (1+c)\theta_{i,\tau}^{(2)}.$$

Then, for any c such that $\sqrt{T}\frac{2c}{2+c} \to \infty$ as $T, N \to \infty$ and $\lambda_{i,\tau}^{(1)} \neq \lambda_{i,\tau}^{(2)}$, it holds that

$$Pr\left(|Z_{i,\tau}| > z_\alpha | H_1\right) \to 1, \tag{4.11}$$

for $T, N \to \infty$, where z_α is the α-th quantile of the standard normal distribution.

Remark 4.2. The condition $\sqrt{T}\frac{2c}{2+c} \to \infty$ as $T, N \to \infty$ in Theorem 4.2 is relatively weak. It implies that for $T, N \to \infty$, the power of the test converges to 1 not only for a constant c, but also for some $c \to 0$ as long as $\sqrt{T}c \to \infty$. In other words, this test even works asymptotically for a local alternative hypothesis where the difference between $\theta_{i,\tau}^{(1)}$ and $\theta_{i,\tau}^{(2)}$ tends to 0, but slower than $1/\sqrt{T}$.

4.3 Implementation of testing procedure

In this section, the procedure of the autocovariance test is illustrated. For two high-dimensional time series, the test procedure can be summarised into four steps. Firstly, estimates of the factor models for both populations should be conducted, where the number of factors needs to be determined. Secondly, the original high-dimensional observations and the factor models' estimates need to be standardised to fulfil the canonical factor model (4.8). Thirdly, the quantities required to compute the test statistic $\widetilde{Z}_{i,\tau}$ should be estimated from both populations. Furthermore, we can compute the test statistic $\widetilde{Z}_{i,\tau}$ and its corresponding p-value for testing the equivalence of factor models. The details of the estimation and testing procedure are illustrated and discussed as follows.

Step 1: Estimates of the factor model:

For de-meaned high-dimensional time series observations $\left\{\boldsymbol{y}_t^{(k)}\right\}$, we first compute the symmetrized lag-τ sample autocovariance matrix $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)^\top$, where $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau) = \frac{1}{T-\tau-1}\sum_{t=1}^{T-\tau} \boldsymbol{y}_t^{(k)} \boldsymbol{y}_{t+\tau}^{(k)\top}$ is the lag-τ sample autocovariance

matrix of $\left\{\boldsymbol{y}_t^{(k)}\right\}$. By applying spectral (eigenvalue) decomposition on $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)^\top$, we can obtain an estimate of the factor loading matrix as $\widehat{\boldsymbol{Q}}_\tau^{(k)} = \left(\widehat{\boldsymbol{q}}_{1,\tau}^{(k)}, \widehat{\boldsymbol{q}}_{2,\tau}^{(k)}, ..., \widehat{\boldsymbol{q}}_{r_k,\tau}^{(k)}\right)$ with $\widehat{\boldsymbol{q}}_{i,\tau}^{(k)}$ the eigenvector of $\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)\widetilde{\boldsymbol{\Gamma}}_y^{(k)}(\tau)^\top$ associated with the i-th largest eigenvalue $\widehat{\lambda}_{i,\tau}^{(k)}$. To determine the number of factors, we adopt the idea in Lam et al. (2011), and use a ratio-based estimator $\widehat{r}_k = \text{argmin}_{1\leq j\leq R}\, \widehat{\lambda}_{j+1,\tau}^{(k)}/\widehat{\lambda}_{j,\tau}^{(k)}$ where $\widehat{\lambda}_{1,\tau}^{(k)} \geq \widehat{\lambda}_{2,\tau}^{(k)} \geq \cdots \geq \widehat{\lambda}_{N,\tau}^{(k)}$ and R is an integer satisfying $r_k \leq R < N$.

With $\widehat{\boldsymbol{Q}}_\tau^{(k)}$, the factors can then be estimated by $\widehat{\boldsymbol{f}}_t^{(k)} = \widehat{\boldsymbol{Q}}_\tau^{(k)\top}\boldsymbol{y}_t^{(k)}$ and the high-dimensional time series can be recovered by $\widehat{\boldsymbol{y}}_t^{(k)} = \widehat{\boldsymbol{Q}}_\tau^{(k)}\widehat{\boldsymbol{f}}_t^{(k)}$. Hence we have estimates of the factor model that is not in the canonical form (4.8) and the residuals are

$$\widehat{\boldsymbol{u}}_t^{(k)} = \boldsymbol{y}_t^{(k)} - \widehat{\boldsymbol{Q}}_\tau^{(k)}\widehat{\boldsymbol{f}}_t^{(k)}. \tag{4.12}$$

Moreover, to standardise the estimated factor model into the canonical form (4.8), we also need to estimate the variance of $u_{j,t}^{(k)}$ by $\left(\widehat{\sigma}_u^{(k)}\right)^2 = \frac{1}{NT-1}\sum_{j=1}^{N}\sum_{t=1}^{T}\left(\widehat{u}_{j,t}^{(k)} - \overline{\widehat{u}}_{j,t}^{(k)}\right)^2$.

Remark 4.3. It is clear that for two high-dimensional time series where the estimated numbers of factors are different, i.e., $\widehat{r}_1 \neq \widehat{r}_2$, one can conclude that the two high-dimensional data follow different factor models where $\mathcal{M}\left(\boldsymbol{Q}^{(1)}\right) \neq \mathcal{M}\left(\boldsymbol{Q}^{(2)}\right)$ and the numbers of spiked eigenvalues for their autocovariance matrices are different. However, if we are interested in testing the equivalence for particular spiked eigenvalue of the autocovariance matrices for two high-dimensional data, it is still possible to perform the autocovariance test even if $\widehat{r}_1 \neq \widehat{r}_2$. The intuition is to test whether the low-dimensional representations of both high-dimensional time series have the same variance in certain directions, though the data cannot be fully projected into the same eigenspace.

Step 2: Standardise the estimated factor model to satisfy the canonical form conditions:

With $\widehat{\boldsymbol{Q}}_\tau^{(k)}$ and $\left(\widehat{\sigma}_u^{(k)}\right)^2$, we can now standardise the estimated factor models (4.12) to fulfil the canonical form condition. Firstly, write $\boldsymbol{M}_\tau^{(k)} = \left(\widehat{\boldsymbol{Q}}_\tau^{(k)}, \boldsymbol{0}_{N-\widehat{r}_k}\right)$ for an $N \times N$ matrix. Then we can define $\widetilde{\boldsymbol{y}}_t^{(k)} := \boldsymbol{M}_\tau^{(k)\top} \boldsymbol{y}_t^{(k)} / \widehat{\sigma}_u^{(k)}$ for the normalised data and $\widetilde{\boldsymbol{u}}_t^{(k)} := \widehat{\boldsymbol{u}}_t^{(k)} / \widehat{\sigma}_u^{(k)}$ for the normalised residuals. By left multiplying $\widehat{\boldsymbol{Q}}_\tau^{(k)\top}$ and then dividing by $\widehat{\sigma}_u^{(k)}$, the estimated factor model is reduced to

$$\widetilde{\boldsymbol{y}}_t^{(k)} = \boldsymbol{M}_\tau^{(k)\top} \widehat{\boldsymbol{Q}}_\tau^{(k)} \widehat{\boldsymbol{f}}_t^{(k)} / \widehat{\sigma}_u^{(k)} + \widetilde{\boldsymbol{u}}_t^{(k)},$$

where note that

$$\boldsymbol{M}_\tau^{(k)\top} \widehat{\boldsymbol{Q}}_\tau^{(k)} = \begin{pmatrix} \boldsymbol{I}_{\widehat{r}_k} \\ \boldsymbol{0}_{N-\widehat{r}_k} \end{pmatrix}.$$

Secondly, to normalise $\widehat{\boldsymbol{f}}_t^{(k)}$, we can estimate the variances of $\widehat{\boldsymbol{f}}_t^{(k)}$ by $\left(\widehat{\sigma}_i^{(k)}\right)^2 = \frac{1}{T-1} \sum_{t=1}^{T} \left(\widehat{f}_{i,t}^{(k)} - \overline{\widehat{f}}_{i,t}^{(k)}\right)^2$, for $i = 1, 2, ..., \widehat{r}_k$.

In addition, write $\widetilde{\boldsymbol{f}}_t^{(k)} = \widehat{\boldsymbol{f}}_t^{(k)} / \left(\widehat{\sigma}_u^{(k)} \widehat{\sigma}_i^{(k)}\right)$ for the normalised estimates of factors, and

$$\widetilde{\boldsymbol{Q}}_\tau^{(k)} = \begin{pmatrix} \widehat{\sigma}_1^{(k)} & & \\ & \ddots & \\ & & \widehat{\sigma}_{\widehat{r}_k}^{(k)} \\ & \boldsymbol{0}_{N-\widehat{r}_k} & \end{pmatrix},$$

for the $N \times \widehat{r}_k$ estimated loading matrix. Then we have standardised the estimated factor model to

$$\widetilde{\boldsymbol{y}}_t^{(k)} = \widetilde{\boldsymbol{Q}}_\tau^{(k)} \widetilde{\boldsymbol{f}}_t^{(k)} + \widetilde{\boldsymbol{u}}_t^{(k)}, \tag{4.13}$$

which follows the canonical form defined by (4.8).

Step 3: Estimate unknown parameters in the test statistic:

For standardised data $\left\{\widetilde{\boldsymbol{y}}_t^{(k)}\right\}$ following the estimated factor model (4.13), $\lambda_{i,\tau}^{(k)}$ can be computed as the i-th largest eigenvalue of the symmetrized lag-τ sample autocovariance matrix $\widetilde{\boldsymbol{\Gamma}}_{\widetilde{\boldsymbol{y}}}^{(k)}(\tau)\widetilde{\boldsymbol{\Gamma}}_{\widetilde{\boldsymbol{y}}}^{(k)}(\tau)^\top$, where the sample autocovariance matrix is given by $\widetilde{\boldsymbol{\Gamma}}_{\widetilde{\boldsymbol{y}}}^{(k)}(\tau) = \frac{1}{T-\tau-1}\sum_{t=1}^{T-\tau}\widetilde{\boldsymbol{y}}_t^{(k)}\widetilde{\boldsymbol{y}}_{t+\tau}^{(k)\top}$ and $\gamma_{i,\tau}^{(k)}$ can be estimated from the sample lag-τ autocovariance of the i-th estimated factor $\left\{\widetilde{f}_{i,t}^{(k)}\right\}$. Besides, we also need to estimate the quantities $v_{i,\tau}^{(k)}$ and $\theta_{i,\tau}^{(k)}$, as defined in Test 4.1, for each sample to compute the test statistic. However, since $\left(v_{i,\tau}^{(k)}\right)^2 = \frac{1}{T-\tau}Var\left(\sum_{t=1}^{T-\tau}f_{i,t}^{(k)}f_{i,t+\tau}^{(k)}\right)$ depends on the variance of $\sum_{t=1}^{T-\tau}f_{i,t}^{(k)}f_{i,t+\tau}^{(k)}$ and $\theta_{i,\tau}^{(k)}$ is the asymptotic centring of $\lambda_{i,\tau}^{(k)}$, they cannot be directly estimated from sample observations. Instead, we can use bootstrap to estimate both quantities. It is worth noting that since the bootstrap is conducted on the estimated low-dimensional factor time series $\left\{\widetilde{\boldsymbol{f}}_t^{(k)}\right\}$, the bootstrap estimators are not affected by the increasing dimensions.

Therefore, the sieve bootstrap proposed in Chapter 2 can be utilised for estimating $v_{i,\tau}^{(k)}$ and $\theta_{i,\tau}^{(k)}$. In specific, an AR(p) model can be fitted for each estimated factor $\widetilde{\boldsymbol{f}}_i^{(k)}$ and the residuals can be taken as

$$\widetilde{\epsilon}_{i,t}^{(k)} = \widetilde{f}_{i,t}^{(k)} - \sum_{l=1}^{p}\widetilde{\psi}_{i,l}^{(k)}\widetilde{f}_{i,t-l}^{(k)},$$

where $\left\{\widetilde{\psi}_{i,l}^{(k)},\ l=1,2,...,p\right\}$ are the AR coefficients. Then by resampling from the empirical distribution of centralised residuals $\left(\widetilde{\epsilon}_{i,t}^{(k)} - \bar{\widetilde{\epsilon}}_i^{(k)}\right)$, the bootstrap factors can be generated as

$$f_{i,t}^{(k)b} = \sum_{l=1}^{p}\widetilde{\psi}_{i,l}^{(k)}f_{i,t-l}^{(k)b} + \epsilon_{i,t}^{(k)b},$$

where $\epsilon_{i,t}^{(k)b}$ is the bootstrap residual. Hence, we can estimate $v_{i,\tau}^{(k)}$ by

$$\widetilde{v}_{i,\tau}^{(k)*} = \sqrt{\frac{1}{T-\tau}\left(\frac{1}{B-1}\sum_{b=1}^{B}\left(\sum_{t=1}^{T-\tau}f_{i,t}^{(k)b}f_{i,t+\tau}^{(k)b} - \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{t=1}^{T-\tau}f_{i,t}^{(k)b}f_{i,t+\tau}^{(k)b}\right)\right)^2\right)},$$

where $b = 1, 2, ..., B$ for B bootstrap samples of $\left\{f_{i,t}^{(k)*}\right\}$. In addition, since $\widetilde{\theta}_{i,\tau}^{(k)}$ is an estimate of the asymptotic centring of $\lambda_{i,\tau}^{(k)}$, we can bootstrap $\left\{\widetilde{\boldsymbol{y}}_t^{(k)}\right\}$ by

$$\boldsymbol{y}_t^{(k)b} = \widetilde{\boldsymbol{Q}}_\tau^{(k)} \boldsymbol{f}_t^{(k)b},$$

for B times and estimate $\theta_{i,\tau}^{(k)}$ by

$$\widetilde{\theta}_{i,\tau}^{(k)*} = \frac{1}{B}\sum_{b=1}^{B} \lambda_{i,\tau}^{(k)b},$$

where $\lambda_{i,\tau}^{(k)b}$ is the i-th largest eigenvalue of the symmetrized lag-τ sample autocovariance matrices of $\left\{\boldsymbol{y}_t^{(k)b}\right\}$. In the meantime, since bootstrap is conducted to estimate $v_{i,\tau}^{(k)}$ and $\theta_{i,\tau}^{(k)}$, an alternative estimate of $\gamma_{i,\tau}^{(k)}$ can also be computed based on B bootstrap samples, as

$$\widetilde{\gamma}_{i,\tau}^{(k)*} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{T-\tau-1}\sum_{t=1}^{T-\tau}\left(f_{i,1}^{(k)b} - \frac{1}{T}\sum_{t=1}^{T} f_{i,t}^{(k)b}\right)\left(f_{i,\tau+1}^{(k)b} - \frac{1}{T}\sum_{t=1}^{T} f_{i,t}^{(k)b}\right)\right).$$

Step 4: Compute the test statistic and p-value:

When the first three steps have been conducted on both high-dimensional times series $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$, we can estimate the unknown parameters in (4.3) by

$$\widetilde{\theta}_{i,\tau}^{*} := \frac{T_1\widetilde{\theta}_{i,\tau}^{(1)*} + T_2\widetilde{\theta}_{i,\tau}^{(2)*}}{T_1+T_2},\ \widetilde{v}_{i,\tau}^{*} := \frac{T_1\widetilde{v}_{i,\tau}^{(1)*} + T_2\widetilde{v}_{i,\tau}^{(2)*}}{T_1+T_2},\ \widetilde{\gamma}_{i,\tau}^{*} := \frac{T_1\widetilde{\gamma}_{i,\tau}^{(1)*} + T_2\widetilde{\gamma}_{i,\tau}^{(2)*}}{T_1+T_2},$$

where $\widetilde{\theta}_{i,\tau}^{(k)*}$, $\widetilde{v}_{i,\tau}^{(k)*}$ and $\widetilde{\gamma}_{i,\tau}^{(k)*}$ are computed from two high-dimensional times following the procedure in Step 4.3. Then, the test statistic can be computed as

$$\widetilde{Z}_{i,\tau} := \left(\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}\right)\sqrt{\frac{T_1T_2}{T_1+T_2}}\frac{\widetilde{\gamma}_{i,\tau}^{*}}{2\widetilde{v}_{i,\tau}^{*}\widetilde{\theta}_{i,\tau}^{*}}, \tag{4.14}$$

where $\lambda_{i,\tau}^{(1)}$ and $\lambda_{i,\tau}^{(2)}$ are the i-th ($1 \leq i \leq \widehat{r}_k$) largest eigenvalues of the symmetrized lag-τ sample autocovariance matrix for the standardised data $\left\{\widetilde{\boldsymbol{y}}_t^{(1)}\right\}$ and $\left\{\widetilde{\boldsymbol{y}}_t^{(2)}\right\}$, respectively. Lastly, the p-value of this test statistic $\widetilde{Z}_{i,\tau}$ can be computed as $\Pr\left(z > \left|\widetilde{Z}_{i,\tau}\right|\right) = 2\left(1 - \Phi\left(\left|\widetilde{Z}_{i,\tau}\right|\right)\right)$ for a two-sided test, and $\Pr\left(z > \widetilde{Z}_{i,\tau}\right) = 1 - \Phi\left(\widetilde{Z}_{i,\tau}\right)$ or $\Pr\left(z < \widetilde{Z}_{i,\tau}\right) = \Phi\left(\widetilde{Z}_{i,\tau}\right)$ for one-sided tests, where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of a standard normal random variable.

4.4 Simulation studies

This section uses numerical simulations to investigate the proposed autocovariance test's empirical sizes and powers in various scenarios.

To start, we first of all explore the empirical sizes of the autocovariance test for various orders of factor strengths and ratios between the sample size T and the data dimension N. We assume the high-dimensional observations $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ are generated from the one-factor model $\boldsymbol{y}_t^{(k)} = \boldsymbol{Q}^{(k)}\boldsymbol{f}_t^{(k)} + \boldsymbol{u}_t^{(k)}$ in the canonical form (4.8). Moreover, we assume the factor $\left\{f_{1,t}^{(k)}\right\}$ follow an AR(1) model with mean zero, AR coefficient $\phi_1^{(k)} = 0.5$ and variance normalised to one. In other words, the factors for both time series are generated by

$$f_{1,t}^{(k)} = \phi_1^{(k)} f_{1,t-1}^{(k)} + z_{1,t}^{(k)},\ k = 1, 2, \tag{4.15}$$

where $\phi_1^{(k)} = 0.5$ and $\left\{z_{1,t}^{(k)}\right\}$ are i.i.d. $\mathcal{N}\left(0, \left(\sigma_z^{(k)}\right)^2\right)$ with the variance given by $\left(\sigma_z^{(k)}\right)^2 = 1/\left(1 - \left(\phi_1^{(k)}\right)^2\right) = 3/4$, so that $Var\left(f_{1,t}^{(k)}\right) = 1$. As discussed for the canonical form condition of factor models, the variances $\left\{\left(\sigma_i^{(k)}\right)^2\right\}$ of normalised factors are contained in the loading matrix $\boldsymbol{Q}^{(k)}$. To study the empirical sizes of the autocovariance test under various factor strengths, we firstly refer to Lam et al. (2011) for the definition of factor strength. Lam et al. (2011) define the factor strength through the relationship between the orders of variances and the data dimension N, which is $\left(\sigma_1^{(k)}\right)^2 \asymp N^{1-\delta}$ for $\delta \in [0, 1)$.

Using this definition, $\delta = 0$ refers to the strongest factors with the pervasiveness, and factor strengths drop when δ increases from 0 to 1. In this section, we consider four different cases for factor strengths, where $\delta = 0, 0.1, 0.3$, and 0.5. Specifically, $\left(\sigma_1^{(k)}\right)^2$ in the loading matrix $\boldsymbol{Q}^{(k)}$ that follows canonical form (4.7) is assumed to be $N, N^{0.9}, N^{0.7}$, and $N^{0.5}$, respectively, and $\left\{u_{j,t}^{(k)}\right\}$ are assumed to be i.i.d. $\mathcal{N}(0,1)$. In summary, both N-dimensional time series observations are generated by

$$\boldsymbol{y}_t^{(k)} = \begin{pmatrix} \sigma_1^{(k)} \\ \boldsymbol{0}_{N-1} \end{pmatrix} f_{1,t}^{(k)} + \boldsymbol{u}_t^{(k)}, \tag{4.16}$$

where $\sigma_1^{(k)} = N^{1-\delta}$, $\{u_{j,t}\}$ are i.i.d. $\mathcal{N}(0,1)$, and $\left\{f_{1,t}^{(k)}\right\}$ are generated by (4.15).

In addition to factor strengths, to explore the impact of ratios between sample size T and data dimension N, we generate data with $T = 400, 800$ and $N = 100, 200, 400, 800, 1600$. To compute the empirical sizes, for each combination of T, N and δ, two high-dimensional time series observations are first of all generated as $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$. Then, by utilizing the estimation and testing procedure in Section 4.3, the test statistic $\widetilde{Z}_{i,\tau}$ can be computed by (4.14) where $B = 500$ bootstrap samples are generated to find $\widetilde{\theta}_{i,\tau}^{(k)*}$, $\widetilde{v}_{i,\tau}^{(k)*}$ and $\widetilde{\gamma}_{i,\tau}^{(k)*}$, at the numbers of factors are assumed to be known (i.e., $\widetilde{r}_k = 1$) for both samples. The empirical sizes of a one-sided autocovariance test for $i = 1$, $\tau = 1$, and significant level $\alpha = 0.1$ under various combinations of T, N and δ are computed as the averages of empirical probabilities that $\widetilde{Z}_{1,1}$ is less than z_α or greater than $z_{1-\alpha}$, i.e.,

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\{\widetilde{Z}_{1,1}(m) < z_\alpha\}}, \text{ or } \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\{\widetilde{Z}_{1,1}(m) > z_{1-\alpha}\}},$$

for $M = 500$ Monte Carlo simulations, where $\widetilde{Z}_{1,1}(m)$ is the test statistic computed from the m-th simulation.

As presented in Figure 4.1, despite some minor fluctuations, the empirical sizes of the autocovariance test are close to the nominal significant level $\alpha = 0.1$

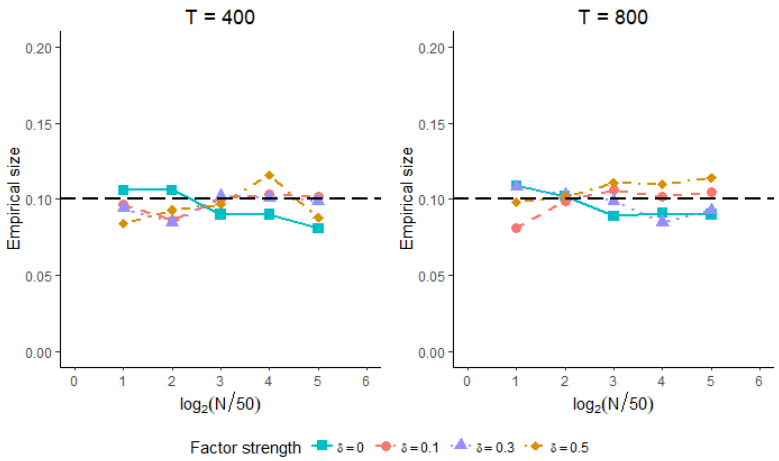


Figure 4.1: Empirical sizes of the autocovariance test in the first scenario with $T = 400, 800$, $N = 100, 200, 400, 800, 1600$, and $\delta = 0, 0.1, 0.3, 0.5$.

for all choices of N, T and δ. That is, when the numbers of factors are known or can be correctly estimated, the nominal type-I errors of the autocovariance test can be verified via empirical simulation studies for $\delta = 0, 0.1, 0.3, 0.5$, $T = 400, 800$, and $N = 100, 200, 400, 800, 1600$. The choice of $\tau = 1$ for the autocovariance test is to acquire the most information on temporal dependence of $\left\{ \boldsymbol{y}_t^{(k)} \right\}$ and to achieve the best accuracy on corresponding estimators $\widetilde{\theta}_{i,\tau}^{(k)*}$, $\widetilde{v}_{i,\tau}^{(k)*}$ and $\widetilde{\gamma}_{i,\tau}^{(k)*}$, while other choices of finite τ may be considered with cautions as $\gamma_{i,\tau}^{(k)}$ tends to 0 when τ increases.

For two high-dimensional time series following factor models that are normalised to the canonical form (4.13), the difference between spiked eigenvalues of the symmetrized lag-τ autocovariance matrix for two time series may come from the difference between variances or auto-correlations of factors in different factor models. Therefore, to empirically investigate the autocovariance test's power, we study two typical scenarios where either variances or auto-correlations of factors are different between two factor models. We are particularly interested in whether the autocovariance test's empirical power grows with the difference between variances or auto-correlations for two high-dimensional time series.

Specifically, to explore the impacts of δ, N and T on empirical powers, we again generate observations from two populations with $T = 400, 800$, $N = 200, 400, 800$, and $\delta = 0, 0.1, 0.3, 0.5$. The data in the first population is generated

by (4.16), which is precisely the same as we study the empirical sizes, while the data in the second population is generated with a different $\sigma_1^{(2)}$ or $\phi_1^{(2)}$ in the factor model. In the first scenario, we examine the impact of difference between $\sigma_1^{(1)}$ and $\sigma_1^{(2)}$ on empirical powers of the autocovariance test, therefore, keep the AR coefficient $\phi_1^{(2)}$ the same as $\phi_1^{(1)}$ (i.e., $\phi_1^{(2)} = \phi_1^{(1)} = 0.5$), and set $\left(\sigma_1^{(2)}\right)^2$ as $1.1\left(\sigma_1^{(1)}\right)^2, 1.3\left(\sigma_1^{(1)}\right)^2, 1.5\left(\sigma_1^{(1)}\right)^2, 1.7\left(\sigma_1^{(1)}\right)^2$, and $1.9\left(\sigma_1^{(1)}\right)^2$, respectively. On the other hand, in the second scenario, we keep $\left(\sigma_1^{(2)}\right)^2 = \left(\sigma_1^{(1)}\right)^2 = N^{1-\delta}$, but set the AR coefficients for the second population to be $\phi_1^{(2)} = 0.9\phi_1^{(1)}, 0.8\phi_1^{(1)}, 0.7\phi_1^{(1)}, 0.6\phi_1^{(1)}$, and $0.5\phi_1^{(1)}$, respectively. By doing that, we can investigate how the empirical powers of the autocovariance test are affected by the difference between auto-correlations of factors in two factor models. Moreover, when generating $\left\{f_{i,t}^{(2)}\right\}$, it is worth to mention that $\left\{z_{1,t}^{(2)}\right\}$ are i.i.d. $\mathcal{N}\left(0, \left(\sigma_z^{(2)}\right)^2\right)$ with $\left(\sigma_z^{(2)}\right) = 1/\left(1 - \phi_1^{(2)2}\right)$.

To compute the empirical powers, for each combination of T, N and δ, two high-dimensional time series observations are generated as $\left\{\boldsymbol{y}_t^{(1)}\right\}$ and $\left\{\boldsymbol{y}_t^{(2)}\right\}$ first. Then, we can follow the estimation and testing procedure in Section 4.3 and compute the test statistic $\widetilde{Z}_{i,\tau}$ by (4.14), where again $B = 500$ bootstrap samples are generated to find $\widetilde{\theta}_{i,\tau}^{(k)*}$, $\widetilde{v}_{i,\tau}^{(k)*}$, and $\widetilde{\gamma}_{i,\tau}^{(k)*}$ for both samples with the number of factors assumed to be known (i.e., $\widetilde{r}_k = 1$). Lastly, based on $M = 500$ Monte Carlo simulations, the empirical powers of a one-sided autocovariance test for $i = 1$, $\tau = 1$, and $\alpha = 0.1$ can be estimated by the empirical probability that $\widetilde{Z}_{1,1}$ is less than z_α, i.e.,

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\left\{\widetilde{Z}_{1,1}(m) < z_\alpha\right\}},$$

for the first scenario, and the probability that $\widetilde{Z}_{1,1}$ is greater than $z_{1-\alpha}$, i.e.,

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\left\{\widetilde{Z}_{1,1}(m) > z_{1-\alpha}\right\}},$$

for the second scenario, where we have assumed $\mu_{1,1}^{(1)} < \mu_{1,1}^{(2)}$ for various choices

of $\left(\sigma_1^{(2)}\right)^2$ in the first scenario, and $\mu_{1,1}^{(1)} > \mu_{1,1}^{(2)}$ for various choices of $\phi_1^{(2)}$ in the second scenario.

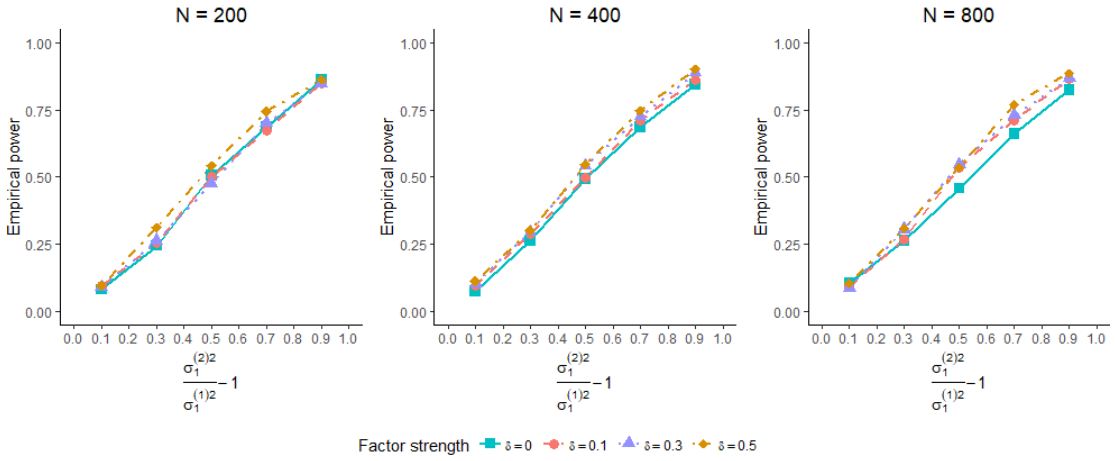


Figure 4.2: Empirical powers of the autocovariance test in the first scenario with $T = 400$, $N = 200, 400, 800$, and $\delta = 0, 0.1, 0.3, 0.5$.

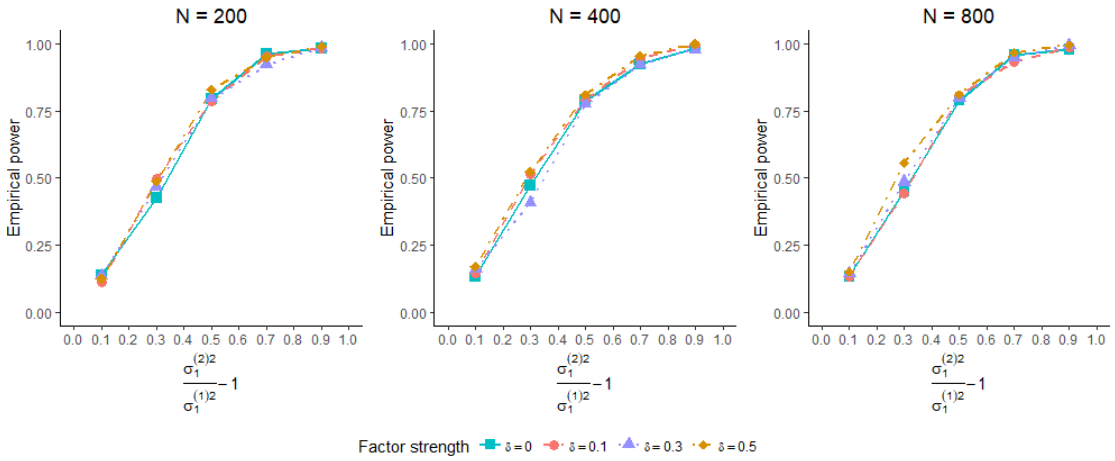


Figure 4.3: Empirical powers of the autocovariance test in the first scenario with $T = 800$, $N = 200, 400, 800$, and $\delta = 0, 0.1, 0.3, 0.5$.

Empirical powers of the autocovariance test in both scenarios with various choices of N, T, and δ are presented in Figures 4.2 to 4.5. As shown in Figures 4.2 and 4.3, it is clear that for all combinations of N and T, empirical powers in the first scenario increase towards 1 when $\left(\sigma_1^{(2)}\right)^2$ increases from $1.1\left(\sigma_1^{(1)}\right)^2$ to $1.9\left(\sigma_1^{(1)}\right)^2$. Therefore, numerical results in Figure 4.2 and 4.3 suggest that the autocovariance test can correctly reject the null hypothesis when two high-dimensional time series follow different factor models with $\left(\sigma_1^{(2)}\right)^2 \neq \left(\sigma_1^{(1)}\right)^2$.

Besides, despite the common trend, for the same difference between $\left(\sigma_1^{(2)}\right)^2$ and $\left(\sigma_1^{(1)}\right)^2$, the empirical powers of one-sided autocovariance tests for $T = 800$ are generally higher than those associated with $T = 400$, which can be justified by $\sqrt{T}$ in (4.3). Also, the powers of stronger factor models with smaller δ are slightly higher than those of weaker factor models with larger δ, especially for $T = 400$.

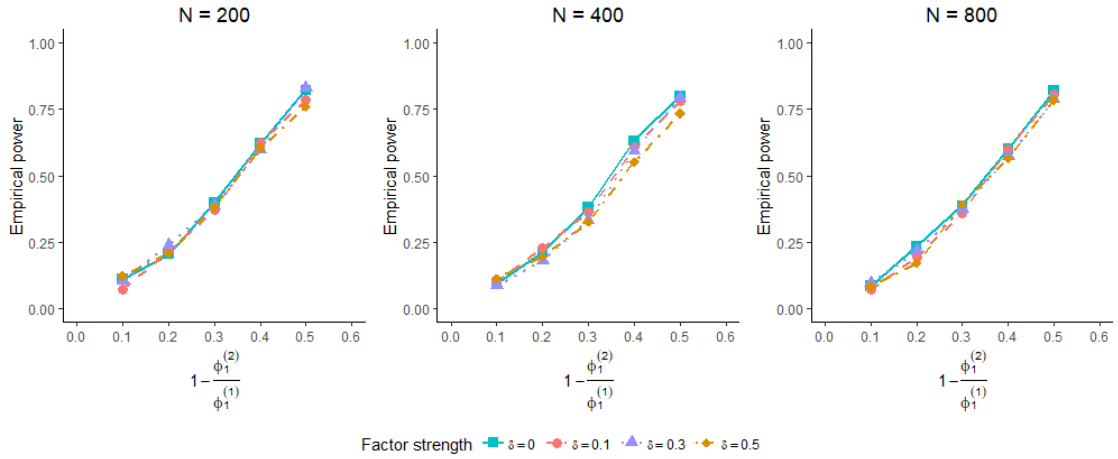


Figure 4.4: Empirical powers of the autocovariance test in the second scenario with $T = 400$, $N = 200, 400, 800$, and $\delta = 0, 0.1, 0.3, 0.5$.

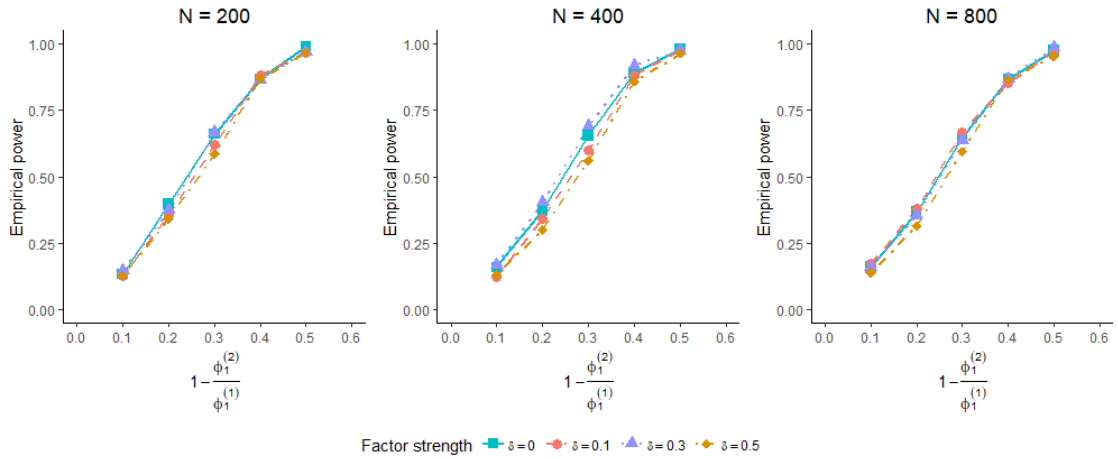


Figure 4.5: Empirical powers of the autocovariance test in the second scenario with $T = 800$, $N = 200, 400, 800$, and $\delta = 0, 0.1, 0.3, 0.5$.

Similarly, as presented in Figure 4.4 and 4.5, for all ratios of N and T, empirical powers in the second scenario also increase towards 1, while $\phi_1^{(2)}$ drops from $\phi_1^{(1)} = 0.5$. As a consequence, Figure 4.4 and 4.5 suggest that the autocovariance

test can correctly reject the null hypothesis when two high-dimensional time series follow different factor models with $\phi_1^{(2)} \neq \phi_1^{(1)}$. However, unlike the first scenario, empirical powers of the one-sided the autocovariance test for relatively weak factor models with large δ, especially $\delta = 0.5$, are slightly lower than those of relatively strong factor models with small δ. In other words, compared with strong factor models, the autocovariance test for weak factor models is slightly less potent in detecting the same proportional changes in auto-correlations of factors for two different factor models.

4.5 Mortality data for multiple countries

To apply the proposed autocovariance test in real-world data, we study age-specific mortality rates for countries worldwide and test whether the mortality rates for different countries share the same factor model with the same spiked eigenvalues of their autocovariance matrices. In the past century, age-specific mortality rates have received massive attention, especially by insurance companies and governments, as accurate forecasting mortality rates are crucial for the pricing of life insurance products and are highly related to social and economic policies. Among many works on forecasting age-specific mortality rates, the Lee-Carter model (Lee and Carter, 1992) is prevalent and has been used globally. Despite some extensions on the original model (see, e.g., Hyndman and Shahid Ullah, 2007; Li et al., 2013), one drawback of the Lee-Carter model is that it only focuses on the death rates of a single country, therefore may produce quite different long-run forecasts of mortality rates for different countries. Recently, joint modelling of mortality rates for multiple countries has become more attractive since the common features extracted for multiple populations can further improve forecasting accuracy. In this sense, correctly classifying countries with similar patterns of mortality rates into the same group for joint modelling and combined statistical analysis becomes critical. In addition to the traditional grouping methods based on socioeconomic status or ethnic group, Tang et al. (2020) emphasise the use of statistical clustering methods on determining the

grouping of countries.

This section uses the proposed autocovariance test to explore whether multiple countries' mortality data have the same spiked eigenvalues of the autocovariance matrices. To achieve this, we collect the total death rates for various countries from the Human Mortality Database (University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), 2018). For the best quality of data, we choose the death rates from age 0 to 90 and require each country's sample size to be relatively large. For some countries such as the Republic of Korea and Chile, the data are only available for a short period, while for some other countries, the data quality cannot be guaranteed due to some historical reasons. As a result, we only study selected countries with total death rates available from 1957 to 2017. Besides, as seen in the first graph of Figure 4.6, the age-specific mortality rates are not stationary for most ages; therefore, they have been pre-processed by taking the logarithm and then differenced, since our method is developed for stationary time series. For countries such as Norway and Iceland, there are quite a few zero death rates for young children due to the relatively small population; hence they are excluded from our study. For all the other countries, zero death rates are replaced by the averages of death rates in adjacent years. In summary, the data we study has dimension $N = 91$ and sample size $T = 60$ for each country. The plots of log mortality rates for Australia are shown in Figure 4.6 as an example.

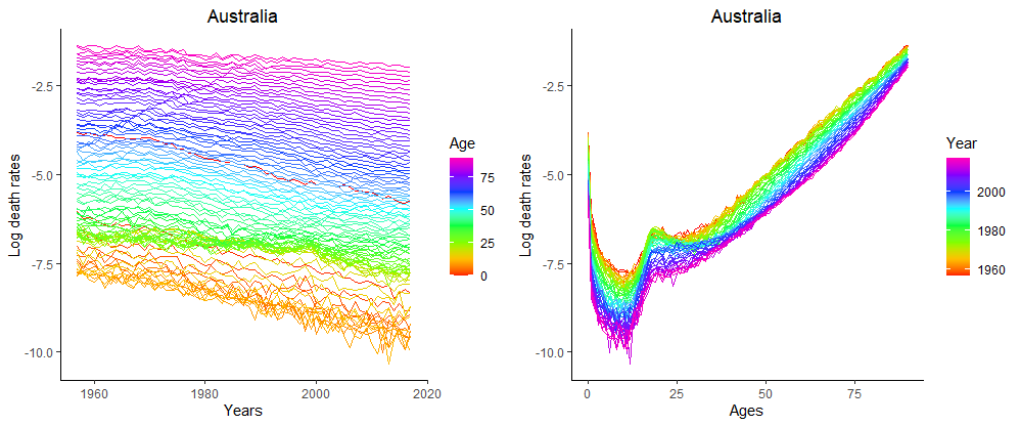


Figure 4.6: Observed time series of log death rates in Australia

According to the estimation and testing procedure in Section 4.3, factor models in canonical form 4.8 are firstly estimated and normalised from the differenced log death rates for each country. In the meantime, the number of factors in the factor model for each country is estimated and compared. As shown in Table 4.1, for most countries, there is only one factor estimated from the differenced log death rates, while there are some exceptions where two, three, and five factors are estimated. For countries with the same number of factors, we can compute the test statistic $\widetilde{Z}_{i,\tau}$ to test the equivalence of factors. For the best accuracy in estimating the number of factors and temporal dependence among death rates, the autocovariance test is performed based on $\tau = 1$ throughout this section.

Table 4.1: Estimated number of factors in the factor model for each country

Estimated number of factors	Countries
1	Australia, Belgium, Bulgaria, Czechia, Finland, Greece, Hungary, Japan, Netherlands, Sweden, Switzerland, U.K., U.S.A.
2	Denmark
3	Canada, France, Italy, Portugal
5	Poland

For countries with one factor in their estimated factor models, the test statistic $\widetilde{Z}_{1,1}$ for each pair of countries can be computed. Meanwhile, the signs of $\widetilde{\gamma}_{1,1}^{(k)}$ are checked for all countries where it is positive for the U.S.A. but negative for all the other countries. Consequently, the factor model for the U.S.A. should be considered different from the rest countries with one factor. For all other countries with one factor, the p-values associated with all test statistics are computed. As illustrated in Figure 4.7, the factor model and the spiked eigenvalues of the autocovariance matrices in the majority of European countries are similar as most p-values of test statistics between two European countries are greater than 0.1. However, the p-values between Finland and Bulgaria, the U.K. and Bulgaria, Finland and Switzerland, the U.K. and Switzerland are relatively small. As a result, extra caution needs to be taken when these countries are included in a combined statistical analysis. Besides, some p-values associated with either

Japan or Australia are also relatively small, which can be considered statistical evidence of geographic impacts.

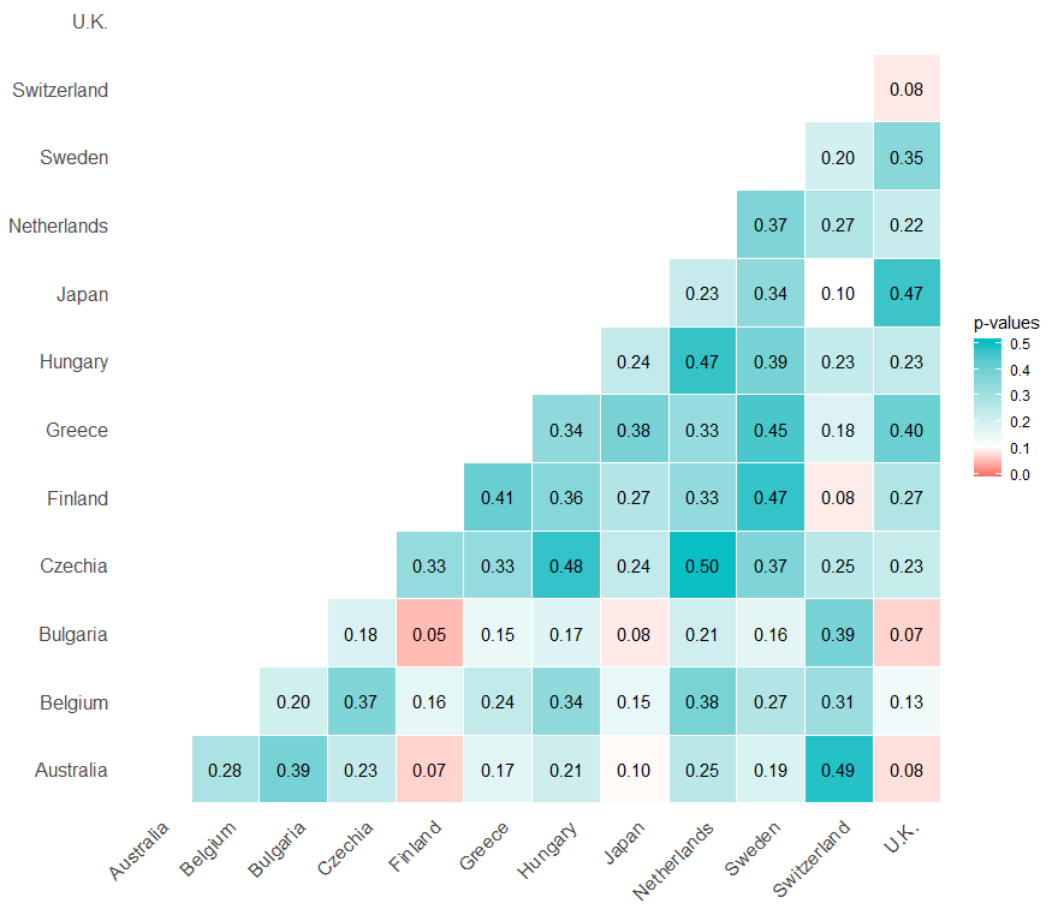


Figure 4.7: p-values of the autocovariance test for each pair of countries that have one factor in the estimated factor model

For countries with three factors, to test on the equivalence of autocovariance through factor models, test statistics between each pair of countries are computed for all three factors as $\widetilde{Z}_{1,1}$, $\widetilde{Z}_{2,1}$ and $\widetilde{Z}_{3,1}$. As depicted in Figure 4.8, the p-values for $\widetilde{Z}_{1,1}$ and $\widetilde{Z}_{2,1}$ between all pairs of countries are relatively large, which suggests that the differences of the first two factors between each pair of countries are not significant (at $\alpha = 0.1$). Nonetheless, p-values for $\widetilde{Z}_{1,3}$ are relatively small between Canada and France, Canada and Italy, and very small between Italy and Portugal. As a result, despite that p-value is 0.09 for $\widetilde{Z}_{1,3}$ between Italy and Portugal, one may suggest considering France, Italy, and Portugal have similar

spiked eigenvalues of their autocovariance matrices in a three-factor model and include them in a combined statistical analysis while leaving Canada for an independent analysis.

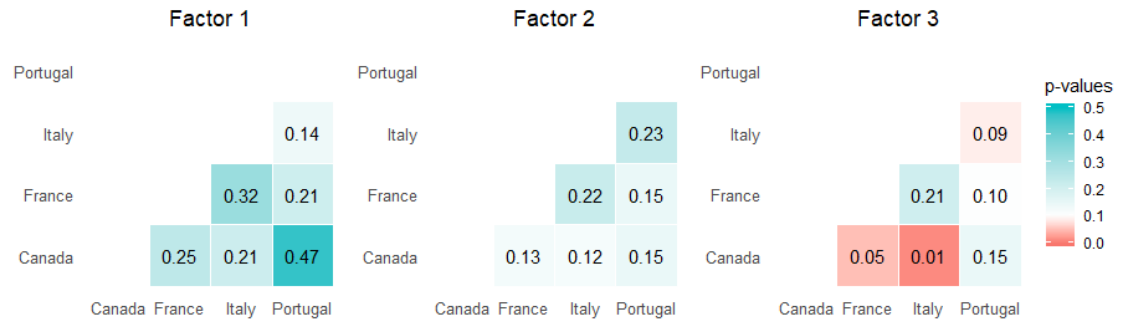


Figure 4.8: p-values of the autocovariance test for each pair of countries that have three factors in the estimated factor model

In practice, as discussed in Remark 4.3, regardless of whether the estimated number of factors of total mortality rates for multiple countries are the same, it is still of interest to test whether the mortality rates for multiple countries have the same low-dimensional representations in the eigenspace spanned by the first eigenvector that is shared by all countries. For this purpose, we perform autocovariance tests on the first factor for all countries except the U.S.A., and the results are illustrated in Figure 4.9. It is then straightforward that, in addition to what has been discussed for those countries with only one factor in their factor models, the first factor of Australia, Bulgaria, and Switzerland also differs from Italy and Poland's first factor, respectively. Consequently, despite the differences between the estimated numbers of factors for Denmark, Canada, France, Italy, Portugal, Poland, and all other countries, the total death rates projected in the eigenspace spanned by the first common eigenvector are not significantly different across these countries.

From the perspective of combined statistical analysis on age-specific mortality rates, p-values of the autocovariance test for each pair of countries in Figure 4.9 can also be considered a measure of dissimilarities between the age-specific death rates for two countries. A relatively small p-value suggests that these

two countries' age-specific death rates are somewhat different. Consequently, a hierarchical clustering method can be developed with the dissimilarities measured by the p-values of the autocovariance test, providing clustering results for combined statistical analysis on age-specific mortality rates.

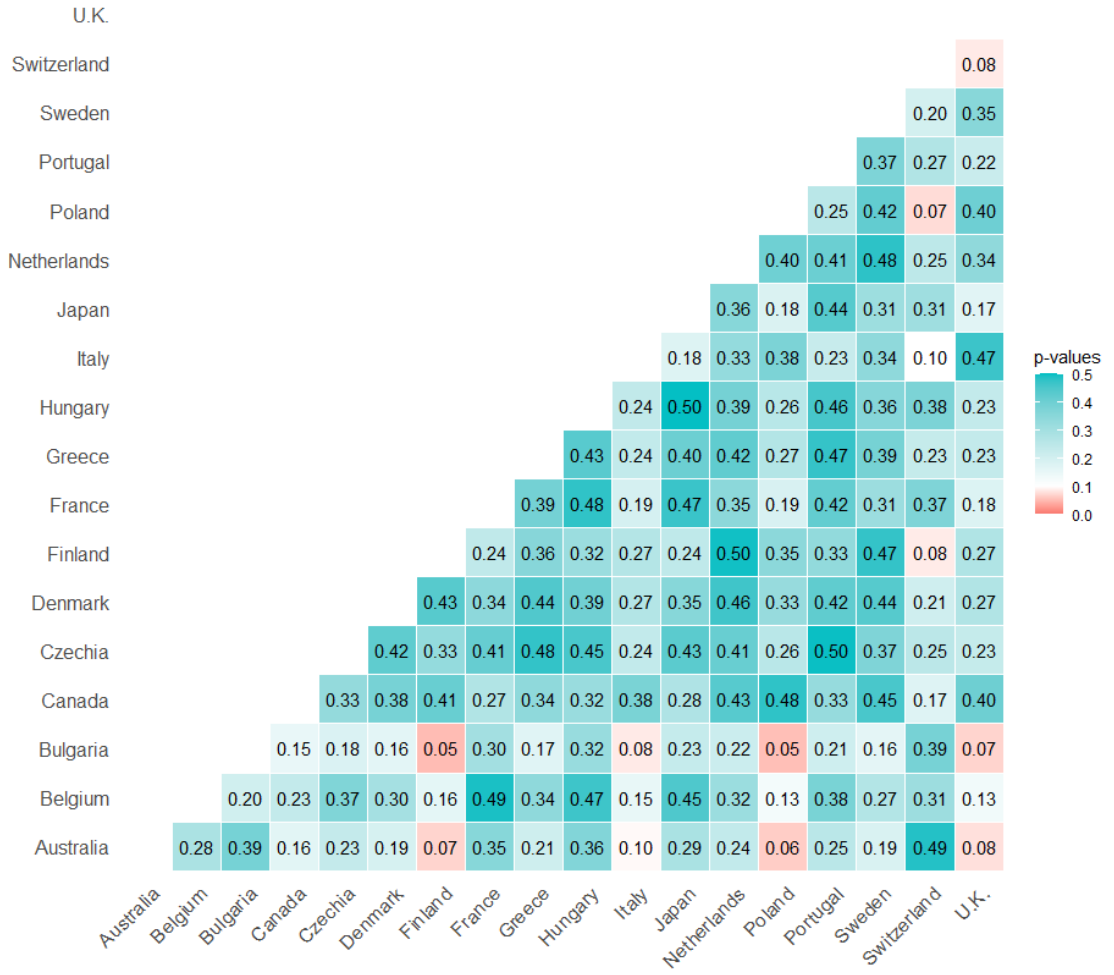


Figure 4.9: p-values of the autocovariance test of the first factor for all countries except U.S.A.

4.6 Conclusions and discussions

Hypothesis testing for the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series is becoming critical since multiple data-sets may be aggregated together for better inference, including improving estimation and forecasting accuracy. In this work, we propose a novel autocovariance test

for comparing the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series. The test statistic is created using a central limit theorem (CLT) of spiked eigenvalues extracted from the symmetrized lag-τ sample autocovariance matrix of high-dimensional time series. Therefore, this novel autocovariance test takes both cross-sectional and temporal dependence of high-dimensional time series into consideration. The proposed autocovariance test is efficient in telling the difference between the spiked eigenvalues of the autocovariance matrices for two high-dimensional time series, even for a local alternative hypothesis, as the p-value of the test statistic increases towards one with sample size T. Simulation studies provide numerical evidence on the autocovariance test's finite-sample performance for high-dimensional time series under various strengths of factors and ratios between data dimensions and sample sizes. Finally, we apply our method to age-specific mortality rates for multiple countries and test whether different countries share the same spiked eigenvalues of the autocovariance matrices. This work is preliminary for combined analysis of human mortality rates for multiple countries since it can provide satistical inference on clustering and grouping countries worldwide. Finally, our work forms an essential part of the statistical inferences for high-dimensional time series and motivates the combined and aggregated analysis for multiple high-dimensional time series.

4.A Appendix A: Technical proof of Theorem 4.2

Proof of Theorem 4.2. Without loss of generality, we only consider the case for $Z_{i,\tau} > 0$ since the case for $Z_{i,\tau} < 0$ can be considered in precisely the same way. For a constant significant level α, to see $Pr(Z_{i,\tau} > z_\alpha | H_1) \to 1$ as $T, N \to \infty$, it is sufficient to show that $Z_{i,\tau} \to \infty$ as $T, N \to \infty$.

To start, we firstly notice that for any $i \in \{1, 2, ..., r\}$ and a finite time lag τ, $\frac{\gamma_{i,\tau}}{2\sqrt{2}v_{i,\tau}}$ does not divergent with N and T, since both $\gamma_{i,\tau}$ and $v_{i,\tau}$ are some constants when $T, N \to \infty$. It then suffices to show $\sqrt{T}\frac{\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}} \to \infty$ when

$T, N \to \infty$. Note that by the definition of $\theta_{i,\tau}$ in (4.4), we can show that

$$\frac{\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}} = \frac{\lambda_{i,\tau}^{(1)}}{\theta_{i,\tau}^{(1)}}\frac{\theta_{i,\tau}^{(1)}}{\theta_{i,\tau}} - \frac{\lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}^{(2)}}\frac{\theta_{i,\tau}^{(2)}}{\theta_{i,\tau}} = \frac{\lambda_{i,\tau}^{(1)}}{\theta_{i,\tau}^{(1)}}\frac{2+2c}{2+c} - \frac{\lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}^{(2)}}\frac{2}{2+c}, \tag{4.17}$$

where the second equation follows from the fact that $\theta_{i,\tau}^{(1)} = (1+c)\theta_{i,\tau}^{(2)}$ and $\theta_{i,\tau} = \frac{\theta_{i,\tau}^{(1)} + \theta_{i,\tau}^{(2)}}{2} = \frac{2+c}{2}\theta_{i,\tau}^{(2)}$. Moreover, under Assumptions 4.1, we know from Lemma 4.1 that for $k = 1$ and 2,

$$\sqrt{T}\frac{\gamma_{i,\tau}^{(k)}}{2v_{i,\tau}^{(k)}}\frac{\lambda_{i,\tau}^{(k)} - \theta_{i,\tau}^{(k)}}{\theta_{i,\tau}^{(k)}} \Rightarrow \mathcal{N}(0,1),$$

as $T, N \to \infty$ where $\theta_{i,\tau}^{(k)}$ is the asymptotic centring of $\lambda_{i,\tau}^{(k)}$. As a result,

$$\frac{\lambda_{i,\tau}^{(k)}}{\theta_{i,\tau}^{(k)}} = 1 + o_P\left(\frac{1}{\sqrt{T}}\right),$$

as $T, N \to \infty$, where we stress the fact that $\gamma_{i,\tau}^{(k)}$ and $v_{i,\tau}^{(k)}$ are constant when $T, N \to \infty$. Therefore, (4.17) reduces to

$$\frac{\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}} = \frac{2+2c}{2+c}\left(1 + o_P\left(\frac{1}{\sqrt{T}}\right)\right) - \frac{2}{2+c}\left(1 + o_P\left(\frac{1}{\sqrt{T}}\right)\right) = \frac{2c}{2+c} + o_P\left(\frac{1}{\sqrt{T}}\frac{2c}{2+c}\right),$$

for $T, N \to \infty$, and we conclude that

$$\sqrt{T}\frac{\lambda_{i,\tau}^{(1)} - \lambda_{i,\tau}^{(2)}}{\theta_{i,\tau}} = \sqrt{T}\frac{2c}{2+c} + o_P\left(\frac{2c}{2+c}\right),$$

when $T, N \to \infty$.

Consequently, when $T, N \to \infty$, $Z_{i,\tau} \to \infty$ as long as $\sqrt{T}\frac{2c}{2+c} \to \infty$ and $\lambda_{i,\tau}^{(1)} \neq \lambda_{i,\tau}^{(2)}$. And it is sufficient to show the assertion in this theorem. □

Chapter 5

Conclusions and Future Works

This thesis studies statistical inferences for high-dimensional data, especially time series, based on dimension reduction methods, such as factor models and principal component analysis. Our main contribution is the novel statistical methods proposed in each chapter, they are sieve bootstrap, homogeneity and sub-homogeneity pursuit, and the equivalence test for spiked eigenvalues of autocovariance matrix (the autocovariance test). In particular, the sieve bootstrap and the autocovariance test are proposed based on the spiked eigenstructure of the autocovariance matrix where factor models are introduced to represent the temporal dependence of the original high-dimensional time series by low-dimensional factors. Meanwhile, the CPCA method is developed to simultaneously estimate homogeneity and sub-homogeneity (group-specific information), where the data is assumed to have a more complicated spiked eigenstructure in its covariance matrix. Besides, the work in Chapter 2 is not only a building block of bootstrap methods, but also act as an essential high-dimensional statistical method. For example, when implementing the autocovariance test in Chapter 4, the sieve bootstrap method can be utilised to estimate unknown quantities in the test statistic.

The work presented in this thesis leaves several directions open for future research. In both Chapter 2 and 4, the high-dimensional time series we study are assumed to be stationary or even follow linear models. One of the potential works is then related to extending the existing sieve bootstrap method proposed in Chapter 2 to non-linear time series and non-stationary time series. As discussed in Kreiss et al. (2011) and Bühlmann (1997), when time series are non-linear but

stationary, the sieve bootstrap may be still applicable as long as the autoregressive representation exists. Besides, other bootstrap schemes, such as block bootstrap, are also applicable for non-linear time series; therefore, they can also be applied to the factors for statistical inferences. However, despite that block bootstrap methods may provide a better finite sample performance than sieve bootstrap methods (Bühlmann, 1997), most of the bootstrap methods are not valid for general non-stationary time series. As discussed in Bühlmann (1998) and Kreiss et al. (2011), one of the exceptions is that the sieve bootstrap can correctly estimate the deterministic trend of non-stationary time series and provide valid inferences. Therefore, our bootstrap may be extended to a broader class of time series.

Besides, as discussed in Chapter 4, the p-value of the autocovariance test can be considered a measure of dissimilarity between two populations of high-dimensional time series. Therefore, a hierarchical clustering method (Gordon, 1999) for high-dimensional time series can be proposed, which can motivate our method's applications on not only hypothesis testings but also clusterings for many real data. Moreover, no matter if two high-dimensional time series share the same spiked eigenvalues or not, another potential extension to our work is to adopt canonical correlation analysis (CCA) (Anderson, 2003) on the low-dimensional factors to study whether factors from one high-dimensional time series have the predictability on the other one. In other words, we can use CCA to pursue simultaneous inferences for two high-dimensional time series.

Bibliography

Anderson, T., 2003. *An Introduction to Multivariate Statistical Analysis*. Wiley. (cited on pages 75 and 144)

Anderson, T. W., 1971. *The Statistical Analysis of Time Series*. Wiley, New York. ISBN 978-0-471-02900-7. (cited on pages 8 and 121)

Aue, A.; Norinho, D. D.; and Hörmann, S., 2015. On the prediction of stationary functional time series. *Journal of the American Statistical Association: Theory and Methods*, 110, 509 (Jan. 2015), 378–392. (cited on page 44)

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

Bai, J., 2003b. Inferential theory for factor models of large dimensions. *Econometrica*, 71, 1 (2003), 135–171. (cited on pages 8, 13, 112, and 117)

Bai, J. and Ng, S., 2002a. Determining the number of factors in approximate factor models. *Econometrica*, 70, 1 (2002), 191–221. (cited on pages 1 and 91)

Bai, J. and Ng, S., 2002b. Determining the number of factors in approximate factor models. *Econometrica*, 70, 1 (2002), 191–221. (cited on pages 8, 13, 112, and 117)

Bai, Z. and Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, , 6 (1996), 311–329. (cited on page 112)

Bi, D.; Nie, A. J.; Han, X.; and Yang, Y., 2020. Central limit theorem for spiked eigenvalues of sample autocovariance matrices for high-dimensional time series. https://drive.google.com/drive/folders/1R7-HvOtctj-QWmpNRXkgUrGq4Cui9eBm?usp=sharing. (cited on pages 5, 111, 116, and 121)

Boivin, J. and Ng, S., 2006. Are more data always better for factor analysis? *Journal of Econometrics*, 132, 1 (2006), 169–194. (cited on pages 3 and 75)

Brockwell, P. J. and Davis, R. A., 1991. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edn. ISBN 978-0-387-97429-3. (cited on pages 14 and 62)

Bühlmann, P., 1997. Sieve bootstrap for time series. *Bernoulli*, 3, 2 (Jun. 1997), 123–148. (cited on pages 8, 24, 143, and 144)

Bühlmann, P., 1998. Sieve bootstrap for smoothing in nonstationary time series. *Annals of Statistics*, 26, 1 (Feb. 1998), 48–83. Publisher: Institute of Mathematical Statistics. (cited on page 144)

Bühlmann, P.; Rütimann, P.; van de Geer, S.; and Zhang, C.-H., 2013. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143, 11 (2013), 1835–1858. (cited on page 88)

Cai, T.; Han, X.; and Pan, G., 2017. Limiting laws for divergent spiked eigenvalues and largest non-spiked eigenvalue of sample covariance matrices. *arXiv preprint arXiv:1711.00217*, (2017). (cited on page 76)

Cai, T. T. and Yuan, M., 2011. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39, 5 (Oct. 2011), 2330–2355. (cited on page 23)

Chen, S. X. and Qin, Y.-L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, 38, 2 (Apr. 2010), 808–835. Publisher: Institute of Mathematical Statistics. (cited on page 112)

Cheng, R. and Pourahmadi, M., 1993. Baxter's inequality and convergence of finite predictors of multivariate stochastic processess. *Probability Theory and Related Fields*, 95, 1 (Mar. 1993), 115–124. (cited on pages 16 and 22)

Chiou, J.-M. and Li, P.-L., 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 4 (2007), 679–699. (cited on pages 4, 78, and 88)

Cramér, H. and Wold, H., 1936. Some Theorems on Distribution Functions. *Journal of the London Mathematical Society*, s1-11, 4 (1936), 290–294. _eprint: https://londmathsoc.onlinelibrary.wiley.com/doi/pdf/10.1112/jlms/s1-11.4.290. (cited on page 49)

Deville, J.-C. and Malinvaud, E., 1983. Data analysis in official socio-economic statistics. *Journal of the Royal Statistical Society: Series A (General)*, 146, 4 (1983), 335–352. (cited on page 83)

Donoho, D. L. et al., 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1, 32 (2000), 375. (cited on page 78)

Efron, B., 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1 (Jan. 1979), 1–26. (cited on pages 3, 7, and 8)

El Karoui, N. and Purdom, E., 2018. Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19, 1 (Jan. 2018), 170–235. (cited on pages 3 and 7)

Fama, E. F. and French, K. R., 1997. Industry costs of equity. *Journal of Financial Economics*, 43, 2 (1997), 153–193. (cited on pages 76 and 104)

Fan, J.; Han, F.; and Liu, H., 2014. Challenges of big data analysis. *National Science Review*, 1, 2 (2014), 293–314. (cited on page 78)

Fan, J.; Liao, Y.; and Mincheva, M., 2011. High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39, 6 (Dec. 2011), 3320–3356. (cited on page 9)

Fan, J.; Liao, Y.; and Mincheva, M., 2013a. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 4 (2013), 603–680. (cited on pages 2, 83, 91, 103, and 104)

Fan, J.; Liao, Y.; and Mincheva, M., 2013b. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75, 4 (2013), 603–680. (cited on pages 15 and 58)

Fan, J.; Sun, Q.; Zhou, W.-X.; and Zhu, Z., 2018. Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, (2018), 1–13. (cited on pages 75 and 91)

Gneiting, T. and Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association: Review Article*, 102, 477 (Mar. 2007), 359–378. (cited on page 37)

Gordon, A. D., 1999. *Classification, 2nd Edition*. CRC Press. ISBN 978-1-58488-853-6. Google-Books-ID: _w5AJtbfEz4C. (cited on page 144)

Hall, P.; Müller, H.-G.; and Wang, J.-L., 2006. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34, 3 (Jun. 2006), 1493–1517. (cited on page 10)

Hong, D.; Fessler, J. A.; and Balzano, L., 2018. Optimally weighted pca for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, (2018). (cited on page 83)

Hörmann, S.; Kidziński, Ł.; and Hallin, M., 2015. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 2 (2015), 319–348. (cited on page 44)

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 6 (1933), 417. (cited on pages 2 and 75)

Hubert, L. and Arabie, P., 1985. Comparing partitions. *Journal of Classification*, 2, 1 (1985), 193–218. (cited on pages 89 and 94)

Hyndman, R. J. and Shahid Ullah, M., 2007. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51, 10 (Jun. 2007), 4942–4956. (cited on page 134)

Johnstone, I. M., 2001. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29, 2 (2001), 295–327. (cited on page 76)

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 (2009), 682–693. (cited on page 78)

Johnstone, I. M. and Paul, D., 2018. PCA in High Dimensions: An Orientation. *Proceedings of the IEEE*, 106, 8 (Aug. 2018), 1277–1292. (cited on page 2)

Jolliffe, I., 2002. *Principal Component Analysis*. Springer Verlag, New York. (cited on pages 75 and 89)

Jolliffe, I. T., 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edn. ISBN 978-0-387-95442-4. (cited on pages 1 and 112)

Jolliffe, I. T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 2065 (2016), 20150202. (cited on page 75)

Krampe, J.; Kreiss, J.-P.; and Paparoditis, E., 2019. Bootstrap based inference for sparse high-dimensional time series models. Technical report, arXiv. (cited on page 8)

Kreiss, J.-P., 1988. *Asymptotic statistical inference for a class of stochastic processes*. Ph.D. thesis, Department of Mathematics, University of Hamburg. (cited on pages 3 and 8)

Kreiss, J.-P., 1992. Bootstrap procedures for AR (∞) — processes. In *Bootstrapping and Related Techniques*, Lecture Notes in Economics and Mathematical Systems, 107–113. Springer, Berlin, Heidelberg. (cited on page 19)

Kreiss, J.-P. and Lahiri, S. N., 2012. 1 - Bootstrap Methods for Time Series. In *Handbook of Statistics* (Eds. T. Subba Rao; S. Subba Rao; and C. R. Rao), vol. 30 of *Time Series Analysis: Methods and Applications*, 3–26. Elsevier. (cited on page 3)

Kreiss, J.-P.; Paparoditis, E.; and Politis, D. N., 2011. On the range of validity of the autoregressive sieve bootstrap. *The Annals of Statistics*, 39, 4 (Aug. 2011), 2103–2130. (cited on pages 8, 16, 21, 23, 24, 25, 143, and 144)

Ku, W.; Storer, R. H.; and Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30, 1 (Nov. 1995), 179–196. (cited on page 112)

Kulperger, R. J. and Rao, B. L. S. P., 1989. Bootstrapping a finite state markov chain. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 51, 2 (1989), 178–191. Publisher: Springer. (cited on page 3)

Kunsch, H. R., 1989. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 3 (Sep. 1989), 1217–1241. (cited on pages 3, 8, and 24)

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 (Apr. 2012), 694–726. (cited on pages 2, 114, and 120)

Lam, C.; Yao, Q.; and Bathia, N., 2011. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98, 4 (Dec. 2011), 901–918. Publisher: Oxford Academic. (cited on pages 2, 3, 9, 13, 14, 15, 16, 17, 18, 23, 26, 59, 60, 112, 114, 117, 119, 124, and 128)

Lee, R. D. and Carter, L. R., 1992. Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association: Applications and Case Studies*, 87,

419 (1992), 659–671. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. (cited on page 134)

Li, N.; Lee, R.; and Gerland, P., 2013. Extending the Lee-Carter method to model the rotation of age patterns of mortality-decline for long-term projection. *Demography*, 50, 6 (Dec. 2013), 2037–2051. (cited on page 134)

Li, Q.; Cheng, G.; Fan, J.; and Wang, Y., 2018. Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association*, 113, 521 (2018), 380–389. (cited on page 83)

Li, Z.; Wang, Q.; and Yao, J., 2017. Identifying the number of factors from singular values of a large sample auto-covariance matrix. *Annals of Statistics*, 45, 1 (Feb. 2017), 257–288. Publisher: Institute of Mathematical Statistics. (cited on pages 117, 118, and 119)

Liu, A.; Zhang, Y.; Gehan, E.; and Clarke, R., 2002. Block principal component analysis with application to gene microarray data classification. *Statistics in Medicine*, 21, 22 (2002), 3465–3474. (cited on pages 77 and 82)

Meyer, M. and Kreiss, J.-P., 2015. On the vector autoregressive sieve bootstrap. *Journal of Time Series Analysis*, 36, 3 (2015), 377–397. (cited on pages 8, 16, 19, 21, 22, 23, 24, 25, 49, 65, 66, and 67)

Morales-Jimenez, D.; Johnstone, I. M.; McKay, M. R.; and Yang, J., 2018. Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *arXiv preprint arXiv:1810.10214*, (2018). (cited on page 76)

Pan, G. M. and Zhou, W., 2011. Central limit theorem for Hotelling's T2 statistic under large dimension. *Annals of Applied Probability*, 21, 5 (Oct. 2011), 1860–1910. Publisher: Institute of Mathematical Statistics. (cited on page 112)

Paparoditis, E., 2018. Sieve bootstrap for functional time series. *The Annals of Statistics*, 46, 6B (Dec. 2018), 3510–3538. (cited on pages 8, 9, and 19)

Pearson, K., 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 11 (1901), 559–572. (cited on pages 2 and 75)

Politis, D. N.; Romano, J. P.; and Wolf, M., 1997. Subsampling for heteroskedastic time series. *Journal of Econometrics*, 81, 2 (Dec. 1997), 281–317. (cited on page 49)

Ramsay, J. O. and Silverman, B. W., 2002. *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer-Verlag, New York. (cited on pages 9, 10, and 23)

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 336 (1971), 846–850. (cited on pages 89 and 94)

Shang, H. L., 2018. Bootstrap methods for stationary functional time series. *Statistics and Computing*, 28, 1 (Jan. 2018), 1–10. (cited on pages 44 and 45)

Sowell, F., 1989. A decomposition of block toeplitz matrices with applications to vector time series. Technical report, Carnegie Mellon University. (cited on page 63)

Tan, K. M.; Witten, D.; and Shojaie, A., 2015. The cluster graphical lasso for improved estimation of gaussian graphical models. *Computational Statistics & Data Analysis*, 85 (2015), 23–36. (cited on pages 77 and 82)

Tang, C.; Lin, S. H.; and Yang, Y., 2020. Clustering and forecasting multiple functional time series. Unpublished. (cited on page 134)

Tibshirani, R.; Wainwright, M.; and Hastie, T., 2015. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC. (cited on page 78)

University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), 2018. *Human Mortality Database*. www.mortality.org. (cited on page 135)

Wang, W. and Fan, J., 2017. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45, 3 (Jun. 2017), 1342–1374. (cited on page 10)

Wiener, N. and Masani, P., 1958. The prediction theory of multivariate stochastic processes, II: The linear predictor. *Acta Mathematica*, 99 (1958), 93–137. (cited on pages 3, 8, 21, and 22)

Yao, F.; Müller, H.-G.; and Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association: Theory and Methods*, 100, 470 (Jun. 2005), 577–590. (cited on page 10)

Yuan, M. and Lin, Y., 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 1 (2006), 49–67. (cited on page 99)

Zhang, X. and Wang, J.-L., 2016. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44, 5 (Oct. 2016), 2281–2321. (cited on page 10)

Zou, H.; Hastie, T.; and Tibshirani, R., 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 2 (2006), 265–286. (cited on page 78)