STATISTICAL LEARNING OF TEMPORALLY DEPENDENT HIGH- & MULTI-DIMENSIONAL DATA

by

YI CHEN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Rong Chen

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Statistical Learning of Temporally Dependent High- & Multi-Dimensional Data

by Yi Chen

Dissertation Director: Rong Chen

The growing capabilities in generating and collecting data has risen unique opportunities and challenges in Statistics and the emerging field of Data Science. The availability of data with complex structure, such as temporal dependence and multi-dimensional, provides scientists with more accurate ways to characterize intricate natural or social phenomenon. This thesis deals with statistical models, methods, theory, and algorithms for learning low-rank structures from temporal-dependent multi-dimensional data, including time series with matrix observations, dynamic networks, and multivariate spatial-temporal data. We established a unified framework of modeling such data as matrix-variate time series that faithfully preserves the structural properties and the temporal dependencies that are intrinsic to the data. The focus is to achieve dimension reduction and learn the underlying latent low-rank structure of the data. The models presented in this thesis extend the matrix factor model proposed by Wang et al. (2017) in three directions to fully exploit the structures and properties of the observed data.

Specifically, the constrained matrix factor models provide a general framework for incorporating domain or prior knowledge in the matrix factor model through linear constraints. The proposed framework is shown to be useful in achieving parsimonious parameterization, gaining efficiency in statistical inference, facilitating interpretation of

the latent matrix factor, and identifying specific factors of interest. The factor models for dynamic networks target at a special kind of matrix time series where, at each time point, the observation is a square adjacency matrix whose rows and columns represent the same set of actors in the network. Most available probability and statistical models for dynamic network data are deduced from random graph theory where the networks are characterized on the level of node and edge. Our high-level modeling of the dynamic networks as a time series of relational matrices is less restrictive and more scaleable to high-dimensional dynamic network data which is very common nowadays.

The factor models for multivariate spatial-temporal data are designed to accommodate the smooth functional behavior of the underlying spatial process. The functional matrix factor model aims to explicitly express discrete observations from spatial continuum in the form of a function. It has the advantage of generating models that can describe continuous smooth spatial changes, which then allows for accurate estimates of parameters, effective data noise reduction through curve/surface smoothing, and applicability to data with irregular spatial sampling.

The estimating methods are generally based on moment matching and spectral decomposition of matrices constructed from the empirical auto-cross-covariance of the time series, thus capturing the temporal dynamics presented in the data. The latent low-rank structures are learned directly from the data with little subjective input or any restricted distributional assumptions. For the functional matrix factor model, the functional loadings are approximated non-parametrically. The estimated latent states or factors are of smaller dimensions and can be used as data in second-stage inference and prediction. Theoretical properties of the estimators are established. Simulation studies are carried out to demonstrate the finite-sample performance of the proposed methods and their associated asymptotic properties. The proposed methods are applied to a wide range of real datasets, such as multinational macroeconomic indices data, dynamic global trading networks, and the Comprehensive Climate Dataset among others.

Acknowledgements

I wish to express my sincere appreciation to those who have contributed to this thesis and supported me in one way or another during this amazing journey.

First and foremost, I would like to express my deepest and sincerest gratitude to my advisor **Prof. Rong Chen** who has been a tremendous mentor for me. I would like to thank him for the continuous support of my Ph.D. study and research, for his patience, inspiration, and immense knowledge. His guidance about how to approach research, write, give talks, and cool down my neurosis has been invaluable. He did not give up on me even when I was at my worst. Without his help, I could not have walked out of the most difficult time of my life and have it all – finished my Ph.D. thesis, found an academic job, and raised a three years old daughter – at the same time. I am also extending my heartfelt thanks to his wife for hosting several wonderful parties and for being an excellent example of balancing motherhood and a successful career.

I would especially like to thank **Prof. Ruey S. Tsay** for his insightful comments and suggestions to my research, and also for the hard question which incented me to widen my research from various perspectives. I remain amazed that despite his busy schedule, he was able to go through the final draft of our paper and marked comments and suggestions on almost every page. He is an inspiration.

I am extremely grateful to **Prof. Qiwei Yao** who has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I really appreciate his willingness to speak with me in short notice every time and going through every question I have. His advice on research as well as on my career have been invaluable.

My heartfelt thanks goes to **Prof. Regina Y. Liu** for her advices and encouragements during my doctoral research, especially at the hard time when I experienced

postpartum depression. She has always been an excellent role model for me as a successful woman statistician and professor.

I would also like to thank **Prof. Han Xiao**, **Prof. Tirthankar Dasgupta** and **Prof. Yuan Liao** for serving as my committee members. I also want to thank them for letting my defense be an enjoyable moment, and for their brilliant comments and suggestions.

Dozens of people have helped and taught me immensely at Rutgers University: **Prof. Tirthankar Dasgupta**, **Prof. John E. Kolassa**, **Prof. Minge Xie**, and **Prof. Harry D. Crane** have all been wonderfully encouraging, giving feedback on talks, papers and proposals, and preparing me in various ways as a researcher.

Nobody has been more important to me in the pursuit of my dreams than the members of my family. I would like to thank my parents for their unconditional love. Most importantly, I wish to thank my loving and supportive hunsband, Xiang, and my beautiful daughter, Victoria, who provides unending inspiration.

Dedication

To Xiang and Victoria.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Scientific studies of various natural and social phenomenons crucially depends on the analysis of data collected to characterize the phenomenons in the fields. Nowadays complex data that can better depict the real world are becoming widely available thanks to the development of information technology and the migration of human activities towards electronic devices and Internet. Temporal dependent and multi-dimensional data are of special interests because of their ability to capture the temporal feature and congregate multiple aspects of real-world phenomenons. For example, macroeconomic indicators reported by different countries can be viewed as a time series of two dimensional matrices whose rows and columns correspond to countries and macro indicators respectively. Economic data are inherently serially and cross-sectionally correlated. Thus it is essential to pertain the temporal dependence and analyze multiple times series collectively. Since the country and the macro indicator are two distinctive aspects of the data, it is also important to model these two dimensions differently. Other examples include dynamic network data and multivariate spatial temporal data. The first have their network features recored in the bilateral relationships between the actors and their dynamic features captured in the temporal dependence, while the latter are intrinsically composed of three different dimensions, namely variable, space, and time. As a introduction, Section 1.1 will discuss matrix-variate time series, dynamic networks, and multivariate spatial temporal data in detail. It will also provide a general review of current research on analyzing such data.

While a wide range of statistical tools and techniques for data analysis already exist, the increasing availability of complex data structures calls for new approaches that can faithfully preserve the inherent structures. This thesis presents a unified

framework of modeling such data as matrix-variate time series that faithfully preserve the intrinsic multi-dimensional feature and the temporal dependences. The focus is to achieve dimension reduction and to learn the underlying low-rank latent structure. The methods are derived from factor models in unsupervised statistical learning. In Section 1.2, we will give a preliminary introduction to the factor models.

1.1 Temporal Dependent Multi-Dimensional Data

1.1.1 Matrix-Variate Time Series

In many fields, such as economics, finance and social science, high-dimensional matrix-variate data are becoming readily available. Matrix-variate time series is defined as a sequence of observations in matrix form observed over time. For example, every quarter, countries report a set of economic indicators such as GDP, consumer prices and interest rate et al. At each time point, the obervation is a matrix whose rows represent the countries and whose columns represents the macroeconomic indices. Also, multi-corporate financial data are usually arranged as time series of matrices whose rows represent the companies and columns represents the financial indices. See Figure 1.1 for an illustration of matrix time series.

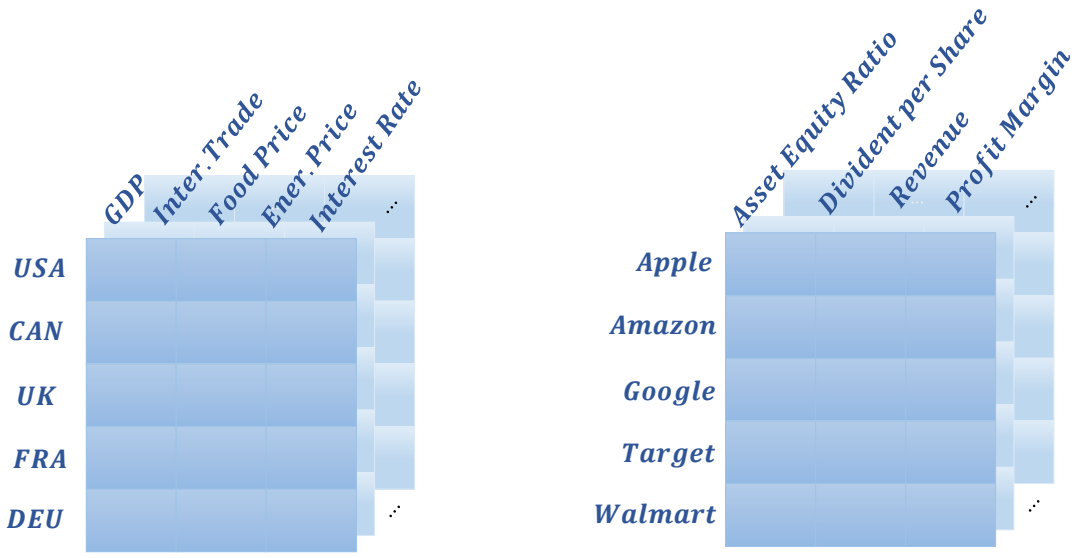


Figure 1.1: Illustration of Matrix Time Series.

Each cell of the matrix time series represents a univariate time series. For example, the first cell represents the univariate time series of the GDP of the United States. Every row or every column corresponds to a multivariate time series. For example, the first rows is the time series of multiple macroeconomic indices of the United States and the first column is the multivariate time series of the GDP of multiple countries. However, it is preferred to analyze multiple time series simultaneously and also preserve the matrix structure because the correlations between variables are valuable information and the row variables and the column variables are correlated in different ways.

Development of statistical methods for analyzing such data is still in its infancy, and as a result, scientists frequently analyze matrix time series by separately modeling each element series or by 'flattening' them to vector time series Box et al. (2015); Brockwell and Davis (2013); Tsay (2013); Fan and Yao (2005); Lam et al. (2012); Bai and Ng (2002a); Bai (2003a). This destroys the intrinsic matrix structure and misses important patterns in the data. This thesis aims to analyze high-dimensional tensor time series, while preserving the genuine tensor structure, accounting for the temporal dependence, and enabling scalability to high-dimensions.

1.1.2 Dynamic Networks

Nowadays in a variety of fields, such as economics and social studies, researchers observe high-dimensional matrix-variate time series where observation at each time point is in a matrix structure. One special kind of such data is a time series of square matrices that describe pairwise relationships among a set of entities. For example, international trade commodity flow data between n countries over a period of time can be represented as matrix time series $\{\boldsymbol{X}_t\}_{t=1:T}$, where $\boldsymbol{X}_t$ is a $n \times n$ matrix, and each element $x_{ij,t}$ is the directed level of trade from nation i to nation j at time t. The i-th row represents data for which nation i is the exporter and the column j represents data for which j is the importer. Since the data are based on pairs of nations, the diagonal representing the relationships of nations with themselves is generally absent.

Such *network/relational data* that consist of measurements made on pairs of entities have been researched from various aspects over the last decades. Developed statistical

models for *static network analysis* include the class of exponential random graph models (ERGMs) that are analogous to standard regression models (Wasserman and Pattison (1996); Robins et al. (2007); Lusher et al. (2012)), the class of stochastic block models that are, in their most basic form, essentially a mixture of classical random graph models (Holland et al. (1983); Nowicki and Snijders (2001); Daudin et al. (2008); Airoldi et al. (2008); Karrer and Newman (2011)) and the class of latent network models that use both observed and unobserved variables in modeling the presence or absence of network edges (Hoff et al. (2002); Hoff (2008, 2005, 2009, 2015b,a); Cranmer et al. (2016)). All models mentioned thus far only consider a 'snapshot' $\boldsymbol{X}_t$ of a dynamic process in a given 'slice' of time t, and thus not able to discover the dynamic pattern of the network nor to answer scientific questions concerned with the evolution of networks over time.

Statistical research on *dynamic network analysis* is less developed compared to the existing literature on modeling of static network graphs. While there has been a substantial amount of work done in the past decades on the mathematical and probabilistic modeling of dynamic processes on network graphs (see Barrat et al. (2008)), there has been comparatively much less work on the statistical side. Snijders and colleagues (Snijders (2001); Huisman and Snijders (2003); Snijders (2005, 2006); Snijders et al. (2007, 2010)) developed an actor-based model for network evolution that incorporated individual level attributes. The approach is based on an economic model of rational choice, whereby actors make decisions to maximize individual utility functions. Hanneke et al. (2010) and Krivitsky and Handcock (2014) introduced a class of temporal exponential random graph models for logitudinal network data (i.e. the networks are observed in panels). They model the formation and dissolution of edges in a separable fashion, assuming an exponential family model for the transition probability from a network at time t to a network at time $t+1$. Westveld and Hoff (2011) represent the network and temporal dependencies with a random effects model, resulting in a stochastic process defined by a set of stationary covariance matrices. Xing et al. (2010) extends an ealier work on a mixed membership stochastic block model for static network (Airoldi et al.

(2008)) to the dynamic senario by using a state-space model where the mixed memembership is characterized through the observation function and the dynamics of the latent 'tomographic' states are defined by the state function. Estimation is based on the maximum likelihood principal using a variational EM algorithm. They deduced from random graph theory Crane et al. (2016); Krivitsky and Handcock (2014). These methods are deduced from random graph theory and model the relational data at relation (edge) or entity (node) level, and thus often confronted with computational challenges, overparametrization, and overfitting issues when dealing with high-dimensional matrix time series, which are very common in economics and social networks nowadays.

In contrast to the pre-existing research in dynamic network analysis, the approach we propose in Chapter 3 is more time series oriented, in that a dynamic network is treated as a time series of matrix observations – the relational matrices– instead of the traditional nodes and edges characterization. We adopt a matrix factor model where the observed surface dynamic network is assume to be driven by a latent dynamic network with lower dimensions. The linear relationship between the surface network and the latent network is characterized by unknown but deterministic loading matrices. The latent network and the corresponding loadings are estimated via an eigenanalysis of a positive definite matrix constructed from the auto-cross-covariances of the network times series, thus capturing the dynamics presenting in the network. Since the dimension of the latent network is typically small or at least much smaller than the surface network, the proposed model often result in a concise description of the whole network series, achieving the objective of dimension reduction. The resulting latent network of much smaller dimensions can also be used for downstream microscope analysis of the dynamic network.

Different from Xing et al. (2010) that summarize the relational data by the relationships between a small number of groups, we impose neither any distributional assumptions on the underlying network nor any parametric forms on its covariance function. The latent network is learned directly from the data with little subjective input. The meaning of the nodes of the latent network in our model is automatically learned from the data and is not confined to the 'groups' to which the actors belong,

which provide a more flexible interpretation of the data. Additionally, our modeling framework is very flexible and extendable: Using a matrix factor model framework, it can accommodate continuous and ordinal relational data. It can be extended to incorporate prior information on the network structure or include exogenous and endogenous covariate as explanatory variables of the relationships.

1.1.3 Multivariate Spatial-Temporal Process

The increasing availability of multivariate data referenced over geographic regions and time in various applications has created unique opportunities and challenges for those practitioners seeking to capitalize on their full utility. For example, United States Environmental Protection Agency publishes daily from more than 20,000 monitoring stations a collection of environmental and meteorological measurements such as temperature, pressure, wind speed and direction and various pollutants. Such data naturally constitute a tensor (multi-dimensional array) with three modes (dimensions) representing space, time and variates, respectively. Simultaneously modeling the dependencies between different variates, regions, and times is of great potential to reduce dimensions, produce more accurate estimation and prediction and further provide a deeper understanding of the real world phenomenon. At the same time, methodological issues arise because these data exhibit complex multivariate spatio-temporal covariances that may involve non-stationarity and potential dependencies between spatial locations, time points and different processes. Traditionally, researchers mainly restrict their analysis to only two dimensions while fixing the third: time series analysis applied to a slice of such data at one location focus on temporal modeling and prediction (Box et al. (2015); Brockwell and Davis (2013); Tsay (2013); Fan and Yao (2005)); spatial statistical models for a slice of such data at one time point address spatial dependence and prediction over unobserved locations (Cressie (2015)); and univariate spatio-temporal statistics concentrate on only one variable observed over space and time (Cressie and Wikle (2015)).

Since physical processes rarely occur in isolation but rather influence and interact

with one another, multivariate spatio-temporal models are increasingly in demand because the dependencies between multiple variables, locations and times can provide valuable information for understanding real world phenomenons. Various multivariate spatio-temporal conditional autoregressive models have been proposed by Carlin et al. (2003); Congdon (2004); Pettitt et al. (2002); Zhu et al. (2005); Daniels et al. (2006); Tzala and Best (2008), among others. However, these methodologies cannot efficiently model high-dimensional data sets. Additionally, these approaches impose separability and various independence assumptions, which are not appropriate for many settings, as these models fails to capture important interactions and dependencies between different variables, regions, and times (Stein (2005b)). Bradley et al. (2015) introduced a multivariate spatio-temporal mixed effects model to analyze high-dimensional multivariate data sets that vary over different geographic regions and time points. They adopt a reduced rank spatial structure (Wikle (2010)) and model temporal behavior via vector autoregressive components. However, their method only applies to low-dimensional multivariate observations because they model each variable separately. In addition, they assume the random effect term is common across all processes which is unrealistic especially in the case with a large number of variables.

In Chapter 4, we present a new class of multivariate spatio-temporal models that model spatial, temporal and variate dependence simultaneously. The model builds upon the matrix factor models proposed in Wang et al. (2017), while further incorporating the functional structure of the spatial process and dynamics of the latent matrix factor. The spatial dependence is model by the spatial loading functions, the variable dependence is modeled by the variable loading matrix, while the temporal dependence is modeled by the latent factors of first-order autoregressive matrix time series.

Some spatial-factor-analysis models that capture spatial dependence through factor processes have been developed in the literature. Lopes et al. (2008) considers univariate observations but uses factor analysis to reduce (identify) clusters/groups of locations/regions whose temporal behavior is primarily described by a potentially small set of common dynamic latent factors. Also working with the univariate case, Cressie and Johannesson (2008) successfully reduces the computational cost of kriging by using a

flexible family of non-stationary covariance functions constructed from low rank basis functions. See also Wikle (2010). For multivariate spatial data, Cook et al. (1994) introduced the concept of a spatially shifted factor and a single-factor shifted-lag model and Majure and Cressie (1997) discussed graphical methods for identifying shifts. Following the ideas of multiple-lag dynamic factor models that generalize static factor models in the time series setting, Christensen and Amemiya (2001, 2002, 2003) extended the shifted-lag model to a generalized shifted-factor model by adding multiple shifted-lags and developed a systematic statistical estimation, inference, and prediction procedure. The assumption that spatial processes are second-order stationary is required for the moment-based estimation procedure and the theoretical development. Our modeling of the spatial dependence though latent factor processes is different from the aforementioned methods in that we impose no assumptions about the stationarity over space, nor the distribution of data, nor the form of spatial covariance functions. The idea is similar to that of Huang et al. (2016), however we aim at estimating the spatial loading functions instead of the loading matrix and kriging at unsampled location is based on the loading function. In addition, future forecasting in our model reserves the matrix formation of the observation and temporal dependence through the matrix auto-regression of order one.

1.2 Unsupervised Statistical Learning of Latent Structure

1.2.1 Unsupervised Statistical Learning with Factor Models

Unsupervised statistical learning (Hastie et al. (2009)) focuses on methods that search for patterns in the data and extract useful information without training samples of previously solved samples. Factor models are one powerful approach of unsupervised statistical learning to reduce the dimensionality and extract the latent structure of the data. They provide a flexible way of describing correlations among the observed variables by assuming that the co-movements of a high-dimensional observed data was driven by a few unobserved (latent) common variables (factors). The latent factors present a low-dimensional summary of the observed data and can be considered as

concise and de-noised descriptions of the underlying processes that have generated the data. The representations can then be used to understand the data generation processes or predict the unobserved data entities.

In the context of temporal dependent or time series data, let p be the number of cross-section units and T be the number of time series observations. For $i = 1, \ldots, p$, $t = 1, \ldots, T$, a factor model is defined as

$$x_{ti} = \boldsymbol{\lambda}_i' \boldsymbol{f}_t + e_{it}, \tag{1.1}$$

where $\boldsymbol{f}_t = (f_{t1}, \ldots, f_{tr})'$ is a r-dimensional latent factor and $\boldsymbol{\lambda}_i' = (\lambda_{i1}, \ldots, \lambda_{ir})$ is the factor loadings of variable i on r factors.

Let $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)$. We have model (1.1) in vector form

$$\boldsymbol{x}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{e}_t. \tag{1.2}$$

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$ be the $T \times p$ data matrix and $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)'$ be the $T \times r$ factor matrix, the factor model (1.1) is written in matrix form

$$\boldsymbol{X} = \boldsymbol{F} \boldsymbol{\Lambda}' + \boldsymbol{E}. \tag{1.3}$$

The classic factor model, later referred to as the *strict* factor model (Chamberlain and Rothschild (1983)), assumes that (i) $\boldsymbol{f}_t$ and $\boldsymbol{e}_t$ in (1.2) are generally assumed to be serially (across t) and cross-sectionally (across i) uncorrelated; (ii) p is fixed while T increases to infinity, or vice versa; (iii) both $\boldsymbol{f}_t$ and $\boldsymbol{e}_t$ are normally distributed. See Anderson (2003). In the past decades, a large community of researchers have extended the classical factor models in various ways by relaxing the three mentioned assumptions: dynamic factor models for time series explicitly recognize the factor that data being analyzed are being serially correlated; large dimension factor models allow the sample size in both dimensions increases to infinity in the asymptotic theory; and approximate factor models allow the noise term $\boldsymbol{e}_t$ to be "weakly" correlated serially (across t) and cross-sectionally (across i).

Our focus is on the high-dimensional approximate dynamic factor models, where $p, T \to \infty$ at the same time, $\boldsymbol{e}_t$ are allowed to be cross-sectionally, and the observations

are temporal dependent, that is $\boldsymbol{x}_t$ are serially correlated. Dynamic factor models defined in econometrics and finance literatures attempt to separate the common factors that affect the dynamics of most original component series from the idiosyncratic series that at most affect the dynamics of a few original time series. (See Chamberlain (1983), Chamberlain and Rothschild (1983), Bai (2003b), Bai and Ng (2002b), Bai and Ng (2007), Forni et al. (2000), Forni et al. (2004).) Such definition is appealing in analyzing economic and financial phenomena. But the fact that idiosyncratic part may exhibit serial correlations poses technical difficulties in both identification and inference. These factor models are only asymptotically identifiable because the rigorous definition of the common factors can only be established when the dimension of time series goes to infinity. On the other hand, Pan and Yao (2008), Lam et al. (2011a), and Lam and Yao (2012) adopt a different approach from a dimension-reduction point of view. Different from the aforementioned econometric factor model, they decompose a high-dimensional time series into two parts: a dynamic part driven by, hopefully, a lower-dimensional factor time series, and a static part which is a vector white noise. Since the white noise exhibits no serial correlations, the decomposition is unique in the sense that both the dimension of the factor process and the factor loading space are identifiable for any finite sample size. This decomposition is conceptually simple and makes the tasks of model identification and statistical inference much easier. We will follow this approach by Lam et al. (2011a) and Lam and Yao (2012) in defining the noise term.

1.2.2 Factor Models for Matrix-variate Time Series

High-dimensional matrix-variate time series have been widely observed nowadays in a variety of scientific fields including economics, meteorology, and ecology. For example, the World Bank and the International Monetary Fund collect and publish macroeconomic data of more than thirty variables spanning over one hundred years and over two hundred countries covering a variety of demographic, social, political, and economic topics. These data neatly form a matrix-variate time series with rows representing the countries and columns representing various macroeconomic indexes. Typical factor

analysis of such data either converts the matrix into a vector or modeling the row or column vectors separately (See Chamberlain (1983), Chamberlain and Rothschild (1983), Bai (2003b), Bai and Ng (2002b), Bai and Ng (2007), Forni et al. (2000), Forni et al. (2004), Pan and Yao (2008), Lam et al. (2011a), and Lam and Yao (2012)). However, the components of matrix-variates are often dependent among rows and columns with certain well-defined structure. Vectorizing a matrix-valued response, or modeling the row or column vectors separately may overlook some intrinsic dependency and fail to capture the matrix structure. Wang et al. (2017) propose a matrix factor model that maintains and utilizes the matrix structure of the data to achieve significant dimension reduction.

Let $\{\boldsymbol{Y}_t\}_{t=1,\ldots,T}$ be a matrix-variate time series, where $\boldsymbol{Y}_t$ is a $p_1 \times p_2$ matrix, that is

$$\boldsymbol{Y}_t = (Y_{\cdot 1,t}, \cdots, Y_{\cdot p_2,t}) = \begin{pmatrix} Y'_{1\cdot,t} \\ \vdots \\ Y'_{p_1\cdot,t} \end{pmatrix} = \begin{pmatrix} y_{11,t} & \cdots & y_{1p_2,t} \\ \vdots & \ddots & \vdots \\ y_{p_1 1,t} & \cdots & y_{p_1 p_2,t} \end{pmatrix}.$$

Wang et al. (2017) propose the following factor model for $\boldsymbol{Y}_t$,

$$\boldsymbol{Y}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t \boldsymbol{\Gamma}' + \boldsymbol{U}_t, \qquad t = 1, 2, \ldots, T, \tag{1.4}$$

where $\boldsymbol{F}_t$ is a $k_1 \times k_2$ unobserved matrix-variate time series of common fundamental factors, $\boldsymbol{\Lambda}$ is a $p_1 \times k_1$ row loading matrix, $\boldsymbol{\Gamma}$ is a $p_2 \times k_2$ column loading matrix, and $\boldsymbol{U}_t$ is a $p_1 \times p_2$ matrix of random errors. In Equation (1.4), $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ and $(c\boldsymbol{\Lambda}, \boldsymbol{\Gamma}/c)$ are equivalent if $c \neq 0$.

In Model (1.4), it is assumed that $vec(\boldsymbol{U}_t) \sim WN(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ and is independent of the factor process $vec(\boldsymbol{F}_t)$. That is, $\{\boldsymbol{U}_t\}_{t=1}^T$ is a white noise matrix-variate time series and the common fundamental factors $\boldsymbol{F}_t$ drive all dynamics and co-movement of $\boldsymbol{Y}_t$. $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ reflect the importance of common factors and their interactions. Wang et al. (2017) provide several interpretations of the loading matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$. Essentially, $\boldsymbol{\Lambda}$ ($\boldsymbol{\Gamma}$) can be viewed as the row (column) loading matrix that reflects how each row (column) in $\boldsymbol{Y}_t$ depends on the factor matrix $\boldsymbol{F}_t$. The interaction between the row and column is introduced through the multiplication of these terms.

However, in factor analysis of matrix time series and many other types of high-dimensional data, the problem of factor interpretations is of paramount importance. Furthermore, it is important in many practical applications to obtain specific latent factors related to certain domain theories, and with the aid of these specific factors to predict future values of interest more accurately. For example, financial researchers may be interested in extracting the latent factors of level, slope, and curvatures of the interest-rate yield curve and in predicting future equity prices based on those factors (Diebold et al. (2005), Diebold et al. (2006), Rudebusch and Wu (2008), and Bansal et al. (2014)).

In many applications, relevant prior or domain knowledge is available or data themselves exhibit certain specific structure. Additional covariates may also have been measured. For example, in business and economic forecasting, sector or group information of variables under study is often available. Such *a priori* information can be incorporated to improve the accuracy and inference of the analysis and to produce more parsimonious and interpretable factors. In other cases, the existing domain knowledge may intrigue researchers' interest in some specific factors. The theories and prior experience may provide guidance for specifying the measurable variables related to the specific factors of interest. It is then desirable to build proper constraints based on those measurable variables in order to effectively obtain the factors of interest.

To address these important issues and practical needs, we extend the matrix factor model of Wang et al. (2017) to impose natural constraints among the column and row variables to incorporate prior knowledge or to induce specific factors. Incorporating *a priori* information in parameter estimation has been widely used in statistical analysis, such as the constrained maximum likelihood estimation, constrained least squares, and penalized least squares. Constrained maximum likelihood estimation with the parameter space defined by linear or smooth nonlinear constraints have been explored in the literature. Hathaway (1985) applies the constrained maximum likelihood estimation to the problem of mixture normal distributions and shows that the constrained estimation avoids the problems of singularities and spurious maximizers facing an unconstrained estimation. Geyer (1991) proposes a general approach applicable to many

models specified by constraints on the parameter space and illustrates his approach with a constrained logistic regression of the incidence of Down's syndrome on maternal age. Penalty methods have also been customarily used to enforce constraints in statistical models including generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators. See, for example, Frank and Friedman (1993), Tibshirani (1996), Liu et al. (2007), Fan and Li (2001), Zou (2006), and Zhang and Lu (2007). The results of these articles show that including the soft constraints as penalizing term enhances the prediction accuracy and improves the interpretation of the resulting statistical model.

For factor models of time series, Tsai and Tsay (2010a) and Tsai et al. (2016) impose constraints, constructed by some empirical procedures, that incorporate the inherent data structure, to both the classical and approximate factor models. Their results show that the constraints are useful tools to obtain parsimonious econometric models for forecasting, to simplify the interpretations of common factors, and to reduce the dimension. Motivated by similar concerns, we consider constrained, multi-term, and partially constrained factor models for high-dimensional matrix-variate time series. Our methods differs from Tsai and Tsay (2010a) in several aspects. First, we deal with matrix factor model and thus have the flexibility to impose row and column constraints. The interaction between the row and column constraints are explored. Second, we adopt a different set of assumptions for factor model defined in Lam et al. (2011a) and Lam and Yao (2012). The matrix-variate time series is decomposed into two parts: a dynamic part driven by a lower-dimensional factor time series and a static part consisting of matrix white noises. Since the white-noise series exhibits no dynamic correlations, the decomposition is unique in the sense that both the dimension of the factor process and the factor loading space are identifiable for a given finite sample size.

Chapter 2

Constrained Factor Models for High-Dimensional Matrix Time Series

High-dimensional matrix-variate time series data are becoming widely available in many scientific fields, such as economics, biology and meteorology. To achieve significant dimension reduction while preserving the intrinsic matrix structure and temporal dynamics in such data, Wang et al. (2017) proposed a matrix factor model that is shown to provide effective analysis. In this paper, we establish a general framework for incorporating domain or prior knowledge in the matrix factor model through linear constraints. The proposed framework is shown to be useful in achieving parsimonious parameterization, facilitating interpretation of the latent matrix factor, and identifying specific factors of interest. Fully utilizing the prior-knowledge-induced constraints results in more efficient and accurate modeling, inference, dimension reduction as well as a clear and better interpretation of the results. In this paper, constrained, multi-term, and partially constrained factor models for matrix-variate time series are developed, with efficient estimation procedures and their asymptotic properties. We show that the convergence rates of the constrained factor loading matrices are much faster than those of the conventional matrix factor analysis under many situations. Simulation studies are carried out to demonstrate finite-sample performance of the proposed method and its associated asymptotic properties. We illustrate the proposed model with three applications, where the constrained matrix-factor models outperform their unconstrained counterparts in the power of variance explanation under the out-of-sample 10-fold cross-validation setting.

The rest of this chapter is organized as follows. Section 2.1 introduces the constrained, multi-term, and partially constrained matrix-variate factor models. Section

2.2 presents estimation procedures for constrained and partially constrained factor models with different constraints. Section 2.3 investigates theoretical properties of the estimators. Section 2.4 presents some simulation results whereas Section 2.5 contains three applications. Technique details are in Section 2.6. Extra information of the data can be found in the Appendix.

2.1 The Constrained Matrix Factor Model

2.1.1 The Model

For consistency in notation, we adopt the following conventions. A bold capital letter $\boldsymbol{A}$ represents a matrix, a bold lower letter $\boldsymbol{a}$ represents a column vector, and a lower letter a represents a scalar. The j-th column vector and the k-th row vector of the matrix $\boldsymbol{A}$ are denoted by $A_{\cdot j}$ and $A_{k\cdot}$, respectively.

Let $\{\boldsymbol{Y}_t\}_{t=1,\ldots,T}$ be a matrix-variate time series, where $\boldsymbol{Y}_t$ is a $p_1 \times p_2$ matrix, that is

$$\boldsymbol{Y}_t = (Y_{\cdot 1,t}, \cdots, Y_{\cdot p_2,t}) = \begin{pmatrix} Y'_{1\cdot,t} \\ \vdots \\ Y'_{p_1\cdot,t} \end{pmatrix} = \begin{pmatrix} y_{11,t} & \cdots & y_{1p_2,t} \\ \vdots & \ddots & \vdots \\ y_{p_1 1,t} & \cdots & y_{p_1 p_2,t} \end{pmatrix}.$$

Wang et al. (2017) propose the following factor model for $\boldsymbol{Y}_t$,

$$\boldsymbol{Y}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t \boldsymbol{\Gamma}' + \boldsymbol{U}_t, \qquad t = 1, 2, \ldots, T, \tag{2.1}$$

where $\boldsymbol{F}_t$ is a $k_1 \times k_2$ unobserved matrix-variate time series of common fundamental factors, $\boldsymbol{\Lambda}$ is a $p_1 \times k_1$ row loading matrix, $\boldsymbol{\Gamma}$ is a $p_2 \times k_2$ column loading matrix, and $\boldsymbol{U}_t$ is a $p_1 \times p_2$ matrix of random errors. In Equation (2.1), $(\boldsymbol{\Lambda}, \boldsymbol{\Gamma})$ and $(c\boldsymbol{\Lambda}, \boldsymbol{\Gamma}/c)$ are equivalent if $c \neq 0$.

In Model (2.1), we assume that $vec(\boldsymbol{U}_t) \sim WN(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ and is independent of the factor process $vec(\boldsymbol{F}_t)$. That is, $\{\boldsymbol{U}_t\}_{t=1}^T$ is a white noise matrix-variate time series and the common fundamental factors $\boldsymbol{F}_t$ drive all dynamics and co-movement of $\boldsymbol{Y}_t$. $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ reflect the importance of common factors and their interactions. Wang et al. (2017) provide several interpretations of the loading matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$. Essentially, $\boldsymbol{\Lambda}$ ($\boldsymbol{\Gamma}$) can be viewed as the row (column) loading matrix that reflects how each row (column) in

$\boldsymbol{Y}_t$ depends on the factor matrix $\boldsymbol{F}_t$. The interaction between the row and column is introduced through the multiplication of these terms.

The definition of common factors in Model (2.1) is similar to that of Lam et al. (2011a). This decomposition facilitates model identification in finite samples and simplifies the procedure of model identification and statistical inference. However, under the definition, both the "common factors" defined in the traditional factor models and the serially correlated idiosyncratic components will be identified as factors. This poses challenges to the interpretation of the estimated factors, which are usually of special interest in many applications. Moreover, when the dimensions p_1 and p_2 are sufficiently large, interpretation of the estimated common factors $\widehat{\boldsymbol{F}}_t$ becomes difficult because of the uncertainty and dependence involved in the estimates of the loading matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$.

To mitigate the aforementioned difficulties and, more importantly, to incorporate natural and known constraints among the column and row variables, we consider the following constrained and partially constrained matrix factor models.

A *constrained matrix factor model* can be written as

$$\boldsymbol{Y}_t = \boldsymbol{H}_R \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}' \boldsymbol{H}_C' + \boldsymbol{U}_t, \tag{2.2}$$

where $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ are pre-specified full column-rank $p_1 \times m_1$ and $p_2 \times m_2$ constraint matrices, respectively, and $\boldsymbol{R}$ and $\boldsymbol{C}$ are $m_1 \times k_1$ row loading matrix and $m_2 \times k_2$ column loading matrix, respectively. For meaningful constraints, we assume $k_1 \leq m_1 << p_1$ and $k_2 \leq m_2 << p_2$. Compared with the matrix factor model in (2.1), we set $\boldsymbol{\Lambda} = \boldsymbol{H}_R \boldsymbol{R}$ and $\boldsymbol{\Gamma} = \boldsymbol{H}_C \boldsymbol{C}$ with $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ given. The number of parameters in the left loading matrix $\boldsymbol{R}$ is $m_1 k_1$, smaller than $p_1 k_1$ of the unconstrained model. The number of parameters in the column loading matrix $\boldsymbol{C}$ also decreases from $p_2 k_2$ to $m_2 k_2$. The constraint matrices $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ are constructed based on prior or domain knowledge of the variables. For example, if $\boldsymbol{H}_R$ consists of orthogonal binary vectors, it represents a classification or grouping of the rows of the observed matrix.

Consider a simplified model with only row constraints $\boldsymbol{Y}_t = \boldsymbol{H}_R \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}' + \boldsymbol{U}_t$. If

$$\boldsymbol{H}_R = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \end{bmatrix}', \tag{2.3}$$

we are effectively imposing the constraint that there are two groups of row variables (say countries) in which the 'row' behavior of each variable in a group is the same. Specifically, the model becomes

$$\boldsymbol{Y}_t^{(1)} = \boldsymbol{R}_1 \boldsymbol{F}_t \boldsymbol{C}' + \boldsymbol{U}_t^{(1)} \quad \text{and} \quad \boldsymbol{Y}_t^{(2)} = \boldsymbol{R}_2 \boldsymbol{F}_t \boldsymbol{C}' + \boldsymbol{U}_t^{(2)}$$

where $\boldsymbol{Y}_t^{(1)}$ consists of the first $p_1^{(1)}$ rows of $\boldsymbol{Y}_t$ – all the countries in the first group, and $\boldsymbol{Y}_t^{(2)}$ consists of the rest of the rows in the second group. In this case, $\boldsymbol{R}_1$ is a $1 \times k_1$ row vector that is common to all rows in the first group $\boldsymbol{Y}_t^{(1)}$. Comparing to the general matrix factor model (2.2), the constrained model imposes the constraint that the loading matrix $\boldsymbol{\Lambda}$ have the form $\boldsymbol{\Lambda} = [\boldsymbol{R}_1' \ \cdots \ \boldsymbol{R}_1' \ \boldsymbol{R}_2' \ \cdots \ \boldsymbol{R}_2']'$. The countries within the same group have the same row loadings. Note that the two groups still share the same factor matrix $\boldsymbol{F}_t$ and the same column loading matrix $\boldsymbol{C}$. The two groups related to the global common factor $\boldsymbol{F}_t$ differently. The smaller loading matrix $\boldsymbol{R}$ of dimension $2 \times m_1$, instead of the unconstrained $p_1 \times m_1$ loading matrix, provides a much simpler interpretation. More complicated constraints can be used. See Appendix 2.1.2 for an illustration of some constraint matrices.

If there are two "distinct" sets of constraints and the factors corresponding to these two sets do not interact, Model (2.2) can be extended to a *multiple-term matrix factor model* as

$$\boldsymbol{Y}_t = \boldsymbol{H}_{R_1} \boldsymbol{R}_1 \boldsymbol{F}_{1t} \boldsymbol{C}_1' \boldsymbol{H}_{C_1}' + \boldsymbol{H}_{R_2} \boldsymbol{R}_2 \boldsymbol{F}_{2t} \boldsymbol{C}_2' \boldsymbol{H}_{C_2}' + \boldsymbol{U}_t. \tag{2.4}$$

For example, countries can be grouped according to their geographic locations, such as European and Asian countries, and also grouped according to their economic characteristics, such as natural resource based and manufacture based economies, and the corresponding factors may not interact with each other.

Note that (2.4) can be rewritten as (2.2), with $\boldsymbol{H}_R = \begin{bmatrix} \boldsymbol{H}_{R_1} & \boldsymbol{H}_{R_2} \end{bmatrix}$, $\boldsymbol{H}_C =$

$\begin{bmatrix} \boldsymbol{H}_{C_1} & \boldsymbol{H}_{C_2} \end{bmatrix}$,

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & 0 \\ 0 & \boldsymbol{R}_2 \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 & 0 \\ 0 & \boldsymbol{C}_2 \end{bmatrix}, \text{ and } \boldsymbol{F}_t = \begin{bmatrix} \boldsymbol{F}_{1t} & 0 \\ 0 & \boldsymbol{F}_{2t} \end{bmatrix}.$$

Hence (2.4) is a special case of (2.2) with the strong assumption that the factor matrix is block diagonal. Such a simplification can greatly enhance the interpretation of the model.

Remark 1. The pre-specified constraint matrices $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$ do not have to be orthogonal. Neither does the pair $\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}_{C_2}$. An estimation procedure is presented in Remark 2 in Section 2.2.3. The rates of convergence will change as a result of information loss from the estimation procedure to deal with the nonorthogonality of $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$. Since we can always transform non-orthogonal constraint matrices to some orthogonal constraint matrices, we shall focus on the case when $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$ (or $\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}_{C_2}$) are orthogonal.

In many applications, prior or domain knowledge may not be sufficiently comprehensive or may only provide a partial specification of the constraint matrices. In the above example, it is possible that the countries within a group react to one set of factors the same way, but differently to another set of factors. In such cases, a partially constrained factor model would be more appropriate. Specifically, a *partially constrained matrix factor model* can be written as

$$\boldsymbol{Y}_t = \begin{bmatrix} \boldsymbol{H}_{R_1}\boldsymbol{R}_1 & \boldsymbol{\Lambda}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{11,t} & \boldsymbol{F}_{12,t} \\ \boldsymbol{F}_{21,t} & \boldsymbol{F}_{22,t} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}_1'\boldsymbol{H}_{C_1}' \\ \boldsymbol{\Gamma}_2' \end{bmatrix} + \boldsymbol{U}_t,$$

where $\boldsymbol{H}_{R_1}$, $\boldsymbol{R}_1$, $\boldsymbol{H}_{C_1}$ and $\boldsymbol{C}_1$ are defined similarly as those in (2.4). $\boldsymbol{F}_{ij,t}$'s are common matrix factors corresponding to the interactions of the row and column loading space spanned by the columns of $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ and their complements, $\boldsymbol{\Lambda}_2$ is $p_1 \times q_1$ row loading matrix and $\boldsymbol{\Gamma}_2$ is a $p_2 \times q_2$ column loading matrix. Again, we have $q_1 < p_1$ and $q_2 < p_2$. We further assume that $vec(\boldsymbol{F}_{ij,t})$'s are independent with $vec(\boldsymbol{U}_t)$. $\boldsymbol{H}_{R_1}'\boldsymbol{\Lambda}_2 = \boldsymbol{0}$ and $\boldsymbol{H}_{C_1}'\boldsymbol{\Gamma}_2 = \boldsymbol{0}$, because all the row loadings that are in the space of $\boldsymbol{H}_{R_1}$ and all the column loadings that are in the space of $\boldsymbol{H}_{C_1}$ could be absorbed into the first parts of

loading matrices. Thus, we could explicitly rewrite the model as

$$\boldsymbol{Y}_t = \begin{bmatrix} \boldsymbol{H}_{R_1}\boldsymbol{R}_1 & \boldsymbol{H}_{R_2}\boldsymbol{R}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{11,t} & \boldsymbol{F}_{12,t} \\ \boldsymbol{F}_{21,t} & \boldsymbol{F}_{22,t} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}_1'\boldsymbol{H}_{C_1}' \\ \boldsymbol{C}_2'\boldsymbol{H}_{C_2}' \end{bmatrix} + \boldsymbol{U}_t, \tag{2.5}$$

where $\boldsymbol{H}_{R_2}$ is a $p_1 \times (p_1 - m_1)$ constraint matrix satisfying $\boldsymbol{H}_{R_1}'\boldsymbol{H}_{R_2} = \boldsymbol{0}$, $\boldsymbol{H}_{C_2}$ is a $p_2 \times (p_2 - m_2)$ constraint matrix satisfying $\boldsymbol{H}_{C_1}'\boldsymbol{H}_{C_2} = \boldsymbol{0}$, $\boldsymbol{R}_2$ is $(p_1 - m_1) \times q_1$ row loading matrix, and $\boldsymbol{C}_2$ is a $(p_2 - m_2) \times q_2$ column loading matrix.

In the special case when $\boldsymbol{F}_{21,t} = \boldsymbol{0}$ and $\boldsymbol{F}_{12,t} = \boldsymbol{0}$, model (2.5) can be further simplified as

$$\boldsymbol{Y}_t = \boldsymbol{H}_{R_1}\boldsymbol{R}_1\boldsymbol{F}_{11,t}\boldsymbol{C}_1'\boldsymbol{H}_{C_1}' + \boldsymbol{H}_{R_2}\boldsymbol{R}_2\boldsymbol{F}_{22,t}\boldsymbol{C}_2'\boldsymbol{H}_{C_2}' + \boldsymbol{U}_t. \tag{2.6}$$

Model (2.6) is different from the multi-term model of (2.4) in that the matrix $\boldsymbol{H}_{R_2}$ in (2.5) is induced from $\boldsymbol{H}_{R_1}$ while $\boldsymbol{H}_{R_2}$ in (2.4) is an informative constraint, with a lower dimension.

In the special case when $\boldsymbol{H}_{C_1} = \boldsymbol{I}_{p_1}$ (there is no column constraint), model (2.5) becomes

$$\boldsymbol{Y}_t = \begin{bmatrix} \boldsymbol{H}_{R_1}\boldsymbol{R}_1 & \boldsymbol{H}_{R_2}\boldsymbol{R}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{F}_{1,t} \\ \boldsymbol{F}_{2,t} \end{bmatrix} \boldsymbol{C}' + \boldsymbol{U}_t,$$

where $\boldsymbol{F}_{1,t} = [\boldsymbol{F}_{11,t}, \boldsymbol{F}_{12,t}]$ and $\boldsymbol{F}_{2,t} = [\boldsymbol{F}_{21,t}, \boldsymbol{F}_{22,t}]$. The left loading matrix still spans the entire p_1 dimensional space, but the first part of loading matrix $\boldsymbol{R}_1$ has a clearer interpretation.

The partially constrained matrix factor model (2.5) incorporates partial information $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{C_1}$ in the unconstrained model (2.1) without ignoring the possible remainders. If we include all four matrix factors in the four subspaces divided by the interactions of $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{C_1}$ and their complements, the number of parameters in (2.5) is the same as that in the unconstrained model (2.1). However, as shown by the theorems in Section 2.3, the rates of convergence are much faster than those of the unconstrained matrix factor model. Furthermore, in most applications, inclusion of only two matrix-factor terms is adequate in explaining high percentage of variability, as exemplified by the three applications in Section 2.5.

The benefits of partially constrained matrix factor models are two-folds. Firstly, it is capable of picking up, from the complement space of $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$, the factors that are unknown to researchers. In this case, the dimensions of $\boldsymbol{F}_{22,t}$ are typically much smaller than those of $\boldsymbol{F}_{11,t}$ even though the loading matrices $\boldsymbol{R}_2$ and $\boldsymbol{C}_2$ still have large numbers of rows $(p_1 - m_1)$ and $(p_2 - m_2)$, respectively, since the constraint part should have accommodated the main and key common factors. The spirit is similar to the two-step estimation of Lam and Yao (2012) in which one fits a second-stage factor model to the residuals obtained by subtracting the common part of the first-stage factor model.

The second benefit is that the partially constrained matrix factor model is able to identify matrix factors whose dimensions are completely explained by the pre-specified constraint matrices. Specifically, $\boldsymbol{F}_{11,t}$ represents the factor matrix with row and column factors affecting the observed matrix-variate in the way as specified by the constraints $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ completely. Consider the multinational macroeconomic index example. If $\boldsymbol{H}_R$ is built from the country classification information, how the rows in $\boldsymbol{F}_{11,t}$ affect the observations can be completely explained by the country groups instead of individual countries and the row factors in $\boldsymbol{F}_{11,t}$ have a clearer interpretation related to the classification. In many practical applications, researchers are interested in obtaining specific latent factors related to some domain theories and use these specific factors to predict future values of interest as guided by domain theories. For example, in the yield curve example in Section 2.1.2, economic theory implies that the level, slope, and curvature factors affect the observations in the way specified by, for example, $\boldsymbol{H}_R = [\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3]$, where $\boldsymbol{h}_1 = (1, 1, 1, 1, 1)'$, $\boldsymbol{h}_2 = (1, 1, 0, -1, -1)'$, and $\boldsymbol{h}_3 = (-1, 0, 2, 0, -1)$. Then the estimation method in Section 2.2 is capable of isolating $\boldsymbol{H}_{R_1}\boldsymbol{R}_1\boldsymbol{F}_{11,t}\boldsymbol{C}_1'\boldsymbol{H}_{C_1}'$ and correctly estimating the loadings and the specified level, slope, and curvature factors in the constrained spaces. Thus, the constrained factor model can serve as a method to identify and isolate specific factors suggested by domain theories or prior knowledge.

2.1.2 Constraint Matrix

We first consider discrete **covariate-induced constraint matrices**, using dummy variables. Continuous covariate may be segmented into regimes. As an illustration we consider the following toy example of corporate financial matrix-valued time series. Suppose we have 8 companies, which can be grouped according to their industrial classification (Tech and Retail) and also their market capitalization (Large and Medium). The two groups form 2×2 combinations as shown in Table 2.1,

		Market Cap	
		1. Large	2. Medium
Industry	1. Tech	Apple, Microsoft	Brocade, FireEye
	2. Retail	Walmart, Target	JC Penny, Kohl's

	Industry	Market Cap
Apple	1	1
Microsoft	1	1
Brocade	1	2
FireEye	1	2
Walmart	2	1
Target	2	1
JC Penny	2	2
Kohl's	2	2

Table 2.1: Groups of companies by industry and market cap.

Table 2.2 shows some possible constraint matrices utilizing only industrial classification. To combine both industrial classification and market cap information, we first consider an additive model constraint on the $8 \times k_1$ ($k_1 \leq 3$) loading matrix $\mathbf{\Lambda}$ in model (2.1). The additive model constraint means that the i-th row of $\mathbf{\Lambda}$, that is, the loadings of k_1 row factors on the i-th variable, must have the form $\boldsymbol{\lambda}_{i\cdot} = \boldsymbol{u}_{j\cdot} + \boldsymbol{v}_{l\cdot}$, where the i-th variable falls in group $(Industry_j, MarketCap_l)$, k_1-dimensional vectors $\boldsymbol{u}_{j\cdot}$ and $\boldsymbol{v}_{l\cdot}$ are the loadings of k_1 row factors on the j-th market cap group and l-th industrial group, respectively. The most obvious way to express the additive model constraint is to use row constraints $\boldsymbol{H}_R^{(2)}$ in Table 2.2. Then, in the constrained matrix factor model (2.2), $\boldsymbol{H}_R = \boldsymbol{H}_R^{(2)}$ and $\boldsymbol{R} = (\boldsymbol{u}_{1\cdot}, \boldsymbol{u}_{2\cdot}, \boldsymbol{v}_{1\cdot}, \boldsymbol{v}_{2\cdot})'$.

Further, we consider the constraint incorporating an interaction term between industry and market cap grouping information. Now the i-th row of $\mathbf{\Lambda}$ has the form

$$\boldsymbol{H}_R^{(1)}=\begin{array}{|c|}\hline 1\\1\\1\\1\\\hline \\ \\ \\ \\\hline\end{array}\qquad \boldsymbol{H}_R^{(2)}=\begin{array}{|c|c|}\hline 1&1\\1&1\\1&\\1&\\\hline &1\\&1\\&1\\&1\\\hline\end{array}\qquad \boldsymbol{H}_R^{(3)}=\begin{array}{|c|c|c|}\hline 1&1&1\\1&1&1\\1&&1\\1&&1\\\hline &1&-1\\&1&-1\\&1&1\\&1&1\\\hline\end{array}$$

Table 2.2: Illustration of constraint matrices constructed from grouping information by additive model.

$\boldsymbol{\lambda}_{i\cdot} = \boldsymbol{u}_{j\cdot} + \boldsymbol{v}_{l\cdot} + \alpha_{j,l}\boldsymbol{w}$, where $\boldsymbol{w}$ is the k_1-dimensional interaction vector containing loadings of k_1 row factors and α_{ij} is the interaction term determined by $\boldsymbol{u}_{j\cdot}$ and $\boldsymbol{v}_{l\cdot}$ jointly. For example,

$$\alpha_{j,l} = \begin{cases} 1 & if \quad j = l = 1 \text{ or } 2, \\ -1 & if \quad j = 1, l = 2 \text{ or vice versa.} \end{cases}$$

In this case, for the constrained matrix factor model (2.2), $\boldsymbol{H}_R = \boldsymbol{H}_R^{(3)}$ and $\boldsymbol{R} = (\boldsymbol{u}_{1\cdot}, \boldsymbol{u}_{2\cdot}, \boldsymbol{v}_{1\cdot}, \boldsymbol{v}_{2\cdot}, \boldsymbol{w})'$. Note that $\boldsymbol{H}_R^{(2)}$ and $\boldsymbol{H}_R^{(3)}$ here are not full column rank and can be reduced to a full column rank matrix satisfying the requirement in Section 2.2. But the presentations of $\boldsymbol{H}_R^{(2)}$ and $\boldsymbol{H}_R^{(3)}$ are sufficient to illustrate the ideas of constructing complex constraint matrices.

To illustrate a **theory-induced constraint matrix**, we consider the yield curve latent factors model. Nelson and Siegel (1987) propose the Nelson-Siegel representation of the yield curve using a variation of the three-component exponential approximation to the cross-section of yields at any moment in time,

$$y(\tau) = \beta_1 + \beta_2 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + \beta_3 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right),$$

where $\boldsymbol{y}(\tau)$ denotes the set of zero-coupon yields and τ denotes time to maturity.

Diebold and Li (2006) and Diebold et al. (2006) interpret the Nelson-Siegel representation as a dynamic latent factor model where β_1, β_2, and β_3 are time-varying latent factors that capture the level (L), slope (S), and curvature (C) of the yield curve at each period t, while the terms that multiply the factors are respective factor loadings,

that is

$$y(\tau) = L_t + S_t\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + C_t\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right).$$

The factor L_t may be interpreted as the overall level of the yield curve since its loading is equal for all maturities. The factor S_t, representing the slope of the yield curve, has a maximum loading (equal to 1) at the shortest maturity and then monotonically decays through zero as maturities increase. And the factor C_t has a loading that is 0 at the shortest maturity, increases to an intermediate maturity and then falls back to 0 as maturities increase. Hence, S_t and C_t capture the short-end and medium-term latent components of the yield curve. The coefficient λ controls the rate of decay of the loading of C_t and the maturity where S_t has maximum loading.

Multinational yield curve can be represented as a matrix time series $\{\boldsymbol{Y}_t\}_{t=1,\ldots,T}$, where rows of $\boldsymbol{Y}_t$ represent time to maturity and columns of $\boldsymbol{Y}_t$ denotes countries. To capture the characteristics of loading matrix specific to the level, slope, and curvature factors, we could set row loading constraint matrix to, for example, $\boldsymbol{H}_R = [\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3]$, where $\boldsymbol{h}_1 = (1,1,1,1,1)'$, $bh_2 = (1,1,0,-1,-1)'$ and $\boldsymbol{h}_3 = (-1,0,2,0,-1)$. In Section 2.4, we try to mimic multinational yield curve and generate our samples from this type of constraints.

2.2 Estimation Procedure

Similar to all factor models, identification issue exits in the constrained matrix-variate factor model (2.2). Let $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ be two invertible matrices of size $k_1 \times k_1$ and $k_2 \times k_2$. Then the triples $(\boldsymbol{R}, \boldsymbol{F}_t, \boldsymbol{C})$ and $(\boldsymbol{R}\boldsymbol{O}_1, \boldsymbol{O}_1^{-1}\boldsymbol{F}_t\boldsymbol{O}_2^{-1}, \boldsymbol{O}_2\boldsymbol{C})$ are equivalent under Model (2.2). Here, we may assume that the columns of $\boldsymbol{R}$ and $\boldsymbol{C}$ are orthonormal, that is, $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}_{k_1}$ and $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}_{k_2}$, where $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. Even with these constraints, $\boldsymbol{R}$, $\boldsymbol{F}_t$ and $\boldsymbol{C}$ are not uniquely determined in (2.2), as aforementioned replacement is still valid for any orthonormal $\boldsymbol{O}$. However, the column spaces of the loading matrices $\boldsymbol{R}$ and $\boldsymbol{C}$ are uniquely determined. Hence, in the following sections, we focus on the estimation of the column spaces of $\boldsymbol{R}$ and $\boldsymbol{C}$. We denote the row and column factor loading spaces by $\mathcal{M}(\boldsymbol{R})$ and $\mathcal{M}(\boldsymbol{C})$, respectively. For simplicity, we

suppress the matrix column space notation and use the matrix notation directly.

2.2.1 Orthogonal Constraints

We start with the estimation of the constrained matrix-variate factor model (2.2). The approach follows the ideas of Tsai and Tsay (2010a) and Wang et al. (2017). In what follows, we illustrate the estimation procedure for the column space of $\boldsymbol{R}$. The column space of $\boldsymbol{C}$ can be obtained similarly from the transpose of $\boldsymbol{Y}_t$'s. For ease of representation, we assume that the process $\boldsymbol{F}_t$ has mean $\boldsymbol{0}$, and the observation $\boldsymbol{Y}_t$'s are centered and standardized through out this paper.

Suppose we have orthogonal constraints $\boldsymbol{H}_R'\boldsymbol{H}_R = \boldsymbol{I}_{m_1}$ and $\boldsymbol{H}_C'\boldsymbol{H}_C = \boldsymbol{I}_{m_2}$. Define the transformation $\boldsymbol{X}_t = \boldsymbol{H}_R'\boldsymbol{Y}_t\boldsymbol{H}_C$. It follows from (2.2) that

$$\boldsymbol{X}_t = \boldsymbol{R}\boldsymbol{F}_t\boldsymbol{C}' + \boldsymbol{E}_t, \qquad t = 1, 2, \ldots, T, \tag{2.7}$$

where $\boldsymbol{E}_t = \boldsymbol{H}_R'\boldsymbol{U}_t\boldsymbol{H}_C$.

This transformation projects the observed matrix time series into the constrained space. For example, if $\boldsymbol{H}_R$ is the orthonormal matrix corresponding to the group constraint in (2.3), then $\boldsymbol{H}_R^{'}\boldsymbol{Y}_t$ is a $2\times p_2$ matrix, with the first row being the normalized average of the rows of $\boldsymbol{Y}_t$ in the first group and the second row being that in the second group. Such an operation conveniently incorporates the constraints while reduces the dimension of data matrix from $p_1 \times p_2$ to $m_1 \times m_2$, making the analysis more efficient.

Since $\boldsymbol{E}_t$ remains to be a white noise process, the estimation method in Wang et al. (2017) directly applies to the transformed $m_1 \times m_2$ matrix time series $\boldsymbol{X}_t$ in model (2.7). For completeness, we outline briefly the procedure. See Wang et al. (2017) for details.

To facilitate the estimation, we use the QR decomposition $\boldsymbol{R} = \boldsymbol{Q}_1\boldsymbol{W}_1$ and $\boldsymbol{C} = \boldsymbol{Q}_2\boldsymbol{W}_2$. The estimation of column spaces of $\boldsymbol{R}$ and $\boldsymbol{C}$ is equivalent to the estimation of column spaces of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$. Thus model (2.7) can be re-expressed as

$$\boldsymbol{X}_t = \boldsymbol{R}\boldsymbol{F}_t\boldsymbol{C}' + \boldsymbol{E}_t = \boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2' + \boldsymbol{E}_t, \qquad t = 1, 2, \ldots, T, \tag{2.8}$$

where $\boldsymbol{Z}_t = \boldsymbol{W}_1\boldsymbol{F}_t\boldsymbol{W}_2'$, $\boldsymbol{Q}_1'\boldsymbol{Q}_1 = \boldsymbol{I}_{m_1}$, and $\boldsymbol{Q}_2'\boldsymbol{Q}_2 = \boldsymbol{I}_{m_2}$.

Let h be a positive integer. For $i, j = 1, 2, \ldots, m_2$, define

$$\boldsymbol{\Omega}_{zq,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{Z}_t Q_{2,i\cdot}, \boldsymbol{Z}_{t+h} Q_{2,j\cdot}), \text{ and} \tag{2.9}$$

$$\boldsymbol{\Omega}_{x,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(X_{t,\cdot i}, X_{t+h,\cdot j}), \tag{2.10}$$

which can be interpreted as the auto-cross-covariance matrices at lag h between column i and column j of $\{\boldsymbol{Z}_t\boldsymbol{Q}_2'\}_{t=1,\cdots,T}$ and $\{\boldsymbol{X}_t\}_{t=1,\cdots,T}$, respectively. For $h > 0$, both terms do not involve $\boldsymbol{E}_t$ due to the whiteness condition.

For a fixed $h_0 \geq 1$ satisfying Condition 2 in Appendix 2.6, define

$$\boldsymbol{M} = \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \boldsymbol{\Omega}_{x,ij}(h)\boldsymbol{\Omega}_{x,ij}(h)' = \boldsymbol{Q}_1 \left\{ \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \boldsymbol{\Omega}_{zq,ij}(h)\boldsymbol{\Omega}_{zq,ij}(h)' \right\} \boldsymbol{Q}_1'. \tag{2.11}$$

Under Condition 2 in Appendix 2.6, the rank of $\boldsymbol{M}$ is k_1. Since $\boldsymbol{M}$ and the matrix sandwiched by $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_1'$ are positive definite matrices, Equation (2.11) implies that the eigen-space of $\boldsymbol{M}$ is the same as the column space of $\boldsymbol{Q}_1$. Hence, $\mathcal{M}(\boldsymbol{Q}_1)$ can be estimated by the space spanned by the eigenvectors of the sample version of $\boldsymbol{M}$. The normalized eigenvectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{k_1}$ corresponding to the k_1 nonzero eigenvalues of $\boldsymbol{M}$ are uniquely defined up to a sign change. Thus $\boldsymbol{Q}_1$ is unique defined by $\boldsymbol{Q}_1 = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{k_1})$ up to a sign change. We estimate $\widehat{\boldsymbol{Q}}_1 = (\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_{k_1})$ as a representative of $\mathcal{M}(\boldsymbol{Q}_1)$ or $\mathcal{M}(\boldsymbol{R})$

The estimation procedure is based on the sample version of these quantities. For $h \geq 1$ and a prescribed positive integer h_0, define the sample version of $\boldsymbol{M}$ in (2.11) as the following

$$\widehat{\mathbf{M}} = \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \widehat{\boldsymbol{\Omega}}_{x,ij}(h)\widehat{\boldsymbol{\Omega}}_{x,ij}(h)', \text{ where } \quad \widehat{\boldsymbol{\Omega}}_{x,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} X_{t,\cdot i}X_{t+h,\cdot j}'. \tag{2.12}$$

Then, $\mathcal{M}(\boldsymbol{Q}_1)$ can be estimated by $\mathcal{M}(\widehat{\boldsymbol{Q}}_1)$, where $\widehat{\boldsymbol{Q}}_1 = (\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_{k_1})$ and $\widehat{\boldsymbol{q}}_i$ is an eigenvector of $\widehat{\boldsymbol{M}}$, corresponding to its i-th largest eigenvalue. The $\boldsymbol{Q}_2$ is defined similarly for the column loading matrix $\boldsymbol{C}$ and $\mathcal{M}(\widehat{\boldsymbol{Q}}_2)$ and $\widehat{\boldsymbol{Q}}_2$ can be estimated with the same procedure to to the transpose of $\boldsymbol{X}_t$. Consequently, we estimate the normalized factors and residuals, respectively, by $\widehat{\boldsymbol{Z}}_t = \widehat{\boldsymbol{Q}}_1'\boldsymbol{X}_t\widehat{\boldsymbol{Q}}_2$ and $\widehat{\boldsymbol{U}}_t = \boldsymbol{Y}_t - \boldsymbol{H}_R\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Z}}_t\widehat{\boldsymbol{Q}}_2'\boldsymbol{H}_C'$.

The above estimation procedure assumes that the number of row factors k_1 is known. To determine k_1, Wang et al. (2017) used the eigenvalue ratio-based estimator of Lam and Yao (2012). Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_{m_1} \geq 0$ be the ordered eigenvalues of $\widehat{\boldsymbol{M}}$. The ratio-based estimator for k_1 is defined as

$$\widehat{k}_1 = \arg \min_{1 \leq j \leq K} \frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j},$$

where $k_1 \leq K \leq p_1$ is an integer. In practice we may take $K = p_1/2$.

Although the estimation procedure on the transformed series $\boldsymbol{X}_t$ is exactly the same as that of Wang et al. (2017), the asymptotic properties of the estimator are different due to the transformation, as shown in Section 2.3, and $\boldsymbol{X}_t$ is of lower dimension.

2.2.2 Nonorthogonal Constraints

If the constraint matrix $\boldsymbol{H}_R$ (or $\boldsymbol{H}_C$) is not orthogonal, we can perform column orthogonalization and standardization, similar to that in Tsai and Tsay (2010a). Specifically, we obtain

$$\boldsymbol{H}_R = \boldsymbol{\Theta}_R \boldsymbol{K}_R,$$

where $\boldsymbol{\Theta}_R$ is an orthonormal matrix and $\boldsymbol{K}_R$ is a $m_1 \times m_1$ upper triangular matrix with nonzero diagonal elements. $\boldsymbol{H}_C = \boldsymbol{\Theta}_C \boldsymbol{K}_C$ can be obtained in the same way.

Letting $\boldsymbol{X}_t = \boldsymbol{\Theta}_R' \boldsymbol{Y}_t \boldsymbol{\Theta}_C$, $\boldsymbol{R}^* = \boldsymbol{K}_R \boldsymbol{R}$, and $\boldsymbol{C}^* = \boldsymbol{K}_C \boldsymbol{C}$, we have

$$\boldsymbol{X}_t = \boldsymbol{R}^* \boldsymbol{F}_t \boldsymbol{C}^{*'} + \boldsymbol{E}_t, \qquad t = 1, 2, \ldots, T, \tag{2.13}$$

where $\boldsymbol{E}_t = \boldsymbol{\Theta}_R' \boldsymbol{U}_t \boldsymbol{\Theta}_C$. Since $\boldsymbol{E}_t$ remains to be a white noise process, we apply the same estimation method in Section 2.2.1 to obtain $\widehat{\boldsymbol{Q}}_1^*$ and $\widehat{\boldsymbol{Q}}_2^*$ as the representatives of $\mathcal{M}(\widehat{\boldsymbol{R}}^*)$ and $\mathcal{M}(\widehat{\boldsymbol{C}}^*)$. Then the estimators of $\boldsymbol{R}$ and $\boldsymbol{C}$ are $\widehat{\boldsymbol{R}} = \boldsymbol{K}_R^{-1} \widehat{\boldsymbol{Q}}_1^*$ and $\widehat{\boldsymbol{C}} = \boldsymbol{K}_C^{-1} \widehat{\boldsymbol{Q}}_2^*$. Note that $\boldsymbol{K}_R$ and $\boldsymbol{K}_C$ are invertible lower triangular matrices.

2.2.3 Multi-term Constrained Matrix Factor Model

Without loss of generality, we assume that both row and column constraint matrices are orthogonal matrices. If $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$ (or $\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}_{C_2}$) are orthogonal, we obtain,

for $t = 1, 2, \ldots, T$,

$$\begin{aligned} \boldsymbol{H}'_{R_1}\boldsymbol{Y}_t\boldsymbol{H}_{C_1} &= \boldsymbol{R}_1\boldsymbol{F}_{1,t}\boldsymbol{C}'_1 + \boldsymbol{H}'_{R_1}\boldsymbol{U}_t\boldsymbol{H}_{C_1}, \\ \boldsymbol{H}'_{R_2}\boldsymbol{Y}_t\boldsymbol{H}_{C_2} &= \boldsymbol{R}_2\boldsymbol{F}_{2,t}\boldsymbol{C}'_2 + \boldsymbol{H}'_{R_2}\boldsymbol{U}_t\boldsymbol{H}_{C_2}, \end{aligned}$$

where $\boldsymbol{H}'_{R_1}\boldsymbol{U}_t\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}'_{R_2}\boldsymbol{U}_t\boldsymbol{H}_{C_2}$ are white noises. The estimators of $\widehat{\boldsymbol{R}}_1$, $\widehat{\boldsymbol{C}}_1$, $\widehat{\boldsymbol{F}}_{1,t}$, $\widehat{\boldsymbol{R}}_2$, $\widehat{\boldsymbol{C}}_2$ and $\widehat{\boldsymbol{F}}_{2,t}$ can be obtained by applying the estimation procedure described in Section 2.2.1 to $\boldsymbol{H}'_{R_1}\boldsymbol{Y}_t\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}'_{R_2}\boldsymbol{Y}_t\boldsymbol{H}_{C_2}$, respectively.

Remark 2. For multi-term constrained model (2.4), $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$ (or $\boldsymbol{H}_{C_1}$ and $\boldsymbol{H}_{C_2}$) may not necessarily be orthogonal. In this case, we illustrate the estimation procedure for the column loadings, while the row loading estimators for $\widehat{\boldsymbol{R}}_1$ and $\widehat{\boldsymbol{R}}_2$ can be obtained from the same procedure applied to the transpose of $\boldsymbol{Y}_t$. Define projection matrices $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}} = \boldsymbol{I} - \boldsymbol{H}_{R_1}\boldsymbol{H}'_{R_1}$ and $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}} = \boldsymbol{I} - \boldsymbol{H}_{R_2}\boldsymbol{H}'_{R_2}$, which represent the projections onto the spaces perpendicular to the column spaces of $\boldsymbol{H}_{R_1}$ and $\boldsymbol{H}_{R_2}$, respectively. Left multiplying equations (2.4) by $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}$ and $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}}$, respectively, and taking transpose of the resulting matrices, we have $\boldsymbol{Y}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}} = \boldsymbol{H}_{C_1}\boldsymbol{C}_1\boldsymbol{F}'_{1,t}\boldsymbol{R}'_1\boldsymbol{H}'_{R_1}\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}} + \boldsymbol{U}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}$ and $\boldsymbol{Y}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}} = \boldsymbol{H}_{C_2}\boldsymbol{C}_2\boldsymbol{F}'_{2,t}\boldsymbol{R}'_2\boldsymbol{H}'_{R_2}\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}} + \boldsymbol{U}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}}$, where $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}\boldsymbol{U}_t$ and $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}}\boldsymbol{U}_t$ are white noises. The column loading estimators $\widehat{\boldsymbol{C}}_1$ and $\widehat{\boldsymbol{C}}_2$ can be obtained by applying the procedure described in Section 2.2.1 to $\boldsymbol{H}'_{C_1}\boldsymbol{Y}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}$ and $\boldsymbol{H}'_{C_2}\boldsymbol{Y}'_t\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_1}}$, respectively. Note that the $p_1 \times m_1$ matrix $\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}\boldsymbol{H}_{R_1}$ is no longer full rank or orthonormal. However, the row and column loading spaces and latent factors can be fully recovered if the dimension of the reduced constrained loading spaces still larger than the dimensions of the latent factor spaces. However, the rates of convergence will change. For example, the rate of convergence of $\widehat{\boldsymbol{C}}_1$ will depend on $\|\boldsymbol{P}_{\boldsymbol{H}^{\perp}_{R_2}}\boldsymbol{H}_{R_1}\boldsymbol{R}_1\|_2^2$ instead of $\|\boldsymbol{H}_{R_1}\boldsymbol{R}_1\|_2^2$.

2.2.4 Partially Constrained Matrix Factor Model

For the partially constrained matrix factor model (2.5), we assume that $\boldsymbol{H}'_{R_1}\boldsymbol{H}_{R_2} = \boldsymbol{0}$ and $\boldsymbol{H}'_{C_1}\boldsymbol{H}_{C_2} = \boldsymbol{0}$. Define the transformation $\boldsymbol{X}_t^{(lk)} = \boldsymbol{H}'_{R_l}\boldsymbol{Y}_t\boldsymbol{H}_{C_k}$ for $l, k = 1, 2$. Then the transformed data follow the structure,

$$\boldsymbol{X}_t^{(lk)} = \boldsymbol{R}_l\boldsymbol{F}_{lk,t}\boldsymbol{C}'_k + \boldsymbol{E}_t^{(lk)}, \quad l, k = 1, 2,$$

where $\boldsymbol{E}_t^{(lk)} = \boldsymbol{H}'_{R_l}\boldsymbol{U}_t\boldsymbol{H}_{C_k}$ remains white noise processes.

Let $\boldsymbol{M}^{(lk)}$ represent the $\boldsymbol{M}$ matrix defined in (2.11) for each $\boldsymbol{X}_t^{(lk)}$, $l,k = 1,2$. Define $\boldsymbol{M}^{(l\cdot)} = \sum_{k=1}^{2} \boldsymbol{M}^{(lk)}$ for $l = 1,2$, then

$$\boldsymbol{M}^{(l\cdot)} = \boldsymbol{Q}_1^{(l)} \left\{ \sum_{k=1}^{2} \sum_{h=1}^{h_0} \sum_{i=1}^{m_2} \sum_{j=1}^{m_2} \boldsymbol{\Omega}_{zq,ij}^{(lk)}(h) \boldsymbol{\Omega}_{zq,ij}^{(lk)}(h)' \right\} \boldsymbol{Q}_1^{(l)'}, \quad l = 1,2, \tag{2.14}$$

has the same column space as that of $\boldsymbol{R}_l$, for $l = 1,2$, respectively.

The estimators of $\widehat{\boldsymbol{R}}_l$, $l = 1,2$, can be obtained by applying eigen-decomposition on the sample version of $\boldsymbol{M}^{(l\cdot)}$ defined similarly to (2.12). $\boldsymbol{C}_k$, $k = 1,2$, can be obtained by using the same procedure on the transposes of $\boldsymbol{X}_t^{(lk)}$ for $l,k = 1,2$. In the special case of model (2.6) if $\boldsymbol{F}_{21,t} = \boldsymbol{0}$ and $\boldsymbol{F}_{12,t} = \boldsymbol{0}$, the above estimation is essentially the same procedure as those described in Section 2.2.1 applying to $\boldsymbol{X}_t^{(ll)}$ for $l = 1,2$.

This procedure effectively projects the observed matrix time series $\boldsymbol{Y}_t$ into four orthogonal subspaces, based on the constraints obtained from the domain knowledge or some empirical procedure. Because $\boldsymbol{X}_t^{(lk)}, l,k = 1,2$ are orthogonal, they can be analyzed separately. In our setting, we divide a $p_1 \times p_1$ row loading matrix space into two orthogonal $p_1 \times m_1$ and $p_1 \times (p_1 - m_1)$ subspaces. The estimation procedure for the partially constrained model ensures the structural requirement that $\boldsymbol{X}_t^{(l1)}$ and $\boldsymbol{X}_t^{(l2)}$ share the same row loading matrix for the same l without sacrificing the dimension reduction benefit from column space division. More generally, we could divide the space of loading matrix into more than two parts to accommodate each application. Under this partially constrained model, the orthogonality assumption between $\boldsymbol{F}_{lk,t}, l,k = 1,2$ is not important as all are latent variables.

Remark 3. In situations when the prior or domain knowledge captures most major factors, it is reasonable to assume that m_i grows slower than p_i and the row (column) factor strength of the main factor $\boldsymbol{F}_{11,t}$ is no weaker than that of the remainder factor $\boldsymbol{F}_{22,t}$. Improved estimators of $\widehat{\boldsymbol{R}}_l$, $l = 1,2$, can be obtained by applying eigen-decomposition on the sample version of $\boldsymbol{M}^{(l1)}$ defined similarly to (2.12). Improved estimators of $\widehat{\boldsymbol{C}_k}$, $k = 1,2$, can be obtained by using the same procedure on the transposes of $\boldsymbol{X}_t^{(1k)}$ for $k = 1,2$. Here, the estimation procedure discards the noisy part in (2.14) and results in improved estimators.

2.3 Theoretical Properties

In this section, we present the convergence rates for the estimators under the setting that p_1, p_2, m_1, m_2 and T all go to infinity while the dimensions k_1, k_2 and the structure of the latent factor are fixed over time. In what follows, let $\|\boldsymbol{A}\|_2$, $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}\|_{min}$ denote the spectral, Frobenius norm, and the smallest nonzero singular value of $\boldsymbol{A}$, respectively. When $\boldsymbol{A}$ is a square matrix, we denote by $tr(\boldsymbol{A})$, $\lambda_{max}(\boldsymbol{A})$ and $\lambda_{min}(\boldsymbol{A})$ the trace, maximum and minimum eigenvalues of the matrix $\boldsymbol{A}$, respectively. For two sequences a_N and b_N, we write $a_N \asymp b_N$ if $a_N = O(b_N)$ and $b_N = O(a_N)$.

The asymptotic convergence rates are significantly different from those in Wang et al. (2017) due to the constraints. The results reveal more clearly the impact of the constraints on signals and noises and the interaction between them. We only consider the case of the orthogonal constrained model (2.2). Asymptotic properties of nonorthogonal, multi-term, and partially constrained matrix factor model are trivial extensions.

Several regularity conditions (Conditions 1 to 5) are listed in the Appendix. They are similar to those in Wang et al. (2017) and are used to derive the limiting behavior of (2.12) towards its population version. The following condition requires some discussion.

Condition 6.

Factor Strength. There exist constants δ_1 and δ_2 in $[0, 1]$ such that $\|\boldsymbol{H}_R\boldsymbol{R}\|_2^2 \asymp p_1^{1-\delta_1} \asymp \|\boldsymbol{H}_R\boldsymbol{R}\|_{min}^2$ and $\|\boldsymbol{H}_C\boldsymbol{C}\|_2^2 \asymp p_2^{1-\delta_2} \asymp \|\boldsymbol{H}_C\boldsymbol{C}\|_{min}^2$.

Since only $\boldsymbol{Y}_t$ is observed in model (2.2), how well we can recover the factor $\boldsymbol{F}_t$ from $\boldsymbol{Y}_t$ depends on the 'factor strength' reflected by the coefficients in the row and column factor loading matrices $\boldsymbol{H}_R\boldsymbol{R}$ and $\boldsymbol{H}_C\boldsymbol{C}$. For example, in the case of $\boldsymbol{H}_R\boldsymbol{R} = \boldsymbol{0}$ or $\boldsymbol{H}_C\boldsymbol{C} = \boldsymbol{0}$, $\boldsymbol{Y}_t$ carries no information on $\boldsymbol{F}_t$. In the following, we assume $\|\boldsymbol{F}_t\|$ does not change as p_1, p_2, m_1, and m_2 change.

The rates δ_1 and δ_2 in Condition 6 are called the strength for the row factors and the column factors, respectively. If $\delta_1 = 0$, the corresponding row factors are called strong

factors because Condition 6 implies that the factors have impacts on the majority of p_1 vector time series. The amount of information that observed process $\boldsymbol{Y}_t$ carries about the strong factors increases at the same rate as the number of observations or the amount of noise increases. If $\delta_1 > 0$, the row factors are weak, which means the information contained in $\boldsymbol{Y}_t$ about the factors grows more slowly than the noises introduced as p_1 increases. The smaller the $\delta' s$, the stronger the factors. In the strong factor case, the loading matrix is dense. See Lam et al. (2011a) for further discussions.

If we restrict $\boldsymbol{H}_R$ to be orthonormal, $||\boldsymbol{H}_R\boldsymbol{R}||_2^2 = ||\boldsymbol{R}||_2^2 \asymp p_1^{1-\delta_1}$ and there is an interplay between $\boldsymbol{H}_R$ and $\boldsymbol{R}$ as p_1 increases. In order for $\boldsymbol{H}_R$ to remain orthonormal, when p_1 increases, each element of $\boldsymbol{H}_R$ decreases at the rate of $p_1^{-1/2}$. At the same time, each element of $\boldsymbol{R}$ on average increases $\sqrt{p_1^{1-\delta_1}/m_1}$. The column factor loading $||\boldsymbol{H}_C\boldsymbol{C}||_2^2$ behaves in the same way. As p_1 and p_2 increase, each element of the transformed error $\boldsymbol{E}_t$ remains a growth rate of 1 under Condition 3 (see Lemma 1 in Appendix 2.6), but the dimension of $\boldsymbol{E}_t$ is $m_1 \times m_2$ which grows at a slower rate than $p_1 \times p_2$. The factor strength is defined in terms of the observed dimension p_1 and p_2 and the overall loading matrices $\boldsymbol{H}_R\boldsymbol{R}$ and $\boldsymbol{H}_C\boldsymbol{C}$, but clearly how m_1 and m_2 increase with p_1, p_2 is also important because it controls the signal-noise ratio in the constrained model. For example, if $m_i/p_i = c_i < 1$, $i = 1, 2$, that is, the number of members in each group is fixed, then $||\boldsymbol{R}||_2^2||\boldsymbol{C}||_2^2 \asymp m_1^{1-\delta_1}m_2^{1-\delta_2}/c_1^{1-\delta_1}c_2^{1-\delta_2}$, compared to $||\boldsymbol{E}_t||_2^2 \asymp m_1m_2$. If $m_i = p_i^{\alpha_i}$, $\alpha_i < 1$, $i = 1, 2$, then $||\boldsymbol{R}||_2^2||\boldsymbol{C}||_2^2 \asymp m_1^{(1-\delta_1)/\alpha_1}m_2^{(1-\delta_2)/\alpha_2}$ compared to $||\boldsymbol{E}_t||_2^2 \asymp m_1m_2$. Since $c_i < 1$ and $\alpha_i < 1$, the signal-noise ratio is larger than $m_1^{-\delta_1}m_2^{-\delta_2}$, which is the signal-noise ratio of a unconstrained matrix factor model when $p_1 = m_1$ and $p_2 = m_2$.

We have the following theorems for the constrained matrix factor model. Asymptotic properties for the multi-term and the partially constrained models are similar and can be derived easily.

Theorem 1. *Under Conditions 1-6 and* $m_1p_1^{-1+\delta_1}m_2p_2^{-1+\delta_2}T^{-1/2} = o_p(1)$*, as* m_1*,* p_1*,*

m_2, p_2, and T go to ∞, it holds that

$$\begin{aligned}\|\widehat{\boldsymbol{Q}}_1 - \boldsymbol{Q}_1\|_2 &= O_p\left(\max\left(T^{-1/2}, \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right),\\ \|\widehat{\boldsymbol{Q}}_2 - \boldsymbol{Q}_2\|_2 &= O_p\left(\max\left(T^{-1/2}, \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right).\end{aligned}$$

Remark 4. The convergence rate for the unconstrained model is $\Delta_{pT}^{Q} \equiv p_1^{\delta_1}p_2^{\delta_2}T^{-1/2}$ in Wang et al. (2017). The rates for the constrained model under different relations between m_1m_2 and p_1p_2 are shown in Table 2.3.

	$m_1m_2 \asymp p_1p_2$	$p_1^{1-\delta_1}p_2^{1-\delta_2} \sim O_p(m_1m_2)$	$m_1m_2 \sim O_p(p_1^{1-\delta_1}p_2^{1-\delta_2})$
$O_p(\cdot)$	Δ_{pT}^{Q}	$m_1m_2p_1^{-1}p_2^{-1}\Delta_{pT}^{Q}$	$T^{-1/2}$

Table 2.3: Convergence rate of the loading space estimators.

The rate of convergence in Theorem 1 depends on the growth rate of the ratio between m_1m_2 and $p_1^{1-\delta_1}p_2^{1-\delta_2}$, which can be interpreted as the noise-signal ratio. The smaller the noise-signal ratio, the faster the convergence rate. When $p_1^{1-\delta_1}p_2^{1-\delta_2} \sim O_p(m_1m_2)$, the ratio of the convergence rates between the constrained and unconstrained models is of the order of $m_1m_2p_1^{-1}p_2^{-1}$. For example, when $m_1 = p_1^{\alpha_1}$ and $m_2 = p_2^{\alpha_2}$, the rate is $p_1^{\delta_1+\alpha_1-1}p_2^{\delta_2+\alpha_2-1}T^{-1/2}$, and we achieve a better rate than that of the unconstrained case if $\alpha_1 < 1$ or $\alpha_2 < 1$.

When $m_1m_2 \sim O_p(p_1^{1-\delta_1}p_2^{1-\delta_2})$, we achieve the optimal rate $O_p\left(T^{-1/2}\right)$. Note the unconstrained model can only achieve this rate in the case of strong factor. The constrained model can achieve the optimal rate even in the weak factor case. A special case is when the dimensions of the constrained row and column loading spaces m_1 and m_2 are fixed, the convergence rate is $T^{-1/2}$ regardless of the strength condition. Increases of p_1 or p_2 while keeping m_1 and m_2 fixed amount to increases of the sample points in the constrained spaces. When the constrained spaces are properly specified, the additional information introduced from more sample points will accrue and translate into the transformed signal part in (2.7), but the transformed noise gets canceled out by averaging. The noise-signal ratio $\frac{m_1m_1}{p_1^{1-\delta_1}p_2^{1-\delta_2}}$ goes to zero. However, the convergence rate is still bounded below by the convergence rate of the estimated covariance matrix. When $m_1m_2 \asymp p_1p_2$, the convergence rates of the constrained and unconstrained models

are the same. A special case is when $m_1 = c_1 p_1$ and $m_2 = c_2 p_2$, that is, the dimensions of the constrained loading spaces increase with p's linearly.

Remark 5. Under some conditions the convergence rates in Theorem 1 may improve significantly. For example, if $\Sigma_u \equiv Var(vec(\boldsymbol{U}_t))$ is diagonal (i.e. $U_{t,ij}$ and $U_{t,lk}$ are uncorrelated for $(i,j) \neq (l,k)$) and if we have the grouping constraints, then each elements in $\boldsymbol{E}_t$ is a group average. $Var(\boldsymbol{E}_{t,ij})$ is smaller by a factor of $\frac{m_1 m_2}{p_1 p_2}$ and goes to zero when $\frac{m_1 m_2}{p_1 p_2} = o_p(1)$.

Remark 6. The strengths of row factors and column factors δ_1 and δ_2 determine the convergence rate jointly. An increase in the strength of row factors is able to improve the estimation of the column factors loading space and vice versa.

Theorem 2. *Under Conditions 1-6, and if* $m_1 p_1^{-1+\delta_1} m_2 p_2^{-1+\delta_2} T^{-1/2} = o_p(1)$ *and the* $\boldsymbol{M}$ *matrix has* k_1 *distinct positive eigenvalues, then the eigenvalues* $\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{m_1}\}$ *of* $\widehat{\boldsymbol{M}}$*, sorted in the descending order, satisfy*

$$|\widehat{\lambda}_j - \lambda_j| = O_p\left(\max\left(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}\right) \cdot T^{-1/2}\right), \quad for \quad j = 1, 2, \ldots, k_1,$$

$$|\widehat{\lambda}_j| = O_p\left(\max\left(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1^2 m_2^2\right) \cdot T^{-1}\right), \quad for \quad j = k_1 + 1, \ldots, m_1,$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_{m_1}$ *are the eigenvalues of* $\boldsymbol{M}$*.*

Theorem 2 shows that the estimators of the nonzero eigenvalues of $\boldsymbol{M}$ converge more slowly than those of the zero eigenvalues. This provides the theoretical support for the ratio-based estimator of the number of factors described in Section 2.2.1. The assumption that $\boldsymbol{M}$ has k_1 distinct positive eigenvalues is not essential, yet it substantially simplifies the presentation and the proof of the convergence properties.

The convergence rates for the unconstrained model are $\Delta_{pT}^{\lambda} \equiv p_1^{2-\delta_1} p_2^{2-\delta_2} T^{-1/2}$ for the non-zero eigenvalues and $p_1^{\delta_1} p_2^{\delta_2} T^{-1/2} \cdot \Delta_{pT}^{\lambda}$ for the zero eigenvalues, respectively. See Wang et al. (2017). The rates for the constrained model under different relations between $m_1 m_2$ and $p_1 p_2$ are shown in Table 2.4.

In the cases of strong factors or wake factors with $m_1 m_2 \asymp p_1 p_2$, our result is the same as that of Wang et al. (2017). In all other cases, the gap between the convergence rates of nonzero and zero eigenvalues of $\boldsymbol{M}$ is larger in the constrained case.

$O_p(\cdot)$	$m_1m_2 \asymp p_1p_2$	$p_1^{1-\delta_1}p_2^{1-\delta_2} \sim O_p(m_1m_2)$	$m_1m_2 \sim O_p(p_1^{1-\delta_1}p_2^{1-\delta_2})$
Zero	$p_1^{\delta_1}p_2^{\delta_2}T^{-1/2} \cdot \Delta_{pT}^{\lambda}$	$(\frac{m_1m_2}{p_1p_2})^2 p_1^{\delta_1}p_2^{\delta_2}T^{-1/2} \cdot \Delta_{pT}^{\lambda}$	$p_1^{-\delta_1}p_2^{-\delta_2}T^{-1/2} \cdot \Delta_{pT}^{\lambda}$
Non-zero	Δ_{pT}^{λ}	$m_1m_2p_1^{-1}p_2^{-1} \cdot \Delta_{pT}^{\lambda}$	$p_1^{-\delta_1}p_2^{-\delta_2} \cdot \Delta_{pT}^{\lambda}$
Ratio	$p_1^{\delta_1}p_2^{\delta_2}T^{-1/2}$	$m_1m_2p_1^{-1+\delta_1}p_2^{-1+\delta_2}T^{-1/2}$	$T^{-1/2}$

Table 2.4: Convergence rate of estimators for non-zero and zero eigenvalues of $\boldsymbol{M}$.

Let $\boldsymbol{S}_t$ be the dynamic signal part of $\boldsymbol{Y}_t$, i.e. $\boldsymbol{S}_t = \boldsymbol{H}_R\boldsymbol{R}\boldsymbol{F}_t\boldsymbol{C}'\boldsymbol{H}_C' = \boldsymbol{H}_R\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2'\boldsymbol{H}_C'$. From the discussion in Section 2.2.1, $\boldsymbol{S}_t$ can be estimated by

$$\widehat{\boldsymbol{S}}_t = \boldsymbol{H}_R\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Z}}_t\widehat{\boldsymbol{Q}}_2'\boldsymbol{H}_C'.$$

Some theoretical properties of $\widehat{\boldsymbol{S}}_t$ are given below:

Theorem 3. *Under Conditions 1-6 and* $m_1p_1^{-1+\delta_1}m_2p_2^{-1+\delta_2}T^{-1/2} = o_p(1)$, *we have*

$$\begin{aligned}\frac{1}{\sqrt{p_1p_2}}\|\widehat{\boldsymbol{S}}_t - \boldsymbol{S}_t\|_2 = &\quad O_p\left(\max\left(p_1^{-\delta_1/2}p_2^{-\delta_2/2}, \quad m_1p_1^{-1+\delta_1/2}m_2p_2^{-1+\delta_2/2}\right)\cdot\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p_1p_2}}\right),\\ = &\begin{cases} O_p\left(p_1^{-\delta_1/2}p_2^{-\delta_2/2}T^{-1/2} + p_1^{-1/2}p_2^{-1/2}\right), \textit{ if } m_1m_2 \sim O_p(p_1^{1-\delta_1}p_2^{1-\delta_2}),\\ O_p\left(m_1p_1^{-1+\delta_1/2}m_2p_2^{-1+\delta_2/2}T^{-1/2} + p_1^{-1/2}p_2^{-1/2}\right), \textit{ otherwise.}\end{cases}\end{aligned}$$

Theorem 3 shows that as long as m_1m_2 increases slower than p_1p_2 does, we get a faster convergence rate than $O_p\left(p_1^{\delta_1/2}p_2^{\delta_2/2}T^{-1/2} + p_1^{-1/2}p_2^{-1/2}\right)$ – the convergence rate of the unconstrained model in Wang et al. (2017). Note that the estimation of the loading spaces are consistent with fixed p_1 and p_2 in Theorem 1. But the consistency of the signal estimate requires $p_1, p_2 \to \infty$.

As noted in Section 2.2, the row and column factor loading matrices $\boldsymbol{\Lambda} = \boldsymbol{H}_R\boldsymbol{R}$ and $\boldsymbol{\Gamma} = \boldsymbol{H}_C\boldsymbol{C}$ are only identifiable up to a linear space spanned by its columns. Following Lam et al. (2011a) and Wang et al. (2017), we adopt the discrepancy measure used by Chang et al. (2015): for two orthogonal matrices $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ of size $p \times q_1$ and $p \times q_2$, then the difference between the two linear spaces $\mathcal{M}(\boldsymbol{O}_1)$ and $\mathcal{M}(\boldsymbol{O}_2)$ is measured by

$$\mathcal{D}(\mathcal{M}(\boldsymbol{O}_1), \mathcal{M}(\boldsymbol{O}_2)) = \left(1 - \frac{1}{\max(q_1, q_2)}tr\left(\boldsymbol{O}_1\boldsymbol{O}_1'\boldsymbol{O}_2\boldsymbol{O}_2'\right)\right)^{1/2}. \tag{2.15}$$

Clearly, $\mathcal{D}(\mathcal{M}(\boldsymbol{O}_1), \mathcal{M}(\boldsymbol{O}_2))$ assumes values in [0,1]. It equals to 0 if and only if $\mathcal{M}(\boldsymbol{O}_1) = \mathcal{M}(\boldsymbol{O}_2)$ and equals to 1 if and only if $\mathcal{M}(\boldsymbol{O}_1) \perp \mathcal{M}(\boldsymbol{O}_2)$. If $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ are vectors, (2.15) is the cosine similarity measure. The following Theorem 4 shows that the error in estimating loading spaces goes to zero as p_1, p_2 and T go to infinity and the convergence rate is of the same order as that for estimated $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$.

Theorem 4. *Under Conditions 1-6 and if* $m_1 p_1^{-1+\delta_1} m_2 p_2^{-1+\delta_2} T^{-1/2} = o_p(1)$*, then*

$$\begin{aligned} \mathcal{D}(\mathcal{M}(\widehat{\mathbf{\Lambda}}), \mathcal{M}(\mathbf{\Lambda})) &= \mathcal{D}(\mathcal{M}(\widehat{\mathbf{\Gamma}}), \mathcal{M}(\mathbf{\Gamma})) \\ &= O_p\left(\max\left(T^{-1/2}, \frac{m_1}{p_1^{1-\delta_1}} \frac{m_2}{p_2^{1-\delta_2}} T^{-1/2}\right)\right). \end{aligned}$$

Asymptotic theories for estimators of nonorthogonal, multi-term constrained factors model are trivial extensions of the above properties for the orthogonal constrained model.

2.4 Simulations

In this section, we use simulation to study the performance of the estimation methods of Section 2.2 in finite samples. We also compare the results with those of unconstrained models. We employ data generating models under orthogonal full and partial constraints, respectively. In the simulation, we use the Student-t distribution with 5 degrees of freedom to generate the entries in the disturbances $\boldsymbol{U}_t$. Using Gaussian noise shows similar results.

2.4.1 Case 1. Orthogonal Constraints

In this case, the observed data $\boldsymbol{Y}_t$'s are generated according to Model (2.2),

$$\boldsymbol{Y}_t = \boldsymbol{H}_R \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}' \boldsymbol{H}_C' + \boldsymbol{U}_t, \qquad t = 1, \ldots, T,$$

under the following simulation design.

The latent factor process $\boldsymbol{F}_t$ is of dimension $k_1 \times k_2 = 3 \times 2$. The entries of $\boldsymbol{F}_t$ follow $k_1 k_2$ independent $AR(1)$ processes with Gaussian white noise $\mathcal{N}(0,1)$ innovations. Specifically, $vec(\boldsymbol{F}_t) = \mathbf{\Phi}_F \, vec(\boldsymbol{F}_{t-1}) + \boldsymbol{\epsilon}_t$ with $\mathbf{\Phi}_F = diag(-0.5, 0.6, 0.8, -0.4, 0.7, 0.3)$. The dimensions of the constrained row and column loading spaces are $m_1 = 12$ and $m_2 = 3$, respectively. Hence, $\boldsymbol{R}$ is 12×3 and $\boldsymbol{C}$ is 3×2. The entries of $\boldsymbol{R}$ and $\boldsymbol{C}$ are independently sampled from the uniform distribution $U(-p_i^{-\delta_i/2}\sqrt{m_i/p_i},\ p_i^{-\delta_i/2}\sqrt{m_i/p_i})$ for $i = 1, 2$, respectively, so that the condition on the factor strength is satisfied. The disturbance $\boldsymbol{U}_t = \mathbf{\Psi}^{1/2} \boldsymbol{\Xi}_t$ is a white noise process, where the elements of $\boldsymbol{\Xi}_t$ are independent random variables of Student-t distribution with five degrees of freedom and

the matrix $\mathbf{\Psi}^{1/2}$ is chosen so that $\boldsymbol{U}_t$ has a Kronecker product covariance structure $cov(vec(\boldsymbol{U}_t)) = \mathbf{\Gamma}_2 \otimes \mathbf{\Gamma}_1$, where $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are of size $p_1 \times p_1$ and $p_2 \times p_2$ respectively. For $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$, the diagonal elements are 1 and the off-diagonal elements are 0.2.

The effects of factor strength are investigated by varying factor strength parameter (δ_1, δ_2) among $(0,0)$, $(0.5,0)$, $(0.5,0.5)$. For each pair of δ_i's, the dimensions (p_1, p_2) are chosen to be $(20,20)$, $(20,40)$, $(40,20)$ and $(40,40)$. The sample sizes T are $0.5p_1p_2$, p_1p_2, $1.5p_1p_2$ and $2p_1p_2$. For each combination of the parameters, we use 500 realizations. And we use $h_0 = 1$ for all simulations. Estimation error of $\mathcal{M}(\widehat{\boldsymbol{Q}}_i)$ is defined as $\mathcal{D}(\widehat{\boldsymbol{Q}}_i, \boldsymbol{Q}_i)$, where the distance $\mathcal{D}$ is defined in (2.15).

The row constraint matrix $\boldsymbol{H}_R$ is a $p_1 \times 12$ orthogonal matrix. For $p_1 = 20$, $\boldsymbol{H}_R$ is assumed to be a block diagonal matrix $\boldsymbol{I}_4 \otimes \boldsymbol{D}$, where $\boldsymbol{I}_k$ is the identify matrix of dimension k and $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3]$ is a 5×3 matrix with $\boldsymbol{d}_1' = (1,1,1,1,1)/\sqrt{5}$, $\boldsymbol{d}_2' = (-1,-1,0,1,1)/2$, $\boldsymbol{d}_3' = (-1,0,2,0,-1)/\sqrt{6}$. These three $\boldsymbol{d}_j$ vectors can be viewed as the level, slope and curvature, respectively, of a group of five variables. Therefore, the 20 rows are divided into 4 groups of size 5. When we increase p_1 to 40 while keeping $m_1 = 12$ fixed, we double the length of each vector in the columns of $\boldsymbol{D}$, using $\boldsymbol{d}_1' = (1,1,1,1,1,1,1,1,1,1)/\sqrt{10}$, $\boldsymbol{d}_2' = (-1,-1,-1,-1,0,0,1,1,1,1)/\sqrt{8}$ and $\boldsymbol{d}_3' = (-1,-1,0,0,2,2,0,0,-1,-1)/\sqrt{12}$.

The column constraint matrix $\boldsymbol{H}_C$ is a $p_2 \times 3$ orthogonal matrix. For $p_2 = 20$, the three columns of $\boldsymbol{H}_C$ are generated as $\boldsymbol{h}_{c,1} = [\mathbf{1}_7/\sqrt{7}, \mathbf{0}_7, \mathbf{0}_6]'$, $\boldsymbol{h}_{c,2} = [\mathbf{0}_7, \mathbf{1}_7/\sqrt{7}, \mathbf{0}_6]'$, $\boldsymbol{h}_{c,3} = [\mathbf{0}_7, \mathbf{0}_7, \mathbf{1}_6/\sqrt{6}]'$, where $\mathbf{0}_k$ denotes a k-dimensional zero row vector. The constraints represent a 3-group classification. The 20 columns are divided into 3 groups of size 7, 7 and 6 respectively. In increasing p_2 to 40 while keeping $m_2 = 3$ fixed, we double the length of each vector in the columns defined above.

Table 2.5 shows the performance of estimating the true number of factors. We compare the total number of estimated factors $\widehat{k} = \widehat{k}_1\widehat{k}_2$ with the true value $k = k_1k_2 = 6$. The subscripts c and u denote results from the constrained model (2.2) and unconstrained model (2.1), respectively. f_c and f_u denote the relative frequency of correctly estimating the true number of factors k. From the table, we make the following observations. First, when the row and column factors are strong, i.e. $(\delta_1, \delta_2) = (0,0)$,

both constrained and unconstrained models can estimate accurately the number of factors, but the constrained models fare better when the sample size is small. Second, if the strength of the row factors is weak, but the strength of the column factors is strong, i.e. $(\delta_1, \delta_2) = (0.5, 0)$, the unconstrained models fail to estimate the number of factors, but the constrained models continue to perform well. Furthermore, as expected, the performance of the constrained models improves with the sample size. Finally, if the strength of the row and columns factors is weak, i.e. $(\delta_1, \delta_2) = (0.5, 0.5)$, both models encounter difficulties in estimating the correct number of factors for the sample sizes used. This is not surprising as weak signals are hard to detect in general.

				$T = 0.5\,p_1\,p_2$		$T = p_1\,p_2$		$T = 1.5\,p_1\,p_2$		$T = 2\,p_1\,p_2$	
δ_1	δ_2	p_1	p_2	f_u	f_c	f_u	f_c	f_u	f_c	f_u	f_c
0	0	20	20	0.29	0.95	0.77	1	0.95	1	0.99	1
		20	40	0.77	1	0.99	1	1	1	1	1
		40	20	0.81	1	1	1	1	1	1	1
		40	40	1	1	1	1	1	1	1	1
0.5	0	20	20	0	0.2	0	0.49	0	0.78	0	0.92
		20	40	0	0.68	0	0.96	0	0.99	0	1
		40	20	0	0.37	0	0.78	0	0.92	0	0.97
		40	40	0	0.86	0	0.98	0	0.99	0	1
0.5	0.5	20	20	0	0.05	0	0.02	0	0.02	0	0.01
		20	40	0	0.03	0	0.02	0	0.01	0	0
		40	20	0	0.05	0	0.01	0	0	0	0.01
		40	40	0	0.05	0	0	0	0.01	0	0.04

Table 2.5: Relative frequencies of correctly estimating the number of factors k in the case of orthogonal constraints, where p_i are the dimension, T is the sample size, and f_u and f_c denote the results of unconstrained and constrained factor model, respectively.

Figure 2.1 shows the box-plots of the estimation errors in estimating the loading spaces of $\boldsymbol{Q} = \boldsymbol{Q}_2 \otimes \boldsymbol{Q}_1$ using the correct number of factors. The gray boxes are for the constrained models. From the plots, it is seen that when both row and column factors are strong, i.e. $(\delta_1, \delta_2) = (0, 0)$, and the number of factors is properly estimated, the mean and standard deviation of the estimation errors $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$ are small for both models, but the constrained model has a smaller mean estimation error. When row factors are weak, i.e. $(\delta_1, \delta_2) = (0.5, 0)$, and the true number of factors is used, the estimation error of constrained models remains small whereas that of the unconstrained models is substantially larger.

Table 2.6 shows the mean and standard deviations of the estimation errors $\mathcal{D}(\widehat{\boldsymbol{Q}}_i, \boldsymbol{Q}_i)$

for row ($i = 1$) and column ($i = 2$) loading spaces separately for the constrained model (2.2). Column loading spaces are estimated with higher accuracy because the number of column constraints ($p_1 - m_1$) is larger than the number of row constraints ($p_2 - m_2$). From the table, we see that (a) the mean of estimation errors decreases, as expected, as the sample size increases and (b) the mean of estimation errors is inversely proportional to the strength of row factors.

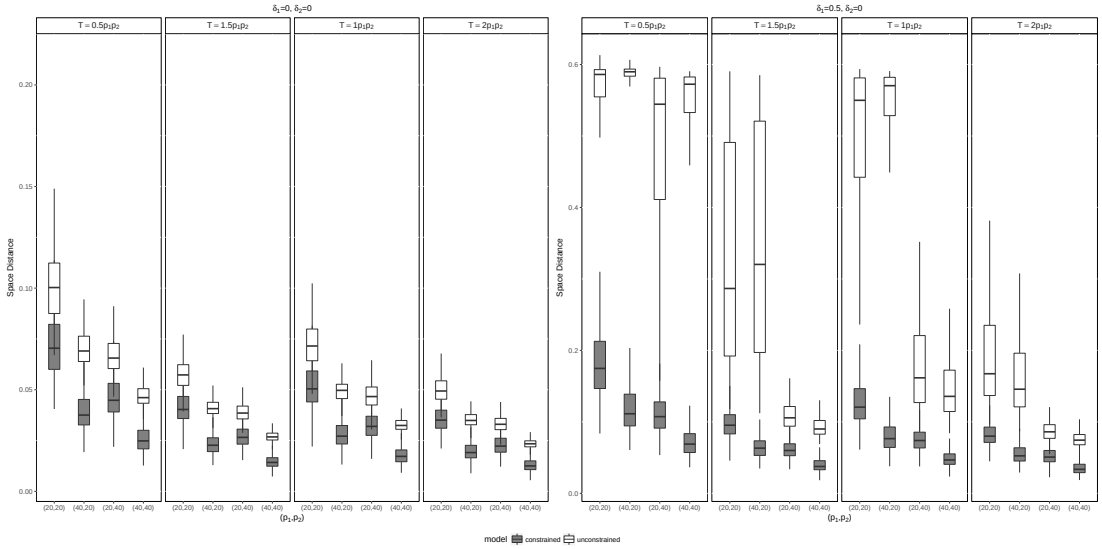


Figure 2.1: Box-plots of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$ for the case of orthogonal constraints. Gray boxes represent the constrained model. The results are based on 500 iterations. See Table 2.17 in Appendix 2.7.2 for plotted values.

				$T = 0.5\,p_1\,p_2$		$T = p_1\,p_2$		$T = 1.5\,p_1\,p_2$		$T = 2\,p_1\,p_2$	
δ_1	δ_2	p_1	p_2	$\mathcal{D}(\widehat{Q}_1, Q_1)$	$\mathcal{D}(\widehat{Q}_2, Q_2)$	$\mathcal{D}(\widehat{Q}_1, Q_1)$	$\mathcal{D}(\widehat{Q}_2, Q_2)$	$\mathcal{D}(\widehat{Q}_1, Q_1)$	$\mathcal{D}(\widehat{Q}_2, Q_2)$	$\mathcal{D}(\widehat{Q}_1, Q_1)$	$\mathcal{D}(\widehat{Q}_2, Q_2)$
0	0	20	20	0.71(0.18)	0.13(0.07)	0.51(0.13)	0.09(0.05)	0.41(0.09)	0.07(0.04)	0.35(0.07)	0.06(0.03)
		20	40	0.46(0.11)	0.08(0.04)	0.32(0.07)	0.05(0.03)	0.27(0.06)	0.04(0.02)	0.23(0.05)	0.04(0.02)
		40	20	0.40(0.12)	0.07(0.04)	0.28(0.07)	0.05(0.03)	0.23(0.06)	0.04(0.02)	0.19(0.05)	0.04(0.02)
		40	40	0.26(0.07)	0.04(0.02)	0.18(0.04)	0.03(0.02)	0.14(0.04)	0.03(0.01)	0.13(0.03)	0.02(0.01)
0.5	0	20	20	1.84(0.75)	0.5(0.23)	1.23(0.35)	0.30(0.15)	0.95(0.23)	0.22(0.11)	0.81(0.18)	0.17(0.09)
		20	40	1.08(0.30)	0.26(0.13)	0.74(0.18)	0.15(0.08)	0.61(0.14)	0.12(0.06)	0.52(0.12)	0.10(0.05)
		40	20	1.18(0.45)	0.28(0.15)	0.78(0.23)	0.17(0.09)	0.64(0.18)	0.13(0.07)	0.54(0.14)	0.11(0.06)
		40	40	0.71(0.21)	0.14(0.08)	0.48(0.13)	0.09(0.05)	0.39(0.1)	0.07(0.04)	0.35(0.09)	0.06(0.03)
0.5	0.5	20	20	5.84(0.62)	2.04(0.53)	5.35(0.75)	1.63(0.42)	4.68(1.17)	1.33(0.34)	4.20(1.31)	1.13(0.32)
		20	40	5.62(0.68)	1.98(0.40)	4.75(1.13)	1.47(0.30)	3.96(1.33)	1.18(0.27)	3.32(1.35)	0.97(0.24)
		40	20	5.53(0.61)	1.52(0.50)	4.68(1.25)	1.00(0.37)	3.64(1.46)	0.76(0.30)	2.87(1.42)	0.61(0.25)
		40	40	5.01(1.01)	1.32(0.38)	3.64(1.47)	0.84(0.29)	2.62(1.46)	0.61(0.20)	1.98(1.14)	0.49(0.19)

Table 2.6: Means and standard deviations (in parentheses) of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$ for constrained factor models. The case of orthogonal constraints is used. The subscripts 1 and 2 denote row and column, respectively. All numbers in the table are 10 times of the true numbers for clear presentation. The results are based on 500 simulations.

To investigate the performance of estimation under different choices of h_0, which is the number of lags used in (2.11), we change the underlying generating model of $vec(\boldsymbol{F}_t)$ to a VAR(2) process without the lag-1 term, $vec(\boldsymbol{F}_t) = \boldsymbol{\Phi}_F vec(\boldsymbol{F}_{t-2}) + \boldsymbol{\epsilon}_t$. Here we only consider the strong factor setting with $\delta_1 = \delta_2 = 0$ and use the sample size $T = 2p_1p_2$ for each combination of p_1 and p_2. All the other parameters are the same as those in Section 2.4.1. Table 2.7 presents the simulation results. Since $vec(\boldsymbol{F}_t)$, and hence $vec(\boldsymbol{Y}_t)$, has zero auto-covariance matrix at lag 1, $\widehat{\boldsymbol{M}}$ under $h_0 = 1$ contains no information on the signal, and, as expected, both the constrained and unconstrained models fail to correctly estimate the number of factors and the loading space. On the other hand, both models are able to correctly estimate the number of factors when $h_0 > 1$ with the constrained model faring better. The fact that $h_0 = 2, 3, 4$ give very similar results shows that the choice of h_0 does not affect the performance much so long as at least one non-zero auto-covariance matrix is included in the calculation. In practice, one can select h_0 by examining the sample cross-correlation matrices of $\boldsymbol{Y}_t$.

	p_1	p_2	$h_0 = 1$	$h_0 = 2$	$h_0 = 3$	$h_0 = 4$
f_c	20	20	0.12	1.00	1.00	1.00
	20	40	0.16	1.00	1.00	1.00
	40	20	0.12	1.00	1.00	1.00
	40	40	0.22	1.00	1.00	1.00
f_u	20	20	0.00	0.89	0.58	0.43
	20	40	0.00	1.00	1.00	0.95
	40	20	0.00	1.00	1.00	0.97
	40	40	0.00	1.00	1.00	1.00
$\mathcal{D}_c(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$	20	20	2.83(1.13)	0.36(0.07)	0.37(0.07)	0.38(0.08)
	20	40	2.69(1.15)	0.23(0.05)	0.23(0.05)	0.24(0.05)
	40	20	2.54(1.21)	0.20(0.05)	0.20(0.05)	0.21(0.06)
	40	40	2.31(1.17)	0.13(0.03)	0.13(0.03)	0.14(0.04)
$\mathcal{D}_u(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$	20	20	4.37(1.29)	0.51(0.07)	0.53(0.07)	0.53(0.08)
	20	40	4.30(1.30)	0.34(0.04)	0.35(0.04)	0.35(0.04)
	40	20	4.36(1.31)	0.36(0.04)	0.37(0.04)	0.37(0.05)
	40	40	4.34(1.34)	0.24(0.02)	0.24(0.03)	0.25(0.03)

Table 2.7: Performance of estimation under different choices of h_0 when $vec(\boldsymbol{F}_t) = \boldsymbol{\Phi}_{\boldsymbol{F}} vec(\boldsymbol{F}_{t-2}) + \boldsymbol{\epsilon}_t$. Metrics reported are relative frequencies of correctly estimating k, means and standard deviations (in parentheses) of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$. Means and standard deviations are multiplied by 10 for ease in presentation. f_u and f_c denote unconstrained and constrained models.

2.4.2 Case 2. Partial Orthogonal Constraints

In this case, the observed data $\boldsymbol{Y}_t$'s are generated using Model (2.5),

$$\boldsymbol{Y}_t = \boldsymbol{H}_R\boldsymbol{R}_1\boldsymbol{F}_t\boldsymbol{C}_1'\boldsymbol{H}_C' + \boldsymbol{L}_R\boldsymbol{R}_2\boldsymbol{G}_t\boldsymbol{C}_2'\boldsymbol{L}_C' + \boldsymbol{U}_t, \qquad t = 1, \ldots, T.$$

Parameter settings of the first part $\boldsymbol{H}_R\boldsymbol{R}_1\boldsymbol{F}_t\boldsymbol{C}_1'\boldsymbol{H}_C'$ are the same as those in Case 1. The latent factor process $\boldsymbol{G}_t$ is of dimension $q_1 \times q_2 = 5 \times 4$. The entries of $\boldsymbol{G}_t$ follow q_1q_2 independent $AR(1)$ processes with Gaussian white noise $\mathcal{N}(0,1)$ innovations, $vec(\boldsymbol{G}_t) = \boldsymbol{\Phi}_G\, vec(\boldsymbol{G}_{t-1}) + \boldsymbol{\epsilon}_t$ with $\boldsymbol{\Phi}_G$ being a diagonal matrix with entries $(-0.7, 0.5, -0.2, 0.9, 0.1, 0.4,$
$0.6, -0.5, 0.7, 0.7, -0.4, 0.4, 0.4, -0.6, -0.6, 0.6, -0.5, -0.3, 0.2, -0.4)$. The row loading matrix $\boldsymbol{L}_R\boldsymbol{R}_2$ is a 20×5 orthogonal matrix, satisfying $\boldsymbol{H}_R'\boldsymbol{L}_R = \boldsymbol{0}$. The column loading matrix $\boldsymbol{L}_C\boldsymbol{C}_2$ is a 20×4 orthogonal matrix, satisfying $\boldsymbol{H}_C'\boldsymbol{L}_C = \boldsymbol{0}$. The entries of $\boldsymbol{R}_2$ and $\boldsymbol{C}_2$ are random draws from the uniform distribution between $-p_i^{-\eta_i/2}\sqrt{p_i/(p_i - m_i)}$ and $p_i^{-\eta_i/2}\sqrt{p_i/(p_i - m_i)}$ for $i = 1, 2$, respectively, so that the conditions on factor strength are satisfied. Factor strength is controlled by the δ_i's.

Model (2.5) could be written in the following form:

$$\boldsymbol{Y}_t = (\boldsymbol{H}_R\boldsymbol{R}_1\ \boldsymbol{L}_R\boldsymbol{R}_2)\begin{pmatrix}\boldsymbol{F}_t & 0\\ 0 & \boldsymbol{G}_t\end{pmatrix}\begin{pmatrix}\boldsymbol{C}_1'\boldsymbol{H}_C'\\ \boldsymbol{C}_2'\boldsymbol{L}_C'\end{pmatrix} + \boldsymbol{U}_t, \qquad t = 1, \ldots, T.$$

In this form, the true number of factors is $k_0 = (k_1 + r_1)(k_2 + r_2)$ and the true loading matrix is $(\boldsymbol{H}_C\boldsymbol{C}_1\ \boldsymbol{L}_C\boldsymbol{C}_2) \otimes (\boldsymbol{H}_R\boldsymbol{R}_1\ \boldsymbol{L}_R\boldsymbol{R}_2)$. Table 2.8 shows the frequency of correctly estimating k_0 based on 500 iterations. In the table, f_u denotes the frequency of correctly estimating k_0 for unconstrained model. f_{con_1} and f_{con_2} denote the same frequency metric for the first matrix factor $\boldsymbol{F}_t$ and second matrix factor $\boldsymbol{G}_t$ of the constrained model. The number of factors in $\boldsymbol{F}_t$ is estimated with a higher accuracy because the dimension of constrained loading space for $\boldsymbol{F}_t$ is $m_1m_2 = 36$, which is smaller than that for $\boldsymbol{G}_t$, $(p_1 - m_1)(p_2 - m_2) = 136$. The result again confirms the theoretical results in Section 2.3. Note that Table 2.8 only contains selected combinations of factor strength parameters δ_i's $(i = 1, \ldots 4)$. The results of all combinations of factor strength are given in Table 2.18 in Appendix 2.7.2.

Figure 2.2 and Figure 2.3 present box-plots of estimation errors under weak and strong factors from 500 simulations, respectively. Again, the results show that the constrained approach efficiently improves the estimation accuracy. The performance of constrained model is good even in the case of weak factors. Moreover, with stronger signals and larger sample sizes, both approaches increase their estimation accuracy.

						$T = 0.5 * p_1 * p_2$			$T = p_1 * p_2$			$T = 1.5 * p_1 * p_2$			$T = 2 * p_1 * p_2$		
δ_1	δ_2	δ_3	δ_4	p_1	p_2	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}
0	0	0	0	20	20	0	0.94	0	0	1.00	0	0	1.00	0	0.01	1.00	0
				20	40	0	1.00	0	0	1.00	0	0.03	1.00	0	0.19	1.00	0
				40	20	0.15	0.99	1.00	0.81	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
				40	40	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0	0	0.5	0	20	20	0	0.94	0	0	1.00	0	0	1.00	0	0	1.00	0
				20	40	0	1.00	0	0	1.00	0	0	1.00	0	0	1.00	0
				40	20	0	0.99	0.54	0	1.00	0.84	0	1.00	0.97	0	1.00	1.00
				40	40	0	1.00	0.98	0	1.00	1.00	0	1.00	1.00	0	1.00	1.00
0.5	0.5	0.5	0.5	20	20	0	0.07	0	0	0.04	0	0	0.01	0	0	0.01	0
				20	40	0	0.07	0	0	0.02	0	0	0.01	0	0	0.01	0
				40	20	0	0.06	0	0	0.01	0	0	0	0	0	0	0
				40	40	0	0.06	0	0	0	0	0	0	0	0	0.03	0

Table 2.8: Relative frequencies of correctly estimating the number of factors for partially constrained factor models. Full tables including all combinations are presented in Table 2.18 in Appendix 2.7.2.

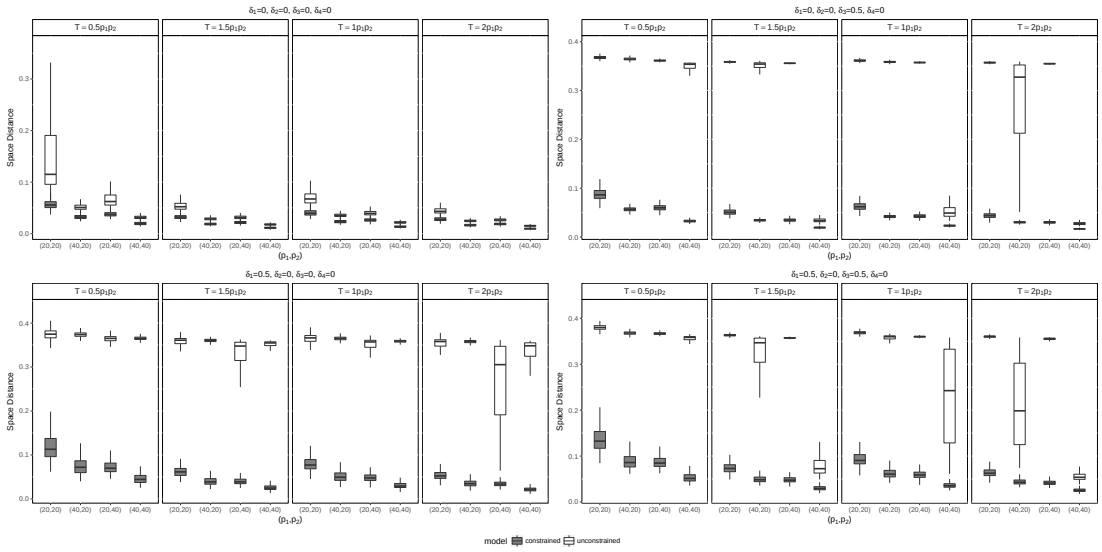


Figure 2.2: The strong factors case. Box-plots of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$ for partially constrained factor models. The gray boxes are for the constrained approach. The results are based on 500 realizations. See Table 2.19 in Appendix 2.7.2 for the plotted values.

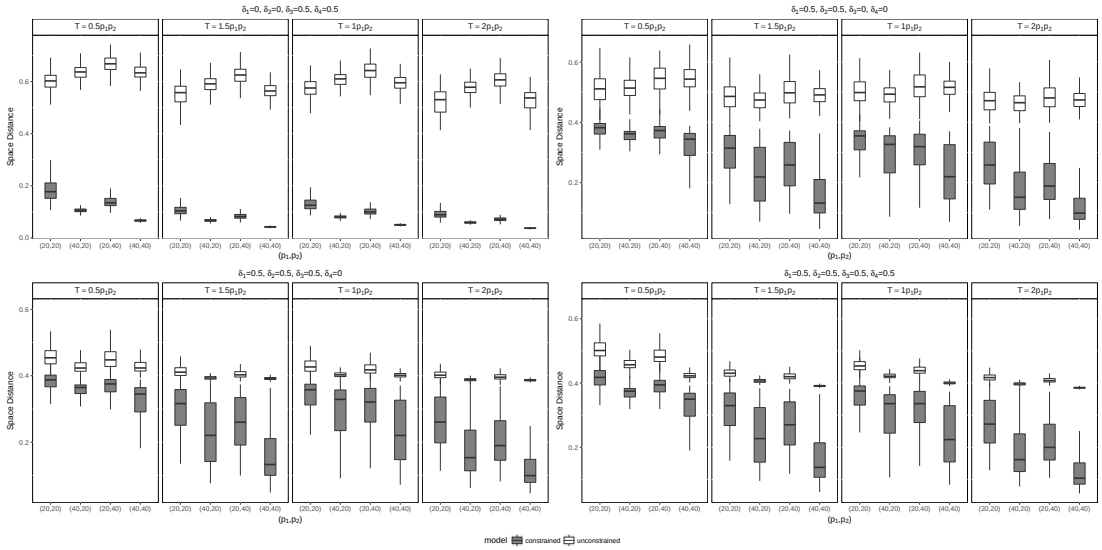


Figure 2.3: The weak factors case. Box-plots of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$ for partially constrained factor models. The gray boxes are for the constrained approach. The results are based on 500 realizations. See Table 2.19 in Appendix 2.7.2 for the plotted values.

2.5 Applications

In this section, we demonstrate the advantages of using constrained matrix-variate factor models with three applications. In practice, the number of common factors (k_1, k_2) and the dimensions of constrained row and column loading spaces (m_1, m_2) must be pre-specified in order to determine an appropriate constrained factor model. The numbers of factors (k_1, k_2) can be determined by any existing methods, such as those in Lam and Yao (2012) and Wang et al. (2017). For any given (k_1, k_2), the dimensions of constrained row and column loading spaces (m_1,m_2) can be determined by either (a) prior or substantive knowledge or (b) an empirical procedure. The results show that even simple grouping information can substantially increase the accuracy in estimation.

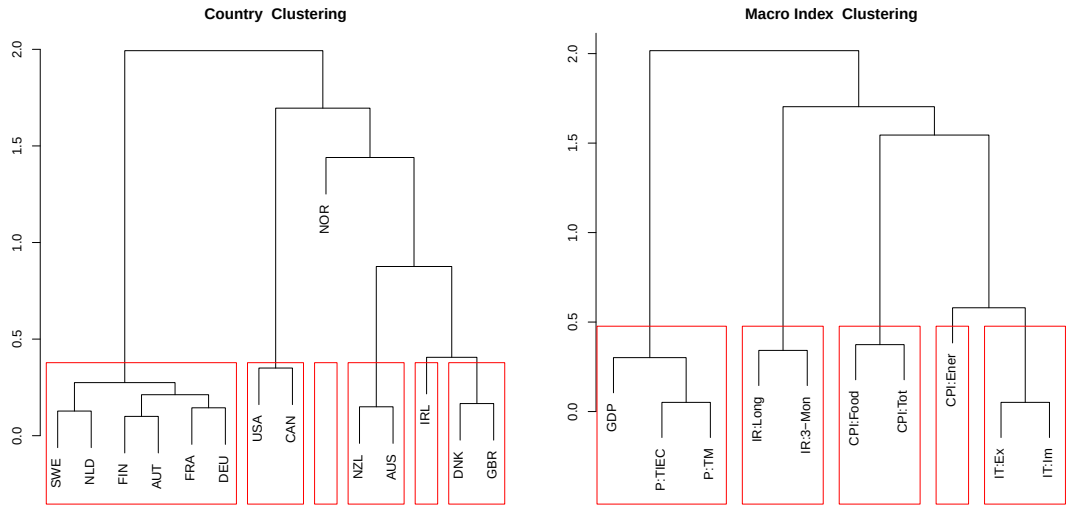
2.5.1 Example 1: Multinational Macroeconomic Indices

We apply the constrained and partially constrained factor models to the macroeconomic indices dataset collected from OECD. The dataset contains 10 quarterly macroeconomic indices of 14 countries from 1990.Q2 to 2016.Q4 for107 quarters. Thus, we have

$T = 107$ and $p_1 \times p_2 = 14 \times 10$ matrix-valued time series. The countries include developed economies from North American, European, and Oceania. The indices cover four major groups, namely production, consumer price, money market, and international trade. Each original univariate time series is transformed by taking the first or second difference or logarithm to satisfy the mixing condition in Condition 4. Countries, detailed descriptions of the dataset, and transformation procedures are given in Tables 2.15 and 2.16 of Appendix 2.7.1.

We first fit an unconstrained matrix factor model which generates estimators of the row loading matrix and the column loading matrix. In the row loading matrix, each row represents a country by its factor loadings for all common row factors, whereas, in the column loading matrix, each row represents a macroeconomic index by its factor loadings for all common column factors. A hierarchical clustering algorithm is employed to cluster countries and macroeconomic indices based on their representations in the common row and column factor spaces, respectively. Figure 2.4 shows the hierarchical clustering results. Based on the clustering result, we construct the row and column constraint matrices. It seems that the row constraint matrix divides countries into 6 groups: (i) United States and Canada; (ii) New Zealand and Australia; (iii) Norway; (iv) Ireland, Denmark, and United Kingdom; (v) Finland and Sweden; (vi) France, Netherlands, Austria, and Germany. The grouping more or less follows geographical partitions with Norway different from all others due to its rich oil production and other distinct economic characteristics. The column constraint matrix divide macroeconomic indices into 5 categories: (i) GDP, production of total industry excluding construction, and production of total manufacturing ; (ii) long-term government bond yields and 3-month interbank rates and yields; (iii) total CPI and CPI of Food; (iv) CPI of Energy; (v) total exports value and total imports value in goods. Again, the grouping agrees with common economic knowledge.

Table 2.9 shows estimates of the row and column loading matrices for constrained and unconstrained 4×4 factor models. The loading matrices are normalized so that the norm of each column is one. They are also varimax-rotated to reveal a clear structure. The values shown are rounded values of the estimates multiplied by 10 for ease in display.



(a) Country Loading Clustering (b) Macroeconomic Index Loading Clustering

Figure 2.4: Macroeconomic series: Clustering loading matrices

From the table, both the row and column loading matrices exhibit similar patterns between unconstrained and constrained models, partially validating the constraints while simplifying the analysis.

Table 2.10 provides the estimates under the same setting as that of Table 2.9 but without any rotation. From the table, it is seen that except for the first common factors of the row loading matrices there exist some differences in the estimated loading matrices between unconstrained and constrained factor models. The results of constrained models convey more clearly the following observations. Consider the row factors. The first row common factor represents the status of global economy as it is a weighted average of all the countries under study. The remaining three row common factors mark certain differences between country groups. For the column factors, the first column common factor is dominated by the price index and interest rates; The second column common factor is mainly the production and international trade; The remaining two column common factors represent interaction between price indices, interest rates, productions, and international trade.

Table 2.11 compares the out-of-sample performance of unconstrained, constrained, and partially constrained factor models using a 10-fold cross validation (CV) for models

Model	Loading	Row	USA	CAN	NZL	AUS	NOR	IRL	DNK	GBR	FIN	SWE	FRA	NLD	AUT	DEU
$R_{unc,rot}$	$\widehat{R}'_{rot}$	1	7	7	1	1	-1	-2	-1	0	1	0	0	0	0	-1
		2	0	1	-2	-1	1	1	1	2	4	3	4	4	4	4
		3	2	-1	5	5	1	5	3	2	-1	1	1	0	0	0
		4	-1	1	1	2	9	-3	0	0	0	1	-1	1	0	0
$R_{con,rot}$	$\widehat{R}'_{rot}H'_R$	1	6	6	0	0	0	2	2	2	-1	-1	0	0	0	0
		2	-1	-1	0	0	0	3	3	3	4	4	3	3	3	3
		3	0	0	7	7	0	1	1	1	1	1	-1	-1	-1	-1
		4	0	0	0	0	10	0	0	0	1	1	0	0	0	0

Model	Loading	Row	CPI:Food	CPI:Tot	CPI:Ener	IR:Long	IR:3-Mon	P:TIEC	P:TM	GDP	IT:Ex	IT:Im
$C_{unc,rot}$	$\widehat{C}'_{rot}$	1	6	7	3	-1	1	0	0	-1	-1	0
		2	-2	1	4	1	-1	0	0	0	6	6
		3	0	0	1	8	6	-1	0	1	0	0
		4	1	-1	0	0	0	6	6	5	0	0
$C_{con,rot}$	$\widehat{C}'_{rot}H'_C$	1	7	7	0	0	0	0	0	0	0	0
		2	0	0	6	0	0	0	0	0	6	6
		3	0	0	0	7	7	0	0	0	0	0
		4	0	0	-2	0	0	6	6	6	1	1

Table 2.9: Estimations of row and column loading matrices (varimax rotated) of constrained and unconstrained matrix factor models for multinational macroeconomic indices. The loadings matrix are multiplied by 10 and rounded to integers for ease in display.

Model	Loading	Row	USA	CAN	NZL	AUS	NOR	IRL	DNK	GBR	FIN	SWE	FRA	NLD	AUT	DEU
R_{unc}	$\widehat{R}'$	1	3	2	2	2	2	2	3	3	3	3	3	3	3	3
		2	4	2	5	5	1	0	1	0	-3	-1	-2	-2	-2	-3
		3	3	6	-2	-2	4	-5	-3	-1	1	0	-1	1	0	0
		4	-4	-3	0	2	8	-1	1	0	-1	1	0	1	0	0
R_{con}	$\widehat{R}'H'_R$	1	1	1	2	2	2	3	3	3	4	4	3	3	3	3
		2	5	5	3	3	4	0	0	0	-2	-2	-2	-2	-2	-2
		3	-1	-1	5	5	-6	0	0	0	0	0	-1	-1	-1	-1
		4	-4	-4	3	3	6	-2	-2	-2	1	1	-1	-1	-1	-1

Model	Loading	Row	CPI:Food	CPI:Ener	CPI:Tot	IR:Long	IR:3-Mon	P:TIEC	P:TM	GDP	IT:Ex	IT:Im
C_{unc}	$\widehat{C}'$	1	1	4	2	4	3	3	3	3	4	4
		2	5	3	6	-1	1	-3	-4	-4	0	0
		3	5	-1	2	-1	1	4	4	3	-4	-4
		4	0	-1	-2	7	5	-2	-2	0	-3	-3
C_{con}	$\widehat{C}'H'_C$	1	6	-2	6	4	4	0	0	0	-2	-2
		2	0	0	0	3	3	5	5	5	3	3
		3	-3	3	-3	5	5	-3	-3	-3	1	1
		4	3	5	3	-1	-1	-2	-2	-2	5	5

Table 2.10: Estimations of row and column loading matrices of constrained and unconstrained matrix factor models for multinational macroeconomic indices. No rotation is used. The loadings matrix are multiplied by 10 and rounded to integers for ease in display.

with different number of factors. Residual sum of squares (RSS), their ratios to the total sum of squares (RSS/TSS), and the number of parameters are means of the 10-fold CV. Clearly, the constrained factor model uses far fewer parameters in the loading matrices yet achieves slightly better results than the unconstrained model. Using the same number of parameters, the partially constrained model is able to reduce markedly the RSS over the unconstrained model.

In this particular application, the constrained matrix factor model with the specified constraint matrices seems appropriate and plausible. If incorrect structures (constraint matrices) are imposed on the model, then the constrained model may become inappropriate. As we can see from the next example, a single orthogonal constraint actually hurts the performance. In cases like this, we need a second or a third constraint to achieve satisfactory performance. Nevertheless, the results from the constrained model are better than those from the unconstrained model.

Model	# Factor 1	# Factor 2	RSS	RSS/TSS	# Parameters
Full	(6,5)		570.50	0.449	134
Constrained	(6,5)		560.31	0.442	61
Partial	(6,5)	(6,5)	454.41	0.358	134
Full	(5,5)		613.26	0.482	120
Constrained	(5,5)		604.63	0.477	55
Partial	(5,5)	(5,5)	516.27	0.407	120
Full	(4,5)		658.15	0.517	106
Constrained	(4,5)		649.85	0.512	49
Partial	(4,5)	(4,5)	576.94	0.454	106
Full	(4,4)		729.46	0.573	96
Constrained	(4,4)		721.96	0.568	44
Partial	(4,4)	(4,4)	657.13	0.517	96
Full	(3,4)		787.80	0.620	82
Constrained	(3,4)		768.64	0.605	38
Partial	(3,4)	(3,4)	719.46	0.567	82
Full	(3,3)		868.43	0.684	72
Constrained	(3,3)		852.76	0.671	33
Partial	(3,3)	(3,3)	813.16	0.640	72

Table 2.11: Results of 10-fold CV of out-of-sample performance for the multinational macroeconomic indices. The numbers shown are average over the cross validation, where RSS and TSS stand for residual and total sum of squares, respectively.

2.5.2 Example 2: Company Financials

In this application, we investigate the constrained matrix-variate factor models for the time series of 16 quarterly financial measurements of 200 companies from 2006.Q1 to 2015.Q4 for 40 observations. Appendix 2.7.3 contains the descriptions of variables used along with their definitions, the 200 companies and their corresponding industry group

and sector information. Data are arranged in matrix-variate time series format. At each t, we observe a 16×200 matrix, whose rows represent financial variables and columns represent companies. Thus we have $T = 40$, $p_1 = 16$ and $p_2 = 200$. The total number of time series is $3,200$. Following the convention in eigenanalysis, we standardize the individual series before applying factor analysis. This data set was used in Wang et al. (2017) for an unconstrained matrix factor model.

The column constraint matrix $\boldsymbol{H}_C$ is constructed based on the industrial classification of Bloomberg. The 200 companies are classified into 51 industrial groups, such as biotechnology, oil & gas, computer, among others. Thus the dimension of $\boldsymbol{H}_C$ is 200×51. Since we do not have adequate prior knowledge on corporate financial, we do no impose any constraint on the row loading matrix. Thus, in this application, we use $\boldsymbol{H}_R = \boldsymbol{I}_{16}$.

We apply the unconstrained model (2.1), the orthogonal constrained model (2.7), and the partial constrained model (2.5) to the data set. Table 2.12 shows the average residual sum of squares (RSS) and their ratios to the total sum of squares (TSS) from a 10-fold CV for models with different number of factors. Again, it is clear, from the table, that the constrained matrix factor models use fewer number of parameters in loading matrices and achieve similar results. If we use the same number of parameters in the loading matrices, variances explained by the constrained matrix factor models are much larger than those of the unconstrained ones, indicating the impact of over-parameterization. This application with $3,200$ time series is typical in high-dimensional time series. The number of parameters involved is usually huge in a unconstrained model. Via the example, we showed that constrained matrix factor models can largely reduce the number of parameters while keeping the same explanation power.

2.5.3 Example 3: Fama-French 10 by 10 Series

Finally, we investigate constrained matrix-variate factor models for the monthly market-adjusted return series of Fama-French 10×10 portfolios from January 1964 to December 2015 for total 624 months and overall $62,400$ observations. The portfolios are the intersections of 10 portfolios formed by size (market equity, ME) and 10 portfolios

Model	# Factor 1	# Factor 2	RSS	RSS/SST	# parameters
Full	(4,10)		8140.32	0.869	2064
	(4,12)		7990.04	0.853	2464
	(4,19)		7587.11	0.810	3864
Constrained	(4,10)		8062.63	0.861	574
Partial	(4,10)	(4,2)	7969.83	0.851	936
	(4,10)	(4,9)	7623.25	0.814	1979
Full	(4, 20)		7539.68	0.805	4064
	(4, 27)		7261.49	0.775	5464
	(4, 39)		6872.18	0.734	7864
Constrained	(4, 20)		7646.70	0.816	1084
Partial	(4, 20)	(4,7)	7292.06	0.779	2191
	(4, 20)	(4,19)	6815.96	0.728	3979
Full	(5,10)		8012.10	0.855	2080
	(5,12)		7849.34	0.838	2480
	(5,19)		7420.04	0.792	3880
Constrained	(5,10)		7942.95	0.848	590
Partial	(5,10)	(5,2)	7849.40	0.838	968
	(5,10)	(5,9)	7472.10	0.798	2011
Full	(5,20)		7368.63	0.787	7960
	(5,23)		7250.73	0.774	4680
	(5,39)		6641.13	0.709	7880
Constrained	(5,20)		7489.20	0.800	1100
Partial	(5,20)	(5,3)	7357.80	0.786	1627
	(5,20)	(5,19)	6595.03	0.704	4011
Full	(5,30)		6960.70	0.743	6080
	(5,34)		6813.93	0.727	6880
	(5,59)		5988.15	0.639	11880
Constrained	(5,30)		7184.53	0.767	1610
Partial	(5,30)	(5,4)	6997.21	0.747	2286
	(5,30)	(5,29)	5936.64	0.634	6011

Table 2.12: Summary of 10-fold CV of out-of-sample analysis for the corporate financial of 16 series for each of 200 companies. The numbers shown are average over the cross validation and RSS and TSS denote, respectively, the residual and total sum of squares.

formed by the ratio of book equity to market equity (BE/ME). Thus, we have $T = 624$ and $p_1 \times p_2 = 10 \times 10$ matrix time series. The series are constructed by subtracting the monthly excess market returns from each of the original portfolio returns obtained from French (2017), so they are free of the market impact.

Using an unconstrained matrix factor model, Wang et al. (2017) carried out a clustering analysis on the ME and BE/ME loading matrices after rotation. Their results

suggest $\boldsymbol{H}_R = [\boldsymbol{h}_{R_1}, \boldsymbol{h}_{R_2}, \boldsymbol{h}_{R_3}]$, where $\boldsymbol{h}_{R_1} = [\mathbf{1}(5)/\sqrt{5}, \mathbf{0}(5)]$, $\boldsymbol{h}_{R_2} = [\mathbf{0}(5), \mathbf{1}(4)/2, 0]$, and $\boldsymbol{h}_{R_3} = [\mathbf{0}(9), 1]$. Therefore, ME factors are classified into three groups of smallest 5 ME's, middle 4 ME's, and the largest ME, respectively. For cases when we need 4 row constraints, we redefine $\boldsymbol{h}_{R_2} = [\mathbf{0}(5), \mathbf{1}(3)/\sqrt{3}, \mathbf{0}(2)]$ and add a fourth column $\boldsymbol{h}_{R_4} = [\mathbf{0}(8), 1, 0]$. For column constraints, $\boldsymbol{H}_C = [\boldsymbol{h}_{C_1}, \boldsymbol{h}_{C_2}, \boldsymbol{h}_{C_3}]$, where $\boldsymbol{h}_{C_1} = [1, \mathbf{0}(9)]$, $\boldsymbol{h}_{C_2} = [0, \mathbf{1}(3)/\sqrt{3}, \mathbf{0}(6)]$, $\boldsymbol{h}_{C_3} = [\mathbf{0}(4), \mathbf{1}(6)]$. Therefore, BE/ME factors are divided into three groups of the smallest BE/ME's, middle 3 BE/ME's, and the 6 largest BE/ME, respectively. For cases when we need 4 column constraints, we redefine $\boldsymbol{h}_{C_3} = [\mathbf{0}(4), \mathbf{1}(4)/2, \mathbf{0}(2)]$ and add a fourth column $\boldsymbol{h}_{C_4} = [\mathbf{0}(8), \mathbf{1}(2)]$.

Table 2.13 shows the estimates of the loading matrices for the constrained and unconstrained 2×2 factor models. The loading matrices are VARIMAX roated for ease in interpretation and normalized so that the norm of each column is one. From the table, the loading matrices exhibit similar patterns, but those of the constrained model convey the following observations more clearly. Consider the row factors, the first factor represents the difference between the average of the 5 smallest ME group and the weighted average of the remaining portfolio whereas the second factor is mainly the average of the medium 4 ME portfolios. For the column loading matrix, the first factor is a weighted average of the smallest BE/ME portfolio and the middle three portfolios. The second factor marks the difference between the smallest BE/ME portfolio from a weighted average of the two remaining groups. Finally, it is interesting to see that the constrained model uses only 16 parameters, yet it can reveal information similar to the unconstrained model that employs 40 parameters. This latter result demonstrates the power of using constrained factor models.

Table 2.14 compares the out-of-sample performance of unconstrained and constrained matrix factor models using a 10-fold CV for models with different number of factors constructed similarly to Table 2.11. In this case, the prediction RSS of the constrained model is slightly larger than that of the unconstrained one with the same number of factors, which may results from the misspecification of the constrained matrices. Testing the adequacy of the constrained matrix is an important research topic that will be addressed in future research. On the other hand, the constrained model uses a much

Model	Loading	Column	Rotated Estimated Loadings									
R_u	$\widehat{R}'$	1	0.43	0.46	0.44	0.43	0.33	0.16	0.05	-0.02	-0.20	-0.23
		2	-0.01	-0.01	-0.05	0.09	0.18	0.39	0.39	0.62	0.51	0.16
	$\widehat{R}'H'_R$	1	0.44	0.44	0.44	0.44	0.44	-0.04	-0.04	-0.04	-0.04	-0.15
		2	0.04	0.04	0.04	0.04	0.04	0.50	0.50	0.50	0.50	0.06
C_u	$\widehat{C}'$	1	0.70	0.48	0.37	0.30	0.14	0.07	0.05	-0.05	-0.09	0.15
		2	0.29	-0.07	-0.10	-0.23	-0.30	-0.32	-0.34	-0.44	-0.48	-0.34
	$\widehat{C}'H'_C$	1	0.78	0.36	0.36	0.36	0	0	0	0	0	0
		2	0.24	-0.18	-0.18	-0.18	-0.37	-0.37	-0.37	-0.37	-0.37	-0.37

Table 2.13: Estimates of the loading matrices of constrained and unconstrained matrix factor modes for Fama-French 10×10 portfolio returns. The loading matrices are varimax rotated and normalized for ease in comparison.

smaller number of parameters than the unconstrained model.

Model	# Factor 1	# Factor 2	RSS	RSS/SST	# Parameters
Full	(3,3)		3064.40	0.500	60
	(3,4)		2905.79	0.474	70
	(3,6)		2644.59	0.431	90
Constrained	(3,3)		3115.16	0.508	24
Partial	(3,3)	(3,3)	2819.06	0.460	60
	(3,3)	(1,1)	3079.79	0.502	36
Full	(3,2)		3316.55	0.541	50
	(3,4)		2905.79	0.474	70
Constrained	(3,2)		3361.03	0.548	18
Partial	(3,2)	(3,2)	3169.79	0.517	50
	(3,2)	(1,1)	3323.25	0.542	31
Full	(2,3)		3269.50	0.533	50
	(2,4)		3152.63	0.514	60
	(2,6)		2976.18	0.431	90
Constrained	(2,3)		3372.79	0.550	18
Partial	(2,3)	(2,3)	3154.36	0.514	50
	(2,3)	(1,2)	3296.73	0.538	37
Full	(2,2)		3473.32	0.567	40
	(2,3)		3269.50	0.533	50
	(2,4)		3152.63	0.514	60
Constrained	(2,2)		3535.56	0.577	16
Partial	(2,2)	(2,2)	3415.25	0.557	40
	(2,2)	(2,1)	3486.15	0.569	33

Table 2.14: Performance of out-of-sample 10-fold CV of constrained and unconstrained factor models using Fama-French 10×10 portfolio return series, where RSS and RSS/TSS denote, respectively, the residual and total sum of squares.

2.6 Proofs

We use the following notations. For $h \geq 0$, let $\mathbf{\Sigma}_{f,u}(h) = Cov(vec(\boldsymbol{F}_t), vec(\boldsymbol{U}_{t+h}))$,

$$\widetilde{\mathbf{\Sigma}}_{f,u}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} vec(\boldsymbol{F}_t)vec(\boldsymbol{U}_{t+h})', \quad \text{and} \quad \widetilde{\mathbf{\Sigma}}_{y}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} vec(\boldsymbol{Y}_t)vec(\boldsymbol{Y}_{t+h})'.$$

The auto-covariance matrices of $\mathbf{\Sigma}_{u,f}(h)$, $\mathbf{\Sigma}_{f}(h)$, $\mathbf{\Sigma}_{u}(h)$ and their sample versions are defined in a similar manner. The following regularity and factor strength conditions are needed.

Condition 1.

No linear combination of the components of $\boldsymbol{F}_t$ is white noise.

Condition 2.

There exists at least one h in $\{1, \ldots, h_0\}$, where $h_0 \geq 1$ is a positive integer, such that $\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \mathbf{\Omega}_{zq,ij}(h)\mathbf{\Omega}_{zq,ij}(h)'$ in equation (2.9) is of full rank.

Condition 1 is natural, as all the white noise linear combinations of $\boldsymbol{F}_t$ should be absorbed into $\boldsymbol{U}_t$, which ensures that there exists at least one $h \geq 1$ for which $\mathbf{\Omega}_{zq,ij}(h)$ is full-ranked. Condition 2 further ensures that $\boldsymbol{M}$ has k_1 positive eigenvalues.

Condition 3.

For $h \geq 0$, the maximum eigenvalue of $\mathbf{\Sigma}_{f,u}(h)$ and $\mathbf{\Sigma}_u$ remains bounded as T, p_1 and p_2 increase to infinity.

In model (2.2), $\boldsymbol{H}_R\boldsymbol{R}\boldsymbol{F}_t\boldsymbol{C}'\boldsymbol{H}_C'$ can be viewed as the signal part of the observation $\boldsymbol{Y}_t$, and $\boldsymbol{U}_t$ as the noise. Condition 3 requires two things. First, each element of $\mathbf{\Sigma}_u$ remains bounded as p_1 and p_2 increase to infinity. Thus each noise component does not goes to infinity so that the signals are not obscured by the noises. Second, as dimensions increase, the covariance matrix of noises does not have information concentrated in a few directions. Thus the noise part does not contain any useful information. This is reasonable since all the common components should be absorbed in the signal.

Condition 4.

The vector-valued process $vec(\boldsymbol{F}_t)$ is α-mixing. For some $\gamma > 2$, the mixing coefficients satisfy the condition that

$$\sum_{h=1}^{\infty} \alpha(h)^{1-2/\gamma} < \infty,$$

where $\alpha(h) = \sup_{\tau} \sup_{A \in \mathcal{F}_{-\infty}^{\tau}, B \in \mathcal{F}_{\tau+h}^{\infty}} |P(A \cap B) - P(A)P(B)|$ and $\mathcal{F}_{\tau}^{s}$ is the σ-field generated by $\{vec(\boldsymbol{F}_t) : \tau \leq t \leq s\}$.

Condition 5.

Let $F_{t,ij}$ be the ij-th entry of $\boldsymbol{F}_t$. Then, $E(|F_{t,ij}|^{2\gamma}) \leq C$ for any $i = 1, \ldots, k_1$, $j = 1, \ldots, k_2$ and $t = 1, \ldots, T$, where C is a positive constant and γ is given in Condition 4. In addition, there exists an integer h satisfying $1 \leq h \leq h_0$ such that $\boldsymbol{\Sigma}_f(h)$ is of rank $k = \max(k_1, k_2)$ and $\|\boldsymbol{\Sigma}_f(h)\|_2 \asymp O(1) \asymp \sigma_k(\boldsymbol{\Sigma}_f(h))$. For $i = 1, \ldots, k_1$ and $j = 1, \ldots, k_2$, $\frac{1}{T-h}\sum_{t=1}^{T-h} Cov(F_{t,i\cdot}, F_{t+h,i\cdot}) \neq \mathbf{0}$ and $\frac{1}{T-h}\sum_{t=1}^{T-h} Cov(F_{t,\cdot j}, F_{t+h,\cdot j}) \neq \mathbf{0}$.

Condition 4 and Condition 5 specify that the latent process $\{\boldsymbol{F}_t\}_{t=1,\ldots,T}$ only needs to satisfy the mixing condition specified in Condition 4 instead of the stationary condition. And we make use of the auto-covariance structure of the latent process $\{\boldsymbol{F}_t\}_{t=1,\ldots,T}$ without assuming any specific model. These two features make our estimation procedure more attractive and general than the standard principal component analysis.

We focus on the case of orthogonal constraints. Results for the non-orthogonal case and the partially-constrained case are similar.

The constrained factor model is $\boldsymbol{Y}_t = \boldsymbol{H}_R \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}' \boldsymbol{H}_C' + \boldsymbol{U}_t$. Suppose we have orthogonal constraints, that is $\boldsymbol{H}_R' \boldsymbol{H}_R = \boldsymbol{I}_{m_1}$ and $\boldsymbol{H}_C' \boldsymbol{H}_C = \boldsymbol{I}_{m_2}$, then the transformed $m_1 \times m_2$ data $\boldsymbol{X}_t = \boldsymbol{H}_R' \boldsymbol{X}_t \boldsymbol{H}_C = \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}' + \boldsymbol{E}_t$, where $\boldsymbol{E}_t = \boldsymbol{H}_R' \boldsymbol{U}_t \boldsymbol{H}_C$ and $\boldsymbol{E}_t$ is still white noise process.

Lemma 1. *Under Condition 3, each element of* $\boldsymbol{\Sigma}_e = Cov(vec(\boldsymbol{E}))$ *is uniformly bounded as* p_1 *and* p_2 *increase to infinity.*

Proof.

$$\begin{aligned}\boldsymbol{\Sigma}_e &= Cov(vec(\boldsymbol{H}_R'\boldsymbol{U}_t\boldsymbol{H}_C)) \\ &= Cov((\boldsymbol{H}_R' \otimes \boldsymbol{H}_C') \cdot vec(\boldsymbol{U}_t)) \\ &= (\boldsymbol{H}_R \otimes \boldsymbol{H}_C)' \cdot \boldsymbol{\Sigma}_u \cdot (\boldsymbol{H}_R \otimes \boldsymbol{H}_C).\end{aligned}$$

Let $\boldsymbol{A} = \boldsymbol{H}_R \otimes \boldsymbol{H}_C$. Since $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ are $p_1 \times m_1$ and $p_2 \times m_2$ orthogonal matrices respectively, $\boldsymbol{A}$ is a $p_1p_2 \times m_1m_2$ orthogonal matrix.

Let e_i be the i-th element of $vec(\boldsymbol{E}_t)$, $A_{\cdot i}$ be the i-th column vector of $\boldsymbol{A}$ for $i = 1, \ldots, m_1m_2$, then the diagonal elements of $\boldsymbol{\Sigma}_e$ are

$$Var(e_i) = A_{\cdot i}'\boldsymbol{\Sigma}_u A_{\cdot i} \leq \ \lambda_{\max}(\boldsymbol{\Sigma}_u) \text{ for } i = 1, \ldots, m_1m_2.$$

Condition 3 assumes $\lambda_{\max}(\boldsymbol{\Sigma}_u) \sim O(1)$, hence $Var(e) \sim O(1)$ for $i = 1, \ldots, m_1m_2$.

And off-diagonal elements of $\boldsymbol{\Sigma}_e$ are

$$Cov(e_i, e_j) \leq Var(e_i)^{\frac{1}{2}} Var(e_j)^{\frac{1}{2}} \sim O(1) \text{ for } i \neq j, \ i, j = 1, \ldots, m_1m_2.$$

Thus, each element of $\boldsymbol{\Sigma}_e$ remains bounded if the maximum eigenvalue of $\boldsymbol{\Sigma}_u = Cov(vec(\boldsymbol{U}))$ is bounded as p_1 and p_2 increase to infinity. □

Lemma 2. *Under the assumption that $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ are orthogonal. Condition 6 also ensures that $\|\boldsymbol{R}\|_2^2 \asymp p_1^{1-\delta_1} \asymp \|\boldsymbol{R}\|_{min}^2$ and $\|\boldsymbol{C}\|_2^2 \asymp p_2^{1-\delta_1} \asymp \|\boldsymbol{C}\|_{min}^2$.*

Proof. For any orthogonal matrix $\boldsymbol{H}$, we have $\|\boldsymbol{HR}\|_2^2 = \|\boldsymbol{R}\|_2^2$ and $\|\boldsymbol{HR}\|_{min}^2 = \|\boldsymbol{R}\|_{min}^2$. And the results follow. □

In the following proofs, we work with the transformed model (2.7), as in $\boldsymbol{X}_t = \boldsymbol{RF}_t\boldsymbol{C}' + \boldsymbol{E}_t$ where $\boldsymbol{X}_t$ and $\boldsymbol{E}_t$ are $m_1 \times m_2$ matrices, $\boldsymbol{F}_t$ is $k_1 \times k_2$ matrix, $\boldsymbol{R}$ is the $m_1 \times k_1$ row loading matrix, and $\boldsymbol{C}$ is the $m_2 \times k_2$ column loading matrix for the transformed model.

We start by defining some quantities used in the proofs. Write

$$\begin{aligned}
\boldsymbol{\Omega}_{s,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{R}\boldsymbol{F}_t C_{i\cdot}, \boldsymbol{R}\boldsymbol{F}_{t+h} C_{j\cdot}),\\
\boldsymbol{\Omega}_{fc,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{F}_t C_{i\cdot}, \boldsymbol{F}_{t+h} C_{j\cdot}),\\
\widehat{\boldsymbol{\Omega}}_{s,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{R}\boldsymbol{F}_t C_{i\cdot} C_{j\cdot}' \boldsymbol{F}_{t+h}' \boldsymbol{R}',\\
\widehat{\boldsymbol{\Omega}}_{se,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{R}\boldsymbol{F}_t C_{i\cdot} E_{t+h,\cdot j}',\\
\widehat{\boldsymbol{\Omega}}_{es,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} E_{t,\cdot j} C_{i\cdot}' \boldsymbol{F}_{t+h}' \boldsymbol{R}',\\
\widehat{\boldsymbol{\Omega}}_{e,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} E_{t,\cdot j} E_{t+h,\cdot j}',\\
\widehat{\boldsymbol{\Omega}}_{fc,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{F}_t C_{i\cdot} C_{j\cdot}' \boldsymbol{F}_{t+h}'.
\end{aligned}$$

The following Lemma 3 from Wang et al. (2017) establishes the entry-wise convergence rate of the covariance matrix estimation of the vectorized latent factor process $vec(\boldsymbol{F}_t)$.

Lemma 3. *Let $F_{t,ij}$ denote the ij-th entry of $\boldsymbol{F}_t$. Under Condition 4 and Condition 5, for any $i,k = 1,\ldots,k_1$ and $j,l = 1,\cdots,k_2$, we have*

$$\left|\frac{1}{T-h}\sum_{t=1}^{T-h}\left(F_{t,ij}F_{t+h,kl} - Cov(F_{t,ij}F_{t+h,kl})\right)\right| = O_p(T^{-1/2}). \tag{2.16}$$

Under the matrix-variate factor Model (2.7), the $\boldsymbol{R}\boldsymbol{F}_t\boldsymbol{C}'$ is the signal and $\boldsymbol{E}_t$ is the noise.

Lemma 4. *Under Conditions 1-6, it holds that*

$$\begin{aligned}
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{s,ij}(h) - \boldsymbol{\Omega}_{s,ij}(h)\|_2^2 &= O_p(p_1^{2-2\delta_1}p_2^{2-2\delta_2}T^{-1}), &(2.17)\\
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{se,ij}(h) - \boldsymbol{\Omega}_{se,ij}(h)\|_2^2 &= O_p(m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}T^{-1}), &(2.18)\\
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{es,ij}(h) - \boldsymbol{\Omega}_{es,ij}(h)\|_2^2 &= O_p(m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}T^{-1}), &(2.19)\\
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{e,ij}(h) - \boldsymbol{\Omega}_{e,ij}(h)\|_2^2 &= O_p(m_1^2 m_2^2 T^{-1}). &(2.20)
\end{aligned}$$

Proof. To prove the convergence rate of $\widehat{\boldsymbol{\Omega}}_{s,ij}(h)$ in (4.32), we first establish the convergence rate of estimating $\boldsymbol{\Omega}_{fc,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{F}_t C_{i\cdot}, \boldsymbol{F}_{t+h} C_{j\cdot})$.

$$
\begin{aligned}
&\|\widehat{\boldsymbol{\Omega}}_{fc,ij}(h) - \boldsymbol{\Omega}_{fc,ij}(h)\|_2^2 \le \|\widehat{\boldsymbol{\Omega}}_{fc,ij}(h) - \boldsymbol{\Omega}_{fc,ij}(h)\|_F^2 \\
&= \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} (\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t - E(\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t)) \cdot vec(C_{i\cdot}C_{j\cdot}') \right\|_2^2 \\
&\le \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} (\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t - E(\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t)) \right\|_F^2 \cdot \|C_{i\cdot}\|_2^2 \cdot \|C_{j\cdot}\|_2^2. \qquad (2.21)
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\widehat{\boldsymbol{\Omega}}_{s,ij}(h) - \boldsymbol{\Omega}_{s,ij}(h)\|_2^2 \\
&= \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\boldsymbol{R}\cdot(\widehat{\boldsymbol{\Omega}}_{fc,ij}(h) - \boldsymbol{\Omega}_{fc,ij}(h))\cdot \boldsymbol{R}'\|_2^2 \\
&\le \|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} (\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t - E(\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t)) \right\|_F^2 \cdot \left(\sum_{i=1}^{m_2} \|C_{i\cdot}\|_2^2 \right)^2 \\
&= \|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} (\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t - E(\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t)) \right\|_F^2 \cdot \|\boldsymbol{C}\|_F^4 \\
&\le k_2^2\, \|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} (\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t - E(\boldsymbol{F}_{t+h}\otimes \boldsymbol{F}_t)) \right\|_F^2 \cdot \|\boldsymbol{C}\|_2^4 \\
&= O_p(p_1^{2-2\delta_1} p_2^{2-2\delta_2} T^{-1}).
\end{aligned}
$$

The first inequality comes from (2.21) and the last inequality follows from Condition 6 and Lemma 1.

To prove the convergence rate of covariance between signal at t and noise at $t+h$ in (4.33), we first establish the convergence rate of covariance between $\boldsymbol{F}_t C_{i\cdot}$ and $E_{t+h,\cdot j}$.

$$
\begin{aligned}
&\left\| \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j} - \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j}\right)\right\|_2^2 \\
&\leq \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} vec(\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j}) - \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(vec(\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j})\right)\right\|_2^2 \\
&\leq \| \frac{1}{T-h}\sum_{t=1}^{T-h} \left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t - E\left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t\right)\right)\cdot vec(C_{i\cdot})\|_2^2 \\
&\leq \| \frac{1}{T-h}\sum_{t=1}^{T-h} \left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t - E\left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t\right)\right)\|_2^2 \cdot \|C_{i\cdot}\|_2^2.
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\widehat{\boldsymbol{\Omega}}_{se,ij}(h) - \boldsymbol{\Omega}_{se,ij}(h)\|_2^2 \\
&= \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{R}\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j} - \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{R}\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j}\right)\right\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\boldsymbol{R}\|_2^2 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} \boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j} - \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_t C_{i\cdot} E'_{t+h,\cdot j}\right)\right\|_2^2 \\
&\leq \|\boldsymbol{R}\|_2^2 \cdot \sum_{j=1}^{m_2} \| \frac{1}{T-h}\sum_{t=1}^{T-h} \left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t - E\left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t\right)\right)\|_2^2 \cdot \sum_{i=1}^{m_2} \|C_{i\cdot}\|_2^2 \\
&\leq \|\boldsymbol{R}\|_2^2 \cdot \sum_{i=1}^{m_2} \| \frac{1}{T-h}\sum_{t=1}^{T-h} \left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t - E\left(E_{t+h,\cdot j}\otimes \boldsymbol{F}_t\right)\right)\|_2^2 \cdot k_2\|\boldsymbol{C}\|_2^2 \\
&= O_p(m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2} T^{-1}).
\end{aligned}
$$

To prove the convergence rate of covariance between noise at t and signal at $t+h$ in (4.34), we use similar arguments and get

$$
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\widehat{\boldsymbol{\Omega}}_{es,ij}(h) - \boldsymbol{\Omega}_{es,ij}(h)\|_2^2 = O_p(m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2} T^{-1/2}).
$$

.

And the convergence rate of $\widehat{\boldsymbol{\Omega}}_{e,ij}(h)$ in (4.35) is given by

$$
\begin{aligned}
&\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \|\widehat{\boldsymbol{\Omega}}_{e,ij}(h) - \boldsymbol{\Omega}_{e,ij}(h)\|_2^2 \\
&= \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E_{t,\cdot i} E'_{t+h,\cdot j}\right\| \\
&= O_p(m_1^2 m_2^2 T^{-1}).
\end{aligned}
$$

□

With the four rates established in Lemma 9, we can study the rate of convergence for the transformed observed covariance matrix $\widehat{\boldsymbol{\Omega}}_{x,ij}(h)$.

Lemma 5. *Under Conditions 1-6, it holds that*

$$\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\|_2^2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1^2m_2^2)\cdot T^{-1}\right). \tag{2.22}$$

Proof. By definition of $\widehat{\boldsymbol{\Omega}}_{x,ij}(h)$ in Section 2.2, we can decompose $\widehat{\boldsymbol{\Omega}}_{x,ij}(h)$ into the following four parts,

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}}_{x,ij}(h) &= \frac{1}{T-h}\sum_{t=1}^{T-h} X_{t,\cdot i}X'_{t+h,\cdot j} \\
&= \frac{1}{T-h}\sum_{t=1}^{T-h}(\boldsymbol{R}\boldsymbol{F}_tC_{i\cdot}+E_{t,i\cdot})(\boldsymbol{R}\boldsymbol{F}_tC_{i\cdot}+E_{t+h,j\cdot})' \\
&= \widehat{\boldsymbol{\Omega}}_{s,ij}(h)+\widehat{\boldsymbol{\Omega}}_{se,ij}(h)+\widehat{\boldsymbol{\Omega}}_{es,ij}(h)+\widehat{\boldsymbol{\Omega}}_{e,ij}(h).
\end{aligned}$$

Thus from Lemma 4, we have

$$\begin{aligned}
&\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\|_2^2 \\
&\le 4\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}(\|\widehat{\boldsymbol{\Omega}}_{s,ij}(h)-\boldsymbol{\Omega}_{s,ij}(h)\|_2^2+\|\widehat{\boldsymbol{\Omega}}_{se,ij}(h)-\boldsymbol{\Omega}_{se,ij}(h)\|_2^2 \\
&\qquad +\|\widehat{\boldsymbol{\Omega}}_{es,ij}(h)-\boldsymbol{\Omega}_{es,ij}(h)\|_2^2+\|\widehat{\boldsymbol{\Omega}}_{e,ij}(h)-\boldsymbol{\Omega}_{e,ij}(h))\|_2^2 \\
&= O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1^2m_2^2)\cdot T^{-1}\right).
\end{aligned}$$

□

Lemma 6. *Under Conditions 1-6 and $m_1p_1^{-1+\delta_1}m_2p_2^{-1+\delta_2}T^{-1/2}=o_p(1)$, it holds that*

$$\|\widehat{\boldsymbol{M}}-\boldsymbol{M}\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1p_1^{1-\delta_1}m_2p_2^{1-\delta_2})\cdot T^{-1/2}\right). \tag{2.23}$$

Proof. By definitions of $\boldsymbol{M}$ in (2.11) and its sample version $\widehat{\boldsymbol{M}}$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{M}}-\boldsymbol{M}\|_2 &= \left\|\sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}(\widehat{\boldsymbol{\Omega}}_{x,ij}(h)\widehat{\boldsymbol{\Omega}}'_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\boldsymbol{\Omega}'_{x,ij}(h))\right\|_2 \\
&\le \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\left(\|(\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h))(\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h))'\|_2+2\|\boldsymbol{\Omega}_{x,ij}(h)\|_2\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\|_2\right) \\
&= \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\|_2^2+2\sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{\Omega}_{x,ij}(h)\|_2\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h)-\boldsymbol{\Omega}_{x,ij}(h)\|_2.
\end{aligned}$$

Now we investigate each item in the above formula.

$$
\begin{aligned}
&\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{\Omega}_{x,ij}(h)\|_2^2 = \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\boldsymbol{\Omega}_{fc,ij}(h)\boldsymbol{R}'\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \|\boldsymbol{\Omega}_{fc,ij}(h)\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E(\boldsymbol{F}_t C_{i\cdot} C_{j\cdot}' \boldsymbol{F}_{t+h}') \right\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(vec(\boldsymbol{F}_t C_{i\cdot} C_{j\cdot}' \boldsymbol{F}_{t+h}')\right) \right\|_2^2 \\
&= \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_{t+h} \otimes \boldsymbol{F}_t\right) \cdot vec(C_{i\cdot} C_{j\cdot}') \right\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_{t+h} \otimes \boldsymbol{F}_t\right) \right\|_2^2 \cdot \|vec(C_{i\cdot} C_{j\cdot}')\|_2^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_{t+h} \otimes \boldsymbol{F}_t\right) \right\|_2^2 \cdot \|C_{i\cdot} C_{j\cdot}'\|_F^2 \\
&= \sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_{t+h} \otimes \boldsymbol{F}_t\right) \right\|_2^2 \cdot \|C_{i\cdot}\|_2^2 \|C_{j\cdot}'\|_2^2 \\
&= \|\boldsymbol{R}\|_2^4 \cdot \left\| \frac{1}{T-h}\sum_{t=1}^{T-h} E\left(\boldsymbol{F}_{t+h} \otimes \boldsymbol{F}_t\right) \right\|_2^2 \cdot \left(\sum_{i=1}^{m_2} \|C_{i\cdot}\|_2^2 \right)^2 \\
&= O_p(p_1^{2-2\delta_1} p_2^{2-2\delta_2}).
\end{aligned}
$$

From Lemma 5, we have

$$
\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h) - \boldsymbol{\Omega}_{x,ij}(h)\|_2^2 = O_p\left(\max(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1^2 m_2^2) \cdot T^{-1}\right),
$$

then

$$
\begin{aligned}
&\left(\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\|\boldsymbol{\Omega}_{x,ij}(h)\right\|_2 \left\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h) - \boldsymbol{\Omega}_{x,ij}(h)\right\|_2 \right)^2 \\
&\leq \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\|\boldsymbol{\Omega}_{x,ij}(h)\right\|_2^2 \cdot \sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\|\widehat{\boldsymbol{\Omega}}_{x,ij}(h) - \boldsymbol{\Omega}_{x,ij}(h)\right\|_2^2 \\
&= O_p(p_1^{2-2\delta_1} p_2^{2-2\delta_2}) \cdot O_p(\max(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1^2 m_2^2) \cdot T^{-1} \\
&= O_p(\max(p_1^{4-4\delta_1} p_2^{4-4\delta_2},\ m_1^2 p_1^{2-2\delta_1} m_2^2 p_2^{2-2\delta_2}) \cdot T^{-1}.
\end{aligned}
$$

Thus, from the above results, Lemma 5 and the condition that

$$m_1 p_1^{-1+\delta_1} m_2 p_2^{-1+\delta_2} T^{-1/2} = o_p(1),$$

we have

$$\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2 = O_p\left(\max(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1^2 m_2^2) \cdot T^{-1}\right) + O_p\left(\max(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}\right) \cdot T^{-1/2}$$
$$= O_p\left(\max(p_1^{2-2\delta_1} p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2}\right).$$

□

Similar to the proof of Lemma 5 in Wang et al. (2017), we have

Lemma 7. *Under Condition 3 and Condition 5, we have*

$$\lambda_i(\boldsymbol{M}) \asymp p_1^{2-2\delta_1} p_2^{2-2\delta_2}, \qquad i = 1, 2, \ldots, k_1,$$

where $\lambda_i(\boldsymbol{M})$ denotes the i-th largest singular value of $\boldsymbol{M}$.

Proof of Theorem 1

Proof. By Lemma 3-7, and Lemma 3 in Lam et al. (2011a), we have

$$\|\widehat{\boldsymbol{Q}}_1 - \boldsymbol{Q}_1\|_2 \le \frac{8}{\lambda_{min}(\boldsymbol{M})} \|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2 = O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}} \frac{m_2}{p_2^{1-\delta_2}} T^{-1/2}\right)\right).$$

Proof for $\|\widehat{\boldsymbol{Q}}_2 - \boldsymbol{Q}_2\|_2$ is similar. □

Proof of Theorem 2

Proof. The proof is similar to that of Theorem 1 of Lam and Yao (2012). Let λ_j and $\boldsymbol{q}_j$ be the j-th largest eigenvalue and eigenvector of $\boldsymbol{M}$, respectively. The corresponding sample versions are denoted by $\widehat{\lambda}_j$ and $\widehat{\boldsymbol{q}}_j$ for the matrix $\widehat{\boldsymbol{M}}$. Let $\boldsymbol{Q}_1 = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{k_1})$, $\boldsymbol{B}_1 = (\boldsymbol{q}_{k_1+1}, \ldots, \boldsymbol{q}_{m_1})$, $\widehat{\boldsymbol{Q}}_1 = (\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_{k_1})$ and $\widehat{\boldsymbol{B}}_1 = (\widehat{\boldsymbol{q}}_{k_1+1}, \ldots, \widehat{\boldsymbol{q}}_{m_1})$.

Eigenvalues λ_j, $j = 1, \ldots, k_1$

For $j = 1, \ldots, k_1$, we have

$$\widehat{\lambda}_j - \lambda_j = \widehat{\boldsymbol{q}}_j' \widehat{\boldsymbol{M}} \widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j' \boldsymbol{M} \boldsymbol{q}_j = I_1 + I_2 + I_3 + I_4 + I_5,$$

where

$$I_1 = (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'(\widehat{\boldsymbol{M}} - \boldsymbol{M})\widehat{\boldsymbol{q}}_j,\ I_2 = (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j), \tag{2.24}$$

$$I_3 = (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}\boldsymbol{q}_j,\ I_4 = \boldsymbol{q}_j'(\widehat{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{q}_j,\ I_5 = \boldsymbol{q}_j'(\widehat{\boldsymbol{M}} - \boldsymbol{M})(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j). \tag{2.25}$$

We have, from Theorem 1,

$$\|\widehat{\boldsymbol{q}}_j' - \boldsymbol{q}_j\|_2 \leq \|\widehat{\boldsymbol{Q}}_1 - \boldsymbol{Q}_1\|_2 = O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right),\ for\ j = 1, \ldots, k_1.$$

And by Lemma 6, $\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}\right) \cdot T^{-1/2}$.

Also from Lemma 7, we have $\|\boldsymbol{M}\|_2 = O_p(p_1^{2-2\delta_1}p_2^{2-2\delta_2})$.

Then,

$$\|I_1\|_2 = \|(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'(\widehat{\boldsymbol{M}} - \boldsymbol{M})\widehat{\boldsymbol{q}}_j\|_2 \leq \|\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j\|_2 \cdot \|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2 \cdot \|\widehat{\boldsymbol{q}}_j\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1^2 m_2^2) \cdot T^{-1}\right)$$

$$\|I_2\|_2 = \|(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)\|_2 \leq \|\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j\|_2^2 \cdot \|\boldsymbol{M}\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1^2 m_2^2) \cdot T^{-1}\right)$$

$$\|I_3\|_2 = \|\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}\boldsymbol{q}_j\|_2 \leq \|\widehat{\boldsymbol{q}}_j' - \boldsymbol{q}_j\|_2 \cdot \|\boldsymbol{M}\|_2 \cdot \|\boldsymbol{q}_j\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2}\right)$$

$$\|I_4\|_2 = \|\boldsymbol{q}_j'(\widehat{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{q}_j\|_2 \leq \|\boldsymbol{q}_j\|_2\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2\|\boldsymbol{q}_j\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2}\right)$$

$$\|I_5\|_2 = \|\boldsymbol{q}_j'(\widehat{\boldsymbol{M}} - \boldsymbol{M})(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)\|_2 \leq \|\boldsymbol{q}_j\|_2\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_2\|\widehat{\boldsymbol{q}}_j' - \boldsymbol{q}_j\|_2$$
$$= O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2}\right).$$

Thus, under the condition that $m_1 p_1^{-1+\delta_1} m_2 p_2^{-1+\delta_2} T^{-1/2} = o_p(1)$, we have

$$|\widehat{\lambda}_j - \lambda_j| = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2}\right),\ for\ j = 1, \ldots, k_1.$$

Eigenvalues λ_j, $j = k_1 + 1, \ldots, p_1$

Similar to proof of Theorem 1 with Lemma 3 in Lam et al. (2011a), we have

$$\|\widehat{\boldsymbol{B}}_1 - \boldsymbol{B}_1\|_2 = O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right).$$

And hence

$$\|\widehat{\boldsymbol{q}}_j' - \boldsymbol{q}_j\|_2 \leq \|\widehat{\boldsymbol{Q}}_1 - \boldsymbol{Q}_1\|_2 = O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right),\ for\ j = k_1 + 1, \ldots, p_1.$$

Define $\widetilde{\boldsymbol{M}} = \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\widehat{\boldsymbol{\Omega}}_{i,j}(h)\boldsymbol{\Omega}_{i,j}'(h)$, then

$$\|\widetilde{\boldsymbol{M}} - \boldsymbol{M}\| = \left\|\sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\left(\widehat{\boldsymbol{\Omega}}_{ij}(h)\boldsymbol{\Omega}_{ij}'(h) - \boldsymbol{\Omega}_{ij}(h)\boldsymbol{\Omega}_{ij}'(h)\right)\right\|_2$$
$$\leq \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2}\left\|\left(\widehat{\boldsymbol{\Omega}}_{ij}(h) - \boldsymbol{\Omega}_{ij}(h)\right)\right\|_2\left\|\boldsymbol{\Omega}_{ij}'(h)\right\|_2$$
$$= O_p(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1 p_1^{1-\delta_1} m_2 p_2^{1-\delta_2}) \cdot T^{-1/2},\ \text{from Lemma 6.}$$

For $j = k_1 + 1, \ldots, p_1$, since $\lambda_j = 0$ we have

$$\widehat{\lambda}_j = \widehat{\boldsymbol{q}}_j'\widehat{\boldsymbol{M}}\widehat{\boldsymbol{q}}_j = K_1 + K_2 + K_3,$$

where $K_1 = \widehat{\boldsymbol{q}}_j'(\widehat{\boldsymbol{M}} - \widetilde{\boldsymbol{M}} - \widetilde{\boldsymbol{M}}' + \boldsymbol{M})\widehat{\boldsymbol{q}}_j$, $K_2 = 2\widehat{\boldsymbol{q}}_j'(\widetilde{\boldsymbol{M}} - \boldsymbol{M})(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)$ and $K_3 = (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)$.

Then,

$$\|K_1\|_2 = \left\| \widehat{\boldsymbol{q}}_j' \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left(\widehat{\boldsymbol{\Omega}}_{ij}(h)\widehat{\boldsymbol{\Omega}}_{ij}'(h) - \widehat{\boldsymbol{\Omega}}_{ij}(h)\boldsymbol{\Omega}_{ij}'(h) - \boldsymbol{\Omega}_{ij}(h)\widehat{\boldsymbol{\Omega}}_{ij}'(h) + \boldsymbol{\Omega}_{ij}(h)\boldsymbol{\Omega}_{ij}'(h) \right) \widehat{\boldsymbol{q}}_j \right\|_2$$

$$= \left\| \widehat{\boldsymbol{q}}_j' \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left(\widehat{\boldsymbol{\Omega}}_{ij}(h) - \boldsymbol{\Omega}_{ij}(h) \right) \left(\widehat{\boldsymbol{\Omega}}_{ij}(h) - \boldsymbol{\Omega}_{ij}(h) \right)' \widehat{\boldsymbol{q}}_j \right\|_2$$

$$\leq \sum_{h=1}^{h_0}\sum_{i=1}^{m_2}\sum_{j=1}^{m_2} \left\| \left(\widehat{\boldsymbol{\Omega}}_{ij}(h) - \boldsymbol{\Omega}_{ij}(h) \right) \right\|_2^2 = O_p(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2},\ m_1^2m_2^2)\cdot T^{-1})$$

$$\|K_2\|_2 = \left\| 2\widehat{\boldsymbol{q}}_j' \cdot \left(\widetilde{\boldsymbol{M}} - \boldsymbol{M} \right) \cdot (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j) \right\|_2 \leq 2 \left\| \widetilde{\boldsymbol{M}} - \boldsymbol{M} \right\|_2 \cdot \|\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1^2m_2^2)\cdot T^{-1} \right)$$

$$\|K_3\|_2 = \left\| (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j)'\boldsymbol{M}(\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j) \right\|_2 \leq \left\| (\widehat{\boldsymbol{q}}_j - \boldsymbol{q}_j) \right\|_2^2 \|\boldsymbol{M}\|_2 = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1^2m_2^2)\cdot T^{-1} \right).$$

Thus, we have

$$|\widehat{\lambda}_j| = O_p\left(\max(p_1^{2-2\delta_1}p_2^{2-2\delta_2}, m_1^2m_2^2)\cdot T^{-1} \right), \qquad for \qquad j = 1, \ldots, k_1.$$

□

Proof of Theorem 3

Proof. $\boldsymbol{S}_t$ is the dynamic signal part of $\boldsymbol{X}_t$, i.e. $\boldsymbol{S}_t = \boldsymbol{H}_R\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2'\boldsymbol{H}_C'$. And its estimator is $\widehat{\boldsymbol{S}}_t = \boldsymbol{H}_R\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{X}_t\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2'\boldsymbol{H}_C'$. We have

$$\begin{aligned}
\widehat{\boldsymbol{S}}_t - \boldsymbol{S}_t = \ & \boldsymbol{H}_R \left(\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{X}_t\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2' - \boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2' \right) \boldsymbol{H}_C' = \boldsymbol{H}_R \left(\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1' \left(\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2' + \boldsymbol{E}_t \right) \widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2' - \boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2' \right) \boldsymbol{H}_C' \\
= \ & \boldsymbol{H}_R \left(\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2'(\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2' - \boldsymbol{Q}_2\boldsymbol{Q}_2') + (\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1' - \boldsymbol{Q}_1\boldsymbol{Q}_1')\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2' + \widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{E}_t\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2' \right) \boldsymbol{H}_C' \\
= \ & I_1 + I_2 + I_3.
\end{aligned}$$

Since $\boldsymbol{H}_R$ and $\boldsymbol{H}_C$ are orthogonal matrices, we have

$$\begin{aligned}
\|I_1\|_2^2 &= \left\| \boldsymbol{H}_R\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2'(\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2' - \boldsymbol{Q}_2\boldsymbol{Q}_2')\boldsymbol{H}_C' \right\|_2^2 \\
&\leq \|\boldsymbol{Z}_t\|_2^2 \left\| (\widehat{\boldsymbol{Q}}_2 - \boldsymbol{Q}_2)\widehat{\boldsymbol{Q}}_2' + \boldsymbol{Q}_2(\widehat{\boldsymbol{Q}}_2 - \boldsymbol{Q}_2)' \right\|_2^2 \\
&\leq 2\|\boldsymbol{Z}_t\|_2^2 \|\widehat{\boldsymbol{Q}}_2 - \boldsymbol{Q}_2\|_2^2
\end{aligned}$$

Thus by Theorem 1, we have

$$\begin{aligned}
\|I_1\| &= O_p\left(p_1^{1/2-\delta_1/2}p_2^{1/2-\delta_2/2} \right) \cdot O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2} \right) \right) \\
&= O_p\left(\max\left(p_1^{1/2-\delta_1/2}p_2^{1/2-\delta_2/2},\ m_1p_1^{-1/2+\delta_1/2}m_2p_2^{-1/2+\delta_2/2} \right) \cdot T^{-1/2} \right).
\end{aligned}$$

Similarity, we have

$$\begin{aligned}\|I_2\|_2 &= \left\|(\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1' - \boldsymbol{Q}_1\boldsymbol{Q}_1')\boldsymbol{Q}_1\boldsymbol{Z}_t\boldsymbol{Q}_2'\right\|_2 \leq 2\|\boldsymbol{Z}_t\|_2\|\widehat{\boldsymbol{Q}}_1 - \boldsymbol{Q}_1\|_2 \\ &= O_p\left(\max\left(p_1^{1/2-\delta_1/2}p_2^{1/2-\delta_2/2},\ m_1p_1^{-1/2+\delta_1/2}m_2p_2^{-1/2+\delta_2/2}\right)\cdot T^{-1/2}\right),\end{aligned}$$

and

$$\|I_3\|_2 = \left\|\widehat{\boldsymbol{Q}}_1\widehat{\boldsymbol{Q}}_1'\boldsymbol{E}_t\widehat{\boldsymbol{Q}}_2\widehat{\boldsymbol{Q}}_2'\right\|_2 \leq \left\|\widehat{\boldsymbol{Q}}_1'\boldsymbol{E}_t\widehat{\boldsymbol{Q}}_2\right\|_2 \leq \|(\widehat{\boldsymbol{Q}}_2'\otimes\widehat{\boldsymbol{Q}}_1')vec(\boldsymbol{E}_t)\|_2 \leq k_1k_2\|\boldsymbol{\Sigma}_e\|_2 = O_p(1).$$

Thus,

$$\|\widehat{\boldsymbol{S}}_t - \boldsymbol{S}_t\|_2 = O_p\left(\max\left(p_1^{1/2-\delta_1/2}p_2^{1/2-\delta_2/2},\ m_1p_1^{-1/2+\delta_1/2}m_2p_2^{-1/2+\delta_2/2}\right)\cdot T^{-1/2}+1\right)$$

□

Proof of Theorem 4

Proof.

$$\mathcal{D}(\widehat{\boldsymbol{Q}}_i,\boldsymbol{Q}_i) = \left(1-\frac{1}{k_i}Tr\left(\widehat{\boldsymbol{Q}}_i\widehat{\boldsymbol{Q}}_i'\boldsymbol{Q}_i\boldsymbol{Q}_i'\right)\right)^{-1/2}, \quad for \quad i=1,2.$$

From Liu and Chen (2016a),

$$\mathcal{D}(\widehat{\boldsymbol{Q}}_i,\boldsymbol{Q}_i) = O_p\left(\|\widehat{\boldsymbol{Q}}_i,\boldsymbol{Q}_i\|_2\right) = O_p\left(\max\left(T^{-1/2},\ \frac{m_1}{p_1^{1-\delta_1}}\frac{m_2}{p_2^{1-\delta_2}}T^{-1/2}\right)\right)$$

for $i=1,2$. Since $\mathcal{D}(\widehat{\boldsymbol{\Lambda}},\boldsymbol{\Lambda}) = \mathcal{D}(\widehat{\boldsymbol{Q}}_1,\boldsymbol{Q}_1)$ and $\mathcal{D}(\widehat{\boldsymbol{\Gamma}},\boldsymbol{\Gamma}) = \mathcal{D}(\widehat{\boldsymbol{Q}}_2,\boldsymbol{Q}_2)$, the result follows.

□

2.7 Appendix

2.7.1 Multinational Macroeconomic Indices Dataset

Table 2.15 lists the short name of each series, its mnemonic (the series label used in the OECD database), the transformation applied to the series, and a brief data description. All series are from the OECD Database. In the transformation column, Δ denote the first difference, $\Delta\ln$ denote the first difference of the logarithm. GP denotes the measure of growth rate last period.

2.7.2 Tables of Simulation Results

Short name	Mnemonic	Tran	description
CPI: Food	CPGDFD	$\Delta^2 \ln$	Consumer Price Index: Food, seasonally adjusted
CPI: Ener	CPGREN	$\Delta^2 \ln$	Consumer Price Index: Energy, seasonally adjusted
CPI: Tot	CPALTT01	$\Delta^2 \ln$	Consumer Price Index: Total, seasonally adjusted
IR: Long	IRLT	Δ	Interest Rates: Long-term gov bond yields
IR: 3-Mon	IR3TIB	Δ	Interest Rates: 3-month Interbank rates and yields
P: TIEC	PRINTO01	$\Delta \ln$	Production: Total industry excl construction
P: TM	PRMNTO01	$\Delta \ln$	Production: Total manufacturing
GDP	LQRSGPOR	$\Delta \ln$	GDP: Original (Index 2010 = 1.00, seasonally adjusted)
IT: Ex	XTEXVA01	$\Delta \ln$	International Trade: Total Exports Value (goods)
IT: Im	XTIMVA01	$\Delta \ln$	International Trade: Total Imports Value (goods)

Table 2.15: Data transformations, and variable definitions

Country	ISO ALPHA-3 Code	Country	ISO ALPHA-3 Code
United States of America	USA	United Kingdom	GBR
Canada	CAN	Finland	FIN
New Zealand	NZL	Sweden	SWE
Australia	AUS	France	FRA
Norway	NOR	Netherlands	NLD
Ireland	IRL	Austria	AUT
Denmark	DNK	Germany	DEU

Table 2.16: Countries and ISO Alpha-3 Codes in Macroeconomic Indices Application

				$T = 0.5\,p_1\,p_2$		$T = p_1\,p_2$		$T = 1.5\,p_1\,p_2$		$T = 2\,p_1\,p_2$	
δ_1	δ_2	p_1	p_2	$\mathcal{D}_u(\widehat{Q},Q)$	$\mathcal{D}_c(\widehat{Q},Q)$	$\mathcal{D}_u(\widehat{Q},Q)$	$\mathcal{D}_c(\widehat{Q},Q)$	$\mathcal{D}_u(\widehat{Q},Q)$	$\mathcal{D}_c(\widehat{Q},Q)$	$\mathcal{D}_u(\widehat{Q},Q)$	$\mathcal{D}_c(\widehat{Q},Q)$
0	0	20	20	1.02(0.2)	0.73(0.18)	0.73(0.12)	0.52(0.13)	0.58(0.08)	0.42(0.09)	0.5(0.07)	0.36(0.07)
		20	40	0.67(0.1)	0.47(0.11)	0.47(0.06)	0.33(0.07)	0.39(0.05)	0.27(0.06)	0.33(0.04)	0.23(0.05)
		40	20	0.71(0.1)	0.41(0.12)	0.5(0.06)	0.28(0.07)	0.41(0.05)	0.24(0.06)	0.35(0.04)	0.2(0.05)
		40	40	0.47(0.06)	0.26(0.07)	0.33(0.03)	0.18(0.04)	0.27(0.03)	0.15(0.04)	0.24(0.02)	0.13(0.03)
0.5	0	20	20	5.64(0.5)	1.92(0.74)	4.94(1.17)	1.27(0.34)	3.34(1.56)	0.98(0.22)	2.09(1.11)	0.83(0.18)
		20	40	4.86(1.19)	1.12(0.3)	1.95(1)	0.76(0.18)	1.12(0.28)	0.62(0.14)	0.89(0.17)	0.53(0.12)
		40	20	5.82(0.26)	1.23(0.44)	5.33(0.87)	0.8(0.22)	3.46(1.6)	0.66(0.18)	1.73(0.81)	0.55(0.14)
		40	40	5.37(0.81)	0.73(0.21)	1.56(0.67)	0.49(0.13)	0.96(0.2)	0.4(0.1)	0.77(0.12)	0.36(0.09)
0.5	0.5	20	20	6.81(0.34)	6.08(0.6)	6.46(0.17)	5.54(0.73)	6.32(0.13)	4.84(1.11)	6.24(0.1)	4.34(1.26)
		20	40	6.67(0.3)	5.86(0.66)	6.39(0.15)	4.93(1.08)	6.26(0.08)	4.12(1.28)	6.2(0.05)	3.47(1.3)
		40	20	6.71(0.28)	5.69(0.61)	6.4(0.13)	4.78(1.23)	6.27(0.07)	3.73(1.43)	6.2(0.05)	2.94(1.4)
		40	40	6.62(0.28)	5.15(0.98)	6.32(0.08)	3.74(1.44)	6.23(0.05)	2.7(1.43)	6.17(0.03)	2.05(1.12)

Table 2.17: Orthogonal constraints case. Means and standard deviations (in parentheses) of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$. $\boldsymbol{D}_u$ for the unconstrained model 2.1. $\boldsymbol{D}_c$ for the constrained model 2.2. All numbers in the table are 10 times of the true numbers for clear presentation. The results are based on 500 iterations.

						$T = 0.5 * p_1 * p_2$			$T = p_1 * p_2$			$T = 1.5 * p_1 * p_2$			$T = 2 * p_1 * p_2$		
δ_1	δ_2	δ_3	δ_4	p_1	p_2	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}	f_u	f_{con_1}	f_{con_2}
0	0	0	0	20	20	0	0.94	0	0	1.00	0	0	1.00	0	0.01	1.00	0
				20	40	0	1.00	0	0	1.00	0	0.03	1.00	0	0.19	1.00	0
				40	20	0.15	0.99	1.00	0.81	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
				40	40	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0	0	0.5	0	20	20	0	0.94	0	0	1.00	0	0	1.00	0	0	1.00	0
				20	40	0	1.00	0	0	1.00	0	0	1.00	0	0	1.00	0
				40	20	0	0.99	0.54	0	1.00	0.84	0	1.00	0.97	0	1.00	1.00
				40	40	0	1.00	0.98	0	1.00	1.00	0	1.00	1.00	0	1.00	1.00
0	0	0.5	0.5	20	20	0	0.94	0	0	1.00	0	0	1.00	0	0	1.00	0
				20	40	0	1.00	0	0	1.00	0	0	1.00	0	0	1.00	0
				40	20	0	0.99	0	0	1.00	0	0	1.00	0	0	1.00	0
				40	40	0	1.00	0	0	1.00	0	0	1.00	0	0	1.00	0
0.5	0	0	0	20	20	0	0.21	0	0	0.53	0	0	0.79	0	0	0.92	0
				20	40	0	0.67	0	0	0.97	0	0	1.00	0	0	1.00	0
				40	20	0	0.34	1.00	0	0.79	1.00	0	0.92	1.00	0	0.95	1.00
				40	40	0	0.87	1.00	0	0.97	1.00	0	0.99	1.00	0	0.99	1.00
0.5	0	0.5	0	20	20	0	0.21	0	0	0.53	0	0	0.79	0	0	0.92	0
				20	40	0	0.67	0	0	0.97	0	0	1.00	0	0	1.00	0
				40	20	0	0.34	0.54	0	0.79	0.84	0	0.92	0.97	0	0.95	1.00
				40	40	0	0.87	0.98	0	0.97	1.00	0	0.99	1.00	0	0.99	1.00
0.5	0	0.5	0.5	20	20	0	0.21	0	0	0.53	0	0	0.79	0	0	0.92	0
				20	40	0	0.67	0	0	0.97	0	0	1.00	0	0	1.00	0
				40	20	0	0.34	0	0	0.79	0	0	0.92	0	0	0.95	0
				40	40	0	0.87	0	0	0.97	0	0	0.99	0	0	0.99	0
0.5	0.5	0	0	20	20	0	0.07	0	0	0.04	0	0	0.01	0	0	0.01	0
				20	40	0	0.07	0	0	0.02	0	0	0.01	0	0	0.01	0
				40	20	0	0.06	1.00	0	0.01	1.00	0	0	1.00	0	0	1.00
				40	40	0	0.06	1.00	0	0	1.00	0	0	1.00	0	0.03	1.00
0.5	0.5	0.5	0	20	20	0	0.07	0	0	0.04	0	0	0.01	0	0	0.01	0
				20	40	0	0.07	0	0	0.02	0	0	0.01	0	0	0.01	0
				40	20	0	0.06	0.54	0	0.01	0.84	0	0	0.97	0	0	1.00
				40	40	0	0.06	0.98	0	0	1.00	0	0	1.00	0	0.03	1.00
0.5	0.5	0.5	0.5	20	20	0	0.07	0	0	0.04	0	0	0.01	0	0	0.01	0
				20	40	0	0.07	0	0	0.02	0	0	0.01	0	0	0.01	0
				40	20	0	0.06	0	0	0.01	0	0	0	0	0	0	0
				40	40	0	0.06	0	0	0	0	0	0	0	0	0.03	0

Table 2.18: Relative frequency of correctly estimating k_1

						$T = 0.5 * p_1 * p_2$		$T = p_1 * p_2$		$T = 1.5 * p_1 * p_2$		$T = 2 * p_1 * p_2$	
δ_1	δ_2	δ_3	δ_4	p_1	p_2	$\mathcal{D}_u(\widehat{Q}, Q)$	$\mathcal{D}_c(\widehat{Q}, Q)$	$\mathcal{D}_u(\widehat{Q}, Q)$	$\mathcal{D}_c(\widehat{Q}, Q)$	$\mathcal{D}_u(\widehat{Q}, Q)$	$\mathcal{D}_c(\widehat{Q}, Q)$	$\mathcal{D}_u(\widehat{Q}, Q)$	$\mathcal{D}_c(\widehat{Q}, Q)$
0	0	0	0	20	20	1.56(0.87)	0.57(0.1)	0.71(0.16)	0.41(0.06)	0.54(0.09)	0.33(0.04)	0.45(0.07)	0.28(0.04)
				20	40	0.71(0.33)	0.38(0.05)	0.4(0.06)	0.27(0.03)	0.32(0.04)	0.22(0.03)	0.27(0.03)	0.19(0.02)
				40	20	0.52(0.07)	0.33(0.05)	0.36(0.04)	0.24(0.03)	0.29(0.03)	0.19(0.03)	0.25(0.02)	0.17(0.02)
				40	40	0.32(0.04)	0.2(0.04)	0.22(0.02)	0.14(0.02)	0.18(0.02)	0.12(0.02)	0.15(0.01)	0.1(0.02)
0	0	0.5	0	20	20	3.68(0.04)	0.88(0.13)	3.61(0.02)	0.63(0.08)	3.59(0.02)	0.51(0.07)	3.57(0.02)	0.44(0.06)
				20	40	3.61(0.02)	0.61(0.06)	3.57(0.01)	0.43(0.04)	3.56(0.01)	0.35(0.03)	3.55(0.02)	0.3(0.03)
				40	20	3.65(0.04)	0.57(0.05)	3.58(0.05)	0.42(0.03)	3.43(0.36)	0.35(0.02)	2.78(0.94)	0.3(0.02)
				40	40	3.36(0.51)	0.33(0.03)	0.59(0.36)	0.24(0.02)	0.35(0.06)	0.2(0.02)	0.28(0.03)	0.17(0.01)
0	0	0.5	0.5	20	20	5.99(0.36)	1.88(0.51)	5.73(0.38)	1.32(0.29)	5.49(0.45)	1.06(0.19)	5.24(0.49)	0.92(0.17)
				20	40	6.67(0.32)	1.42(0.3)	6.42(0.35)	1.02(0.15)	6.24(0.34)	0.83(0.11)	6.06(0.33)	0.72(0.09)
				40	20	6.37(0.29)	1.06(0.09)	6.09(0.28)	0.8(0.06)	5.89(0.31)	0.67(0.04)	5.77(0.29)	0.59(0.04)
				40	40	6.37(0.3)	0.67(0.04)	5.95(0.29)	0.5(0.03)	5.62(0.34)	0.42(0.02)	5.26(0.46)	0.37(0.02)
0.5	0	0	0	20	20	3.72(0.19)	1.22(0.38)	3.61(0.21)	0.8(0.17)	3.55(0.21)	0.63(0.13)	3.47(0.32)	0.55(0.11)
				20	40	3.61(0.17)	0.73(0.17)	3.45(0.33)	0.49(0.1)	3.2(0.59)	0.4(0.08)	2.66(0.9)	0.35(0.06)
				40	20	3.73(0.09)	0.78(0.27)	3.64(0.06)	0.52(0.13)	3.59(0.07)	0.41(0.11)	3.56(0.09)	0.36(0.08)
				40	40	3.65(0.05)	0.46(0.13)	3.57(0.07)	0.31(0.07)	3.49(0.21)	0.26(0.06)	3.29(0.48)	0.22(0.05)
0.5	0	0.5	0	20	20	3.81(0.07)	1.4(0.34)	3.69(0.04)	0.94(0.16)	3.63(0.03)	0.75(0.12)	3.6(0.04)	0.64(0.11)
				20	40	3.67(0.03)	0.87(0.15)	3.6(0.01)	0.6(0.08)	3.57(0.02)	0.49(0.07)	3.54(0.08)	0.42(0.06)
				40	20	3.66(0.09)	0.91(0.24)	3.56(0.13)	0.63(0.11)	3.19(0.58)	0.5(0.09)	2.14(0.92)	0.44(0.07)
				40	40	3.53(0.18)	0.54(0.11)	2.3(1.01)	0.37(0.06)	0.82(0.34)	0.31(0.06)	0.57(0.11)	0.26(0.05)
0.5	0	0.5	0.5	20	20	4.91(0.48)	2.19(0.51)	4.5(0.48)	1.5(0.28)	4.22(0.4)	1.2(0.18)	3.99(0.27)	1.04(0.17)
				20	40	5.69(0.25)	1.56(0.3)	5.45(0.24)	1.11(0.14)	5.23(0.35)	0.9(0.11)	4.85(0.54)	0.78(0.09)
				40	20	5.32(0.29)	1.29(0.2)	5.21(0.28)	0.93(0.09)	4.99(0.44)	0.77(0.07)	4.67(0.56)	0.68(0.06)
				40	40	5.3(0.15)	0.79(0.09)	4.8(0.55)	0.58(0.05)	3.81(0.33)	0.49(0.04)	3.63(0.03)	0.43(0.03)
0.5	0.5	0	0	20	20	5.13(0.47)	3.76(0.4)	5.05(0.46)	3.36(0.5)	4.88(0.44)	2.97(0.68)	4.73(0.38)	2.59(0.76)
				20	40	5.44(0.46)	3.63(0.39)	5.2(0.48)	3.05(0.65)	5.01(0.45)	2.57(0.78)	4.86(0.44)	2.1(0.8)
				40	20	5.17(0.4)	3.49(0.39)	4.91(0.33)	2.93(0.77)	4.75(0.33)	2.26(0.93)	4.64(0.3)	1.82(0.89)
				40	40	5.46(0.41)	3.19(0.6)	5.17(0.36)	2.31(0.92)	4.91(0.31)	1.66(0.89)	4.75(0.29)	1.28(0.77)
0.5	0.5	0.5	0	20	20	4.59(0.31)	3.82(0.4)	4.33(0.27)	3.39(0.5)	4.15(0.21)	3(0.67)	4.05(0.16)	2.62(0.75)
				20	40	4.54(0.34)	3.66(0.39)	4.24(0.25)	3.06(0.64)	4.07(0.18)	2.59(0.78)	3.99(0.15)	2.11(0.79)
				40	20	4.3(0.23)	3.52(0.39)	4.05(0.11)	2.95(0.76)	3.94(0.06)	2.29(0.92)	3.88(0.05)	1.84(0.88)
				40	40	4.3(0.21)	3.2(0.59)	4.03(0.1)	2.32(0.92)	3.92(0.05)	1.67(0.88)	3.87(0.04)	1.29(0.77)
0.5	0.5	0.5	0.5	20	20	5.05(0.28)	4.17(0.43)	4.57(0.22)	3.59(0.48)	4.33(0.17)	3.15(0.63)	4.19(0.13)	2.75(0.72)
				20	40	4.87(0.29)	3.88(0.39)	4.42(0.18)	3.2(0.61)	4.22(0.13)	2.71(0.74)	4.1(0.1)	2.22(0.75)
				40	20	4.61(0.19)	3.63(0.37)	4.23(0.11)	3.03(0.73)	4.07(0.06)	2.37(0.88)	3.98(0.06)	1.93(0.85)
				40	40	4.25(0.13)	3.25(0.58)	4.01(0.05)	2.37(0.9)	3.91(0.03)	1.72(0.86)	3.86(0.02)	1.34(0.75)

Table 2.19: Means and standard deviations (in parentheses) of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{Q}}, \boldsymbol{Q})$. For ease of presentation, all numbers in this table are the true numbers multiplied by 10.

2.7.3 Corporate Financial Data Information

Short Name	Variable Name	Calculation
Profit.M	Profit Margin	Net Income / Revenue
Oper.M	Operating Margin	Operating Income / Revenue
EPS	Diluted Earing per share	from report
Gross.Margin	Gross Margin	Gross Proitt / Revenue
ROE	Return on equity	Net Income / Shareholders Equity
ROA	Return on assets	Net Income / Total Assets
Revenue.PS	Revenue Per Share	Revenue / Shares Outstanding
LiabilityE.R	Liability/Equity Ratio	Total Liabilities / Shareholders Equity
AssetE.R	Asset/Equity Ratio	Total Assets / Shareholders Equity
Earnings.R	Basic Earnings Power Ratio	EBIT / Total Assets
Payout.R	Payout Ratio	Dividend Per Share / EPS Basic
Cash.PS	Cash Per Share	Cash and other / Shares Outstanding
Revenue.G.Q	Revenue Growth over last Quarter	Revenue / Revenue Last Quarter - 1
Revenue.G.Y	Revenue Growth over same Quarter Last Year	Revenue / Revenue Last Year - 1
Profit.G.Q	Profit Growth over last Quarter	Profit / Profit Last Quarter - 1
Profit.G.Y	Profit Growth over same Quarter last Year	Profit / Profit Last Quarter - 1

Table 2.20: Variables in coporate financial data

Chapter 3

Modeling Dynamic Traffic Network with Matrix Factor Models: with Application to International Trade Volume Time Series

Dynamic network analysis has found an increasing interest in the literature because of the importance of different kinds of dynamic social networks, biological networks and economic networks. Most available probability and statistical models for dynamic network data are deduced from random graph theory where the networks are characterized on the node and edge level. They are often very restrictive for applications and unscalable to high-dimensional dynamic network data which is very common nowadays. In this paper, we take a different perspective: the evolving sequence of networks are treated as a time series of network matrices. We adopt a matrix factor model where the observed surface dynamic network is assumed to be driven by a latent dynamic network with lower dimensions. The linear relationship between the surface network and the latent network is characterized by unknown but deterministic loading matrices. The latent network and the corresponding loadings are estimated via an eigenanalysis of a positive definite matrix constructed from the auto-cross-covariances of the network times series, thus capturing the dynamics presenting in the network. The proposed method is able to unveil the latent dynamic structure and achieve the objective of dimension reduction. Different from other dynamic network analytical methods that build on latent variables, our approach does impose any distributional assumptions on the underlying network or any parametric forms of its covariance function. The latent network is learned directly from the data with little subjective input. The estimated low-dimensional latent network as well as the loading matrix can be used as inputs of a second stage analysis. We applied the proposed method to the monthly international

trade flow data from 1982 to 2015. The results unveil an interesting evolution of the latent trading network and the relations between the latent entities and the countries.

The remaining part of this chapter is organized as follows. In Section 3.1, we introduce the dataset of international trade flow and present some exploratory data analysis results. In Section 3.2, we introduce two factor models for network time series data and discuss their interpretations. In Section 3.3, we present an estimation procedure and the theoretical properties on the estimators. In Section 3.4, we study and compare the finite sample properties of the two proposed models on synthetic datasets. In Section 3.5, we apply the proposed factor models to the international trade flow time series from 1981 to 2015.

3.1 International Trade Data and Exploratory Analysis

3.1.1 International Trade Volume Time Series

In the following dynamic network analysis, we make use of data for multilateral imports and exports of commodity goods among 23 countries over the 1982 – 2015 period. Our trade data come from the International Monetary Fund (IMF) *Direction of Trade Statistics* (DOTS) (IMF (2017)), which provides monthly data on the country and area distribution of countries' exports and imports by their partners. The source has been widely used in international trade analysis such as the Bloomberg Trade Flow. Even though IMF-DOTS provides data from 1948-01 to present for 236 countries, the quality of data vary across time and countries. Some countries failed to report their volumes of trade in some or all years. The problem is that these missing cases are concentrated in small and underdeveloped countries or come from Communist countries. In this study, we restrict the sample to 23 countries from three major trading groups, namely NAFTA, EU and APEC, over a 408-month period from 1982-01 to 2015-12. The countries in alphabetic order are Australia, Canada, China Mainland, Denmark, Finland, France, Germany, Hong Kong, Indonesia, Ireland, Italy, Japan, Korea, Malaysia, Mexico, Netherlands, New Zealand, Singapore, Spain, Sweden, Thailand, United Kingdom, United States.

We use the import CIF data of all goods denominated in U.S. dollars since it is generally believed that they are more accurate than export ones (Durand (1953); Linnemann (1966)). This is especially true when we are interested in tracing countries of production and consumption rather than countries of consignment or of purchase and sale (Linnemann (1966)). The figures for exports are determined by imputing them from imports. For example, Canada's exports to France are given as country France imports from Canada. This calculation is done to make world total imports and exports equal. As Linnemann (1966) notes, in order to reduce the effect of incidental transactions of unusual size and of incidental difficulties in trade contract, trade flows were measured as three-month averages, rather than as direct observations of a particular month. For example, the trade flows in 2014-03 are the averages of those in 2014-02, 2014-03, and 2014-04.

3.1.2 Exploratory Analysis

The dynamic trading network can be cast into a time series of adjacency matrix that record the ties (trading volumes) between the nodes (countries) in the network. The length of our network matrix time series is 408 months. At each time, the observation is a square matrix whose rows and columns represent the same set of 23 countries. Each row (column) corresponds to an export (import) country. Each cell in the matrix contains the dollar trading volume that the exporting country exports to the importing country.

Figure 3.1 plots the time series of dollar trading volumes among top 13 countries in GDP in our dataset. These 13 countries are representative of all countries in our dataset. They falls into three major groups: Canada, Mexican, and United States compose the NAFTA group; France, Germany, Italy, Spain, and United Kingdom are in the EU group; Australia, China, Indian, Japan and Korea belong to the APEC group. Overall, all countries spent most of the years enjoying rapid growth as an accelerating wave of globalisation. The world saw largest collapse in the value of good traded at 2009 when the impact of the global financial crisis was at its worst. While the upward trends are shared among all countries, the pattern of trading are more alike

among countries within the same group. For example, the exports time series of the five European countries resembles more to each other than to the exports time series of the Asian countries.

Figure 3.1: Time series plots of the value of good traded among 13 countries over 1982 – 2015. The plots only show the patterns of the time series while the amplitudes are not comparable between plots because the range of the y-axis are not the same.

In order to illustrate the pattern of bilateral relationships, a set of four circular trading plots are shown in Figure 3.2. The direction of flow is indicated by the arrowhead. The size of the flow is determined by the width of the arrow at its base. Numbers on the outer section axis, used to read the size of trading flows, are in billions. Each plot is based on the monthly flows over 1-year period, aggregated to selected annual levels.

Note that the four plots are representative of the bilateral relationship patterns in the 1980's, 1990's, 2000's and 2010's although the plots are based on selected years.

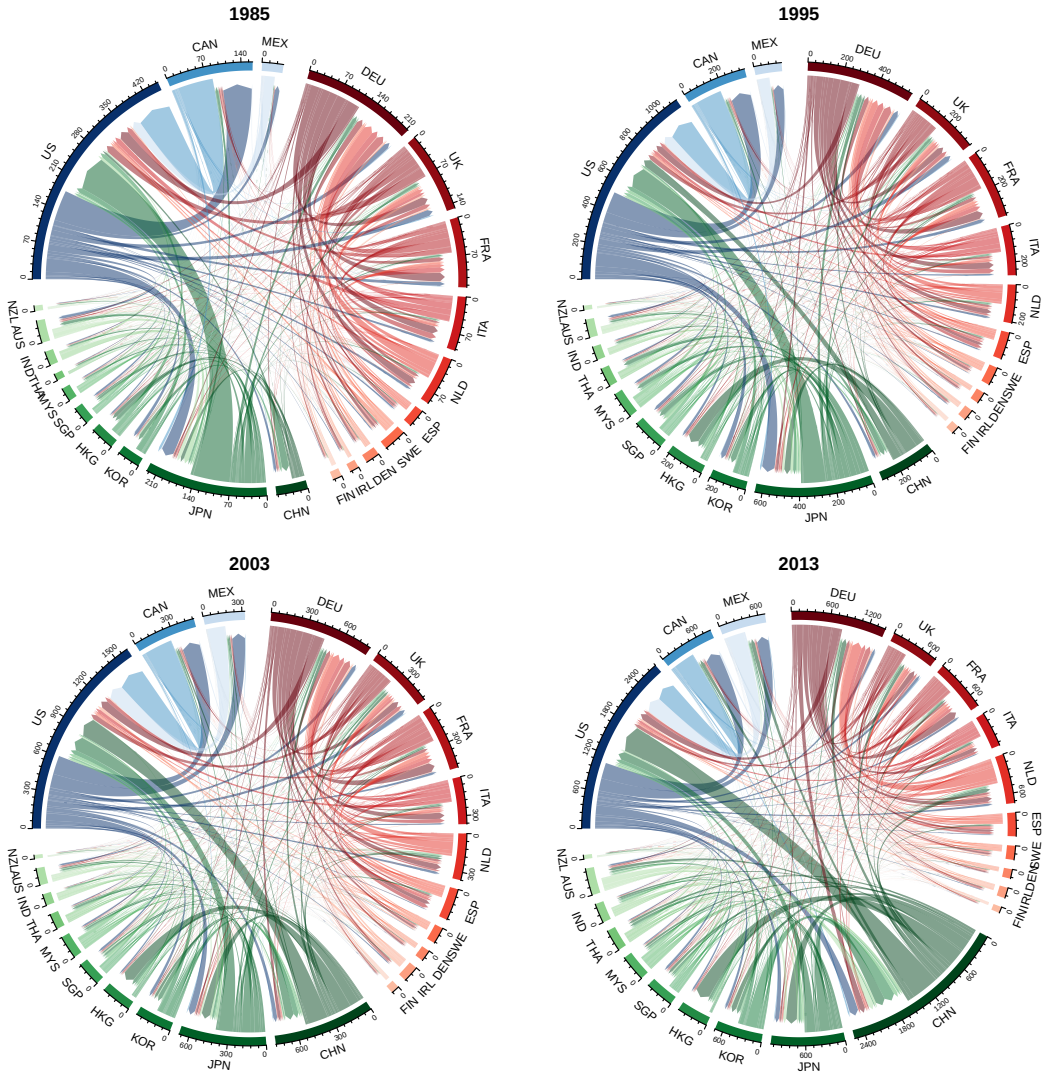


Figure 3.2: Circular trading plots that are representative of the bilateral relationship patterns in the 1980's, 1990's, 2000's and 2010's. The arrowhead indicates the direction of exports. The width of the arrow at its base represents the size of trade flow. Numbers on the outer section axis correspond to the size of trading flows in billion dollars.

For the three groups, most of the trade flows occur within the same group. This phenomenon is most prominent within the EU group where the imports and exports are all in red shade that denotes EU countries in Figure 3.2. The trade flows of NAFTA countries are least confined within the group, mainly because the U.S. alone trades a lot with both EU and APEC countries.

For individual countries, most noticeable are changes in the share and direction of trade of U.S., China, Mexico and Japan. Over years, U.S. maintains the most distinctive one among all countries because of its large size of trading volumes and wide range of trading counter-parties. The destinations of U.S. exports gradually shift from Japan and European countries to China and Mexico. In 1980's Japan accounted for the largest importing and exporting flow among APEC countries. As shown clearly in Figure 3.2, China's slice of pie in global trades grew steadily in size and becomes the largest in the 2010's. Mexico experienced a similar steady growth in global trades although less prominent than that of China. The trading patterns are most stable of the EU countries. The 20 EU countries almost keep the same portions in the size of imports and exports over years.

Based on explanatory statistical analysis and visualization tools, the aforementioned observations are mostly descriptive. Although it is clear that there exist possible lower dimensional latent networks underlying the large scale dynamic networks on the surface, there is few statistical tools available to quantify this latent structure. The new methodology that we proposed in the following section is able to quantify the latent dynamic networks that underpins the observed surface dynamic networks as well as the relationship that connect the latent networks and the surface networks. The propose methods are able to effectively reduce the dimension of the dynamic networks and uncover its core structure. The estimated latent dynamic networks and its relation with the surface networks can be used for testing and predicting the networks.

3.2 Matrix Factor Models for Dynamic Traffic Network

In this section, we propose a new methodology for investigating the evolving structure of dynamic networks. Here we focuses on the traffic flows in the dynamic network such as international import-export trade network, air-passenger volume between cities, and the number of directional interactions among people. The networks in our current considerations are typically dense. We refer to such dynamic network as dynamic traffic network. In the proposed framework, the bilateral relationships in the network at time t is recored in a relational matrix $\boldsymbol{X}_t$ whose rows and columns corresponds to the same

set of actors in the network. The elements of $\boldsymbol{X}_t$ record information of the ties between each pair of the actors. The dynamic features of the networks are characterized by the temporal dependences between consequential observations $\boldsymbol{X}_{t-1}$ and $\boldsymbol{X}_t$. The entire dynamic networks is modeled as a sequences of temporally dependent matrix-variates $\{\boldsymbol{X}_t\}_{1:T}$. An important attribute of this modeling framework is that it capture both the network structure and the temporal dynamics of the dynamic networks at a high level without any distributional assumption, comparing to the most common node-and-edge level modeling.

To formalize the methods, let $\boldsymbol{X}_t$ represent the n by n relational matrix of observed pairwise asymmetrical relationships at time t, $t = 1, \ldots, T$. A general entry of $\boldsymbol{X}_t$, denoted as $x_{ij,t}$, represents the directed relationship of actor i to actor j. For example, in international trade context $x_{ij,t}$ expresses the volume of trade flow from country i to country j at time t; in the transportation context $x_{ij,t}$ represents the fare or length of a trip from location i to location j starting at time t.

Our model for dynamic traffic network can be written as:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{F}_t\boldsymbol{A}' + \boldsymbol{E}_t, \tag{3.1}$$

where $\boldsymbol{A}$ is an $n \times r$ (vertical) matrix of "loadings" of the n actors on a relatively few r ($< n$) components or types of actors. $\boldsymbol{F}_t$ is a small, usually asymmetric, r by r matrix giving the directional relationships among the basic r types, and $\boldsymbol{E}_t$ is simply a matrix of error terms. Loading matrix $\boldsymbol{A}$ relates the observed actors to the latent types and $\boldsymbol{F}_t$ describes the interrelations among the latent types.

The general method of interpreting model (3.1) can be demonstrated by referring to the example of international trade. For discussion, let's consider a 4-dimension solution ($r = 4$). The model (3.1) would describe four basic factors underlying the pattern of international trade behavior for a given set of countries. These latent factors might be thought of as four idealized "types" of countries, types which each real country resembles to various degrees. The factors would be named by examining the loading matrix $\boldsymbol{A}$ to see which individual countries have high loadings on each factor. For example, if the individuals involved were members of major petroleum production countries, a label of

"fuel-exporting type" might be assigned to the factor on which petroleum production countries had high loadings. Other factors such as "agriculture type", "high-tech type" and "industrial type" might emerge from the analysis. Countries do not necessarily belong exclusively to a given "type". They can have moderate loadings on any given factor and high loadings on more than one factor. The factor matrix $\boldsymbol{F}_t$ does not relate to specific countries, but instead provides a general statement of the patterns of trading among the four types of countries. Each element of the $\boldsymbol{F}_t$ matrix would describe how much a given type of country generally tends to trade with another type or the same type if the diagonal element is considered. For example, $F_{ij,t}$ denotes the amount that "agriculture type" generally tend to export to "high-tech type", and $F_{ji,t}$ gives the reverse relation.

An interesting feature of the above model is that, while $\boldsymbol{F}_t$ is allowed to be asymmetric, the left and right loading matrices $\boldsymbol{A}$ are still required to be identical. This provides a description of data in terms of asymmetric relations among a *single* set of types rather than envisioning a different set of types. For example, in our international trade example model (3.1) implies that the countries have the same set of types in their "exporting" role as they have in their "importing" role. A second possible approach, where the left loading matrix may be different from the right one, can be written as:

$$\boldsymbol{X}_t = \boldsymbol{A}_1 \boldsymbol{F}_t \boldsymbol{A}_2' + \boldsymbol{E}_t, \tag{3.2}$$

where $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are the $n \times r$ (vertical) loading matrices of the n row actors and n column actors on r ($< n$) types of actors, respectively. Matrices $\boldsymbol{F}_t$ and $\boldsymbol{E}_t$ are defined the same as in those in (3.1). This formulation is the matrix factor model considered in Wang et al. (2017).

Model (3.1) describes asymmetric relationships among actors in terms of asymmetric relationships among a single set of underlying types of the actors. Model (3.2) is a more general model where there are two sets of underlying types, and the directional relationships are hypothesized to hold from types of one kind to types of the other kind. In the international trade example, model (3.1) would identify a single set of types of countries given in the loading matrix $\boldsymbol{A}$, and provide matrices $\boldsymbol{F}_t$ that describe how

much each type of country tend to trade with each of the other types. In contrast, model (3.2) provides two sets of underlying types: $\boldsymbol{A}_1$ relates to the types of the actors in their row position and $\boldsymbol{A}_2$ relates to the types of the actors in their column position. The $\boldsymbol{F}_t$ then gives the directed relationships from the row types to the column types. In the international trade example, $\boldsymbol{A}_1$ describes the amount of export-related types possessed by countries and $\boldsymbol{A}_2$ describes the amount of the import-related types possessed by countries. And $\boldsymbol{F}_t$ represents the amount that export-related type generally prompts a country to export to a particular import-related type.

When $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are not linear transformation of one another, it can easily be shown that there generally exists no solution of the form given in (3.1) for data generated by (3.2) unless one goes to a higher dimensionality. Consequently, model (3.1) makes a strong claim about a given data set. When the rows and the columns of a given directional relationship matrix can be demonstrated to span the same space, this agreement is a fact unlikely to arise by chance and probably demonstrates the validity of (3.1). With data containing noise, the row and column spaces will probably not match exactly, but a close agreement might still be interpreted as surprising the interesting. However, we will not discuss statistical tests of the fit of these two models in this article, but will demonstrate comparisons of the two models applied to a given set of real data in Section 3.5.

3.3 Estimation Procedure

Similar to all factor models, the latent factors in the proposed model (3.1) for asymmetric directional matrix time series can be linearly transformed into alternative factors with no loss of fit to the data. In general, if $\boldsymbol{H}$ is any nonsingular $r \times r$ transformation matrix, we can define an alternative $\boldsymbol{A}$ matrix, $\overset{*}{\boldsymbol{A}}$, by letting $\overset{*}{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{H}$ and defining the associated $\boldsymbol{F}_t$ matrix $\overset{*}{\boldsymbol{F}}_t = \boldsymbol{H}^{-1}\boldsymbol{F}_t\boldsymbol{H}^{'-1}$. We may assume that the columns of $\boldsymbol{A}$ are orthonormal, that is, $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ denotes the identity matrix of dimension r. Even with these constraints, $\boldsymbol{A}$ and $\boldsymbol{F}_t$ are not uniquely determined in (3.1), as aforementioned linear transformation is still valid for any orthonormal $\boldsymbol{H}$. However, the column space of the loading matrix $\boldsymbol{A}$ is uniquely determined. Hence, in what follows,

we will focus on the estimation of the column space of $\boldsymbol{A}$. We denote the factor loading spaces by $\mathcal{M}(\boldsymbol{A})$. For simplicity, we will depress the matrix column space notation and use the matrix notation directly.

To facilitate the estimation, we use the QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{W}$ to normalize the loading matrices, so that model (3.1) can be re-expressed as

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{F}_t\boldsymbol{A}' + \boldsymbol{E}_t = \boldsymbol{Q}\boldsymbol{Z}_t\boldsymbol{Q}' + \boldsymbol{E}_t, \qquad t = 1, 2, \ldots, T, \tag{3.3}$$

where $\boldsymbol{Z}_t = \boldsymbol{W}\boldsymbol{F}_t\boldsymbol{W}'$ and $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_r$.

Consider column vectors in (3.3), we write

$$X_{t,\cdot j} = \boldsymbol{A}\boldsymbol{F}_t A_{j\cdot} + E_{t,\cdot j} = \boldsymbol{Q}\boldsymbol{Z}_t Q_{j\cdot} + E_{t,\cdot j}, \qquad j = 1, 2, \ldots, n,\ t = 1, 2, \ldots, T. \tag{3.4}$$

We assume that both $\boldsymbol{F}_t$ and $\boldsymbol{E}_t$ are zero mean and thus $E(X_{t,\cdot j}) = 0$. Let h be a positive integer. For $i, j = 1, 2, \ldots, n$, define

$$\boldsymbol{\Omega}_{zq,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{Z}_t Q_{i\cdot}, \boldsymbol{Z}_{t+h} Q_{j\cdot}) \tag{3.5}$$

$$\boldsymbol{\Omega}_{x,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(X_{t,\cdot i}, X_{t+h,\cdot j}), \tag{3.6}$$

which can be interpreted as the auto-cross-covariance matrices at lag h between column i and column j of $\{\boldsymbol{Z}_t\boldsymbol{Q}'\}_{t=1,\cdots,T}$ and $\{\boldsymbol{X}_t\}_{t=1,\cdots,T}$, respectively.

For $h \geq 1$, it follows from (3.4), (3.5) and (3.6) that

$$\boldsymbol{\Omega}_{x,ij}(h) = \boldsymbol{Q}\,\boldsymbol{\Omega}_{zq,ij}(h)\,\boldsymbol{Q}'. \tag{3.7}$$

For a fixed $h_0 \geq 1$ satisfying Condition 2 in Wang et al. (2017) define

$$\boldsymbol{M}_{col} = \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\Omega}_{x,ij}(h)\boldsymbol{\Omega}_{x,ij}(h)' = \boldsymbol{Q}\left\{\sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\Omega}_{zq,ij}(h)\boldsymbol{\Omega}_{zq,ij}(h)'\right\}\boldsymbol{Q}'. \tag{3.8}$$

Similar to the column vector version, we define $\boldsymbol{M}$ matrix for the row vectors of $\boldsymbol{X}_t$'s as following

$$\boldsymbol{M}_{row} = \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\Omega}_{x',ij}(h)\boldsymbol{\Omega}_{x',ij}(h)' = \boldsymbol{Q}\left\{\sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\Omega}_{z'q,ij}(h)\boldsymbol{\Omega}_{z'q,ij}(h)'\right\}\boldsymbol{Q}', \tag{3.9}$$

where

$\mathbf{\Omega}_{z'q,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(\boldsymbol{Z}_t'Q_{i\cdot}, \boldsymbol{Z}_{t+h}'Q_{j\cdot})$ and $\mathbf{\Omega}_{x',ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} Cov(X_{t,i\cdot}, X_{t,j\cdot})$.

Finally, we define $\boldsymbol{M} = \boldsymbol{M}_{col} + \boldsymbol{M}_{row}$, that is

$$\begin{aligned}\boldsymbol{M} &= \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{\Omega}_{x,ij}(h)\mathbf{\Omega}_{x,ij}(h)' + \mathbf{\Omega}_{x',ij}(h)\mathbf{\Omega}_{x',ij}(h)'\right) \\ &= \boldsymbol{Q}\left\{\sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{\Omega}_{zq,ij}(h)\mathbf{\Omega}_{zq,ij}(h)' + \mathbf{\Omega}_{z'q,ij}(h)\mathbf{\Omega}_{z'q,ij}(h)'\right)\right\}\boldsymbol{Q}'. \quad (3.10)\end{aligned}$$

Obviously $\boldsymbol{M}$ is a $n \times n$ non-negative definite matrix. Applying the spectral decomposition to the positive definite matrix sandwiched by $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ on the right side of (3.10), we have

$$\boldsymbol{M} = \boldsymbol{Q}\left\{\sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{\Omega}_{zq,ij}(h)\mathbf{\Omega}_{zq,ij}(h)' + \mathbf{\Omega}_{z'q,ij}(h)\mathbf{\Omega}_{z'q,ij}(h)'\right)\right\}\boldsymbol{Q}' = \boldsymbol{Q}\mathbf{U}\mathbf{D}\mathbf{U}'\boldsymbol{Q}',$$

where $\mathbf{U}$ is a $r \times r$ orthogonal matrix and $\mathbf{D}$ is a diagonal matrix with diagonal elements in descending order. As $\mathbf{U}'\boldsymbol{Q}'\boldsymbol{Q}\mathbf{U} = \boldsymbol{I}_r$, the columns of $\boldsymbol{Q}\mathbf{U}$ are the eigenvectors of $\boldsymbol{M}$ corresponding to its r non-zero eigenvalues. Thus the eigenspace of $\boldsymbol{M}$ is the same as $\mathcal{M}(\boldsymbol{Q}\mathbf{U})$ which is the same as $\mathcal{M}(\boldsymbol{Q})$. Under certain regularity conditions, the matrix $\boldsymbol{M}$ has rank r. Hence, the columns of the factor loading matrix $\boldsymbol{Q}$ can be estimated by the r orthogonal eigenvectors of the matrix $\boldsymbol{M}$ corresponding to its r non-zero eigenvalues and the columns are arranged such that the corresponding eigenvalues are in the descending order.

Now we define the sample versions of these quantities and introduce the estimation procedure. Suppose we have centered the observations $\{\boldsymbol{X}_t\}_{t=1,\ldots,T}$, then for $h \geq 1$ and a prescribed positive integer h_0, let

$$\widehat{\boldsymbol{M}} = \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\widehat{\mathbf{\Omega}}_{x,ij}(h)\widehat{\mathbf{\Omega}}_{x,ij}(h)' + \widehat{\mathbf{\Omega}}_{x',ij}(h)\widehat{\mathbf{\Omega}}_{x',ij}(h)'\right), \quad (3.11)$$

where $\widehat{\mathbf{\Omega}}_{x,ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} X_{t,\cdot i}X_{t+h,\cdot j}'$ and $\widehat{\mathbf{\Omega}}_{x',ij}(h) = \frac{1}{T-h}\sum_{t=1}^{T-h} X_{t,i\cdot}X_{t+h,j\cdot}'$.

A natural estimator for the $\boldsymbol{Q}$ specified above is defined as $\widehat{\boldsymbol{Q}} = \{\widehat{\mathbf{q}}, \cdots, \widehat{\mathbf{q}}_r\}$, where $\widehat{\mathbf{q}}_i$ is the eigenvector of $\widehat{\boldsymbol{M}}$ corresponding to its i-th largest eigenvalue. Consequently,

we estimate the factors and residuals respectively by

$$\widehat{\boldsymbol{Z}}_t = \widehat{\boldsymbol{Q}}' \boldsymbol{X}_t \widehat{\boldsymbol{Q}}, \quad \text{and} \quad \widehat{\boldsymbol{E}}_t = \boldsymbol{X}_t - \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Z}}_t\widehat{\boldsymbol{Q}}' = (\boldsymbol{I}_n - \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}')\boldsymbol{X}_t + \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}'\boldsymbol{X}_t(\boldsymbol{I}_n - \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}'). \quad (3.12)$$

The above estimation procedure assumes the number of row factors r is known. To determine r we could use: (a) the eigenvalue ratio-based estimator in Lam et al. (2012); (b) the Scree plot which is standard in principal component analysis. Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_r \geq 0$ be the ordered eigenvalues of $\widehat{\boldsymbol{M}}$. The ratio-based estimator for r is defined as

$$\widehat{r} = \arg \min_{1 \leq j \leq r_{\max}} \frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j}, \quad (3.13)$$

where $r \leq r_{\max} \leq n$ is an integer. In practice we may take $r_{\max} = n/2$ or $r_{\max} = n/3$.

3.4 Simulation

In this section, we use simulation to study the performance of the estimationn methods in Section 3.3. In the simulations, the observed data $\boldsymbol{X}_{tt=1:T}$ are generated according to model (3.1),

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{F}_t\boldsymbol{A}' + \boldsymbol{E}_t, t = 1, 2, \ldots, T.$$

We choose the dimensions of the latent network $\boldsymbol{F}_t$ to be $r = 3$. The entries of $\boldsymbol{F}_t$ follow r^2 independent AR(1) processes with Gaussian while noise $\mathcal{N}(0, 1)$ innocations. Specifically, $vec(\boldsymbol{F}_t) = \boldsymbol{\Phi}_F vec(\boldsymbol{F}_{t-1}) + \boldsymbol{\epsilon}_t$ with $\boldsymbol{\Phi}_F = diag(0.86, 0.93, 0.81, 0.73, 0.62, 0.61, 0.53, 0.75, 0.78)$. The entries of $\boldsymbol{A}$ are independently sampled from uniform distribution $U(-p^{-\delta/2}, p^{-\delta/2})$ and the factor strength is controlled by parameter δ. The disturbance $\boldsymbol{E}_t$ is a white noise process with mean zero and a Kronecker product covariance structure, that is, $Cov(vec(\boldsymbol{E}_t)) = \boldsymbol{\Gamma}_2 \otimes \boldsymbol{\Gamma}_1$, where $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are both of sized $p \times p$. Both $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ have values 1 on the diagonal and 0.2 on the off-diagonal entries.

We first study the performance of our proposed approach on estimating the loading spaces. Table 3.1 shows the results for estimating the loading spaces $\mathcal{M}(\boldsymbol{A})$. The accuracies are measured by the space distance using the correct dimension of the latent network, that is $r = 3$. Estimators $\widehat{\boldsymbol{A}}_R$, $\widehat{\boldsymbol{A}}_C$ and $\widehat{\boldsymbol{A}}_{RnC}$ are estimated from $\boldsymbol{M}_{row}$, $\boldsymbol{M}_{col}$ and $\boldsymbol{M}$, respectively. The results show that with stronger signals and more data sample

points, the estimation accuracy increases. Moreover, estimator from the combination of row and column information $\boldsymbol{M}$ is the best among three in the sense that it is the closest to the truth.

		$T = 0.5\,n^2$			$T = n^2$			$T = 1.5\,n^2$			$T = 2\,n^2$		
δ	n	$\mathcal{D}(\widehat{A}_R, A)$	$\mathcal{D}(\widehat{A}_C, A)$	$\mathcal{D}(\widehat{A}_{RnC}, A)$	$\mathcal{D}(\widehat{A}_R, A)$	$\mathcal{D}(\widehat{A}_C, A)$	$\mathcal{D}(\widehat{A}_{RnC}, A)$	$\mathcal{D}(\widehat{A}_R, A)$	$\mathcal{D}(\widehat{A}_C, A)$	$\mathcal{D}(\widehat{A}_{RnC}, A)$	$\mathcal{D}(\widehat{A}_R, A)$	$\mathcal{D}(\widehat{A}_C, A)$	$\mathcal{D}(\widehat{A}_{RnC}, A)$
0	20	0.27(0.05)	0.46(0.14)	0.21(0.04)	0.17(0.03)	0.21(0.04)	0.12(0.02)	0.08(0.01)	0.12(0.02)	0.06(0.01)	0.11(0.02)	0.14(0.02)	0.08(0.01)
0	40	0.08(0.01)	0.10(0.01)	0.06(0.01)	0.06(0.01)	0.07(0.01)	0.04(0.01)	0.04(0.00)	0.05(0.01)	0.03(0.00)	0.04(0.00)	0.05(0.01)	0.03(0.00)
0	60	0.03(0.00)	0.05(0.01)	0.03(0.00)	0.03(0.00)	0.04(0.00)	0.02(0.00)	0.02(0.00)	0.03(0.00)	0.02(0.00)	0.02(0.00)	0.03(0.00)	0.02(0.00)
0.5	20	5.54(0.08)	5.76(0.05)	5.61(0.07)	5.08(0.27)	5.60(0.07)	4.81(1.03)	1.30(0.51)	4.98(0.21)	1.24(0.55)	2.39(0.38)	3.49(0.22)	2.43(0.29)
0.5	40	5.70(0.07)	5.65(0.07)	5.59(0.04)	5.54(0.11)	5.13(0.08)	5.50(0.15)	4.18(0.93)	5.66(0.09)	4.22(0.6)	5.70(0.16)	5.76(0.03)	5.75(0.02)
0.5	60	5.69(0.06)	5.49(0.03)	5.71(0.02)	2.79(0.61)	5.6(0.02)	2.50(0.91)	5.11(0.17)	5.59(0.02)	5.19(0.1)	4.71(0.66)	5.08(0.05)	4.88(0.22)

Table 3.1: Means and standard deviations (in parentheses) of the estimation accuracy measured by $\mathcal{D}(\widehat{A}, A)$. For ease of presentation, all numbers in this table are the true numbers multiplied by 10. The results are average of 200 simulations.

Now we present the performance of our proposed approach on estimating the dimension of the latent network $r = 3$. In table 3.2, f_R, f_C, and $f_R nC$ represents the frequency of correctly estimating the dimension using $\boldsymbol{M}_{row}$, $\boldsymbol{M}_{col}$ and $\boldsymbol{M}$, respectively. Again, the results show that with stronger signals and more data sample points, the estimation accuracy increases. Moreover, estimator from the combination of row and column information $\boldsymbol{M}$ is the best among three in the sense that it has the highest frequency of correctly estimating the number of latent dimensions.

		$T = 0.5\,n^2$			$T = n^2$			$T = 1.5\,n^2$			$T = 2\,n^2$		
δ	n	f_R	f_C	f_{RnC}	f_R	f_C	f_{RnC}	f_R	f_C	f_{RnC}	f_R	f_C	f_{RnC}
0	20	0.975	0.3	0.965	0.99	0.72	1	1	1	1	1	1	1
0	40	1	1	1	1	1	1	1	1	1	1	1	1
0	60	1	1	1	1	1	1	1	1	1	1	1	1
0.5	20	0.72	0.085	0.805	0.345	0.3	0	0	0	0	0.005	0.005	0.21
0.5	40	0	0	0.06	0	0	0	0	0	0	0	0	0
0.5	60	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.2: Relative frequencies of correctly estimating the dimension of the latent network. The results are based on 200 simulations.

3.5 Application to International Trade Volume Time Series

By examining the network of international trade, we will show in the following text that we can analyze how countries compare to each other in terms of trade volumes and patterns and how these volumes and patterns evolve as economical cycles and political events unfold. We want to emphasize that our analysis does not draw on

aggregate country statistics such as GNP, production statistics or any other national attributes.

3.5.1 Five-Year Rolling Estimation

To allow for structural changes over time, we break the 408-month period into 30 rolling 5-year periods: 1982 through 1986, 1983 through 1987 and so forth. For each 5-year period, we assume that the loadings are constant and estimate the loading matrix $\boldsymbol{A}$ under model (3.1) and $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ under model (3.2). Fixing the number of factors r, for each of the 30 periods, we estimate 3 loading matrices $\boldsymbol{A}$, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, whose dimensions are $24 \times r$. We index these matrices by the mid-year of the five-year periods. For example, $\boldsymbol{A}$ for period 1982–1986 is indexed with year 1984, $\boldsymbol{A}$ for period 1983–1987 is indexed with year 1985 and so forth.

As noted in Section 3.3, we can only identify the column spaces of the loading matrices because of the rotational indeterminacy. Let $\boldsymbol{A}$ be a matrix whose columns constitute a set of basis of the loading space, then the totality of matrices that represent the column spaces of the loading matrices is $\{\boldsymbol{AH} \mid \boldsymbol{H}$ is any nonsingular $r \times r$ matrix$\}$. Which $\boldsymbol{H}$ we select can depend on which perspective we wish to take toward the interpretation of $\boldsymbol{A}$ and $\boldsymbol{F}_t$. Although in general we might like to seek some kind of approximate simple structure for the columns of $\boldsymbol{A}$, this can be done in different ways, corresponding roughly to different orthogonal or oblique rotation criteria in factor analysis.

In the analyses presented in this article, we will adopt as standard a procedure which applies Varimax to the columns of $\boldsymbol{A}$ *after* they have been scaled to have equal sums of squares; this keeps the columns of $\boldsymbol{A}$ mutually orthogonal. We further standardize the columns of $\boldsymbol{A}$ so that they sum to one. This is feasible because we are dealing with data which contain all positive values, and our columns of $\boldsymbol{A}$ will contain mostly positive entries with only few negative ones. At this moment, we interpret negative entries in $\boldsymbol{A}$ as that an actor load less on a latent type. We note that non-negative matrix decomposition can be employed further to make $\boldsymbol{A}$ with all positive entries.

When the columns of $\boldsymbol{A}$ are standardized to have sums equal to one, the factor

	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Ratio	2	4	1	1	1	1	5	1	2	2	2	2	2	2	2
Scree	2	3	4	4	4	5	5	5	4	3	3	3	3	2	2
$r = 4$	97	94	91	89	85	83	83	84	88	91	90	91	93	94	94
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Ratio	2	2	2	2	2	2	2	2	6	5	2	2	2	2	2
Scree	3	4	3	4	3	3	3	4	4	4	4	4	3	3	4
$r = 4$	90	89	91	91	91	91	90	88	86	86	88	90	92	92	88

Table 3.3: Comparison of estimated latent dimension of $\boldsymbol{F}_t$ in model (3.1) between ratio-based and scree plot methods. Scree plot method chooses the minimal dimension that account for at least 85% variance of the original data. The last line presents the percentage of total variance explained by the $r = 4$ factor model.

matrix $\boldsymbol{F}_t$ can be thought of as a compressed or miniature version of the original observation matrix $\boldsymbol{X}_t$. The sum of all the elements in $\boldsymbol{F}_t$ is equal to the sum of all elements in $\widehat{\boldsymbol{X}}_t$, the part of $\boldsymbol{X}_t$ fit by the model. The factor matrix $\boldsymbol{F}_t$ can be interpreted as expressing relationships among the latent factors in the same units as the original data. That is, the factor matrix $\boldsymbol{F}_t$ can be interpreted as one of the same kind as the original data matrix $\boldsymbol{X}_t$, but describing the relations among the latent types of the actors, rather than the actors themselves. The diagonals for the observed relational matrices $\boldsymbol{X}_t$ are undefined, and will be ignored in the analysis by setting their values to zero. The diagonals for the latent factor matrices $\boldsymbol{F}_t$ can be interpreted as the relationship within the same type, e.g. the import-export between European countries.

3.5.2 Results

We apply the model (3.1) described in Section 3.3 and 3.5.1 to the level of international trade volume data. We use the ratio-based method in (3.13) as well as scree plot to estimate the number of latent dimensions. The comparison between these two methods of estimating latent dimensions in different time periods is shown in Table 3.3. The scree plot method selects the minimal number of dimension that explain at least 85 percents of the variance in the original data. The estimator by (3.13) tends to be smaller than the one given by scree plot. The percentage of total variance explained by the $r = 4$ factor model is shown in the last line.

As shown in Table 3.3, most dimension estimators are smaller than 4 and the factor model with $r = 4$ explains at least 83% of the total variance. Thus, latent dimension

$r = 4$ will be used for illustration. We will focus on the loading matrix $\boldsymbol{A}$, which prescribes the interpretations of the latent types by linking them to the observed actors, and the factor matrix $\boldsymbol{F}_t$, which characterizes the directional relationship between latent types. For visualization, we employ heat map for loading matrix $\boldsymbol{A}$ and network plot for the factor matrix $\boldsymbol{F}_t$. We also use the dendrogram to show the clustering of countries based on their loadings on the latent types. All the plotted values can be found in the supplemental materials.

The estimated loading matrix $\widehat{\boldsymbol{A}}$ has been rotated by Varimax approximation to a simple structure, and its columns are kept orthonormal. We then scale each column of $\widehat{\boldsymbol{A}}$ such that its sum is one. There exist negative values in estimated loadings $\widehat{\boldsymbol{A}}$. However, they are very close to zero and occur rarely, thus we set all negative values to zeros. See supplementary material for plotted values.

Figure 3.3 presents the heat maps of the loadings on top four latent types from 1984 to 2013. Four vertically aligned heat maps correspond to four columns of loading matrix $\widehat{\boldsymbol{A}}$ from year 1984 to 2013. For example, the first columns (denoted by 1984) of the plot (a), (b), (c), and (d) are the four columns of the loading matrix $\widehat{\boldsymbol{A}}_{1984}$ calculated using data from 1982 to 1986; the second columns (denoted by 1985) of the four heat maps correspond to the four columns of the loading matrix $\widehat{\boldsymbol{A}}_{1985}$ calculated using data from 1983 to 1987; and so on.

Most eigen-decomposition algorithms estimate $\widehat{\boldsymbol{A}}$ with columns ranked according to their corresponding eigen-values (accounted variances). The structure of international trade changes over time. The latent factors or types may rank differently in terms of their accounted variances at different time periods. For example, latent type of European countries may account for the largest portion of variance in 1985, but it may rank 3rd in 2001 and even no longer belong to the top four types in 2009. To present the same latent factors or types over time in one heat map, we align the columns of $\widehat{\boldsymbol{A}}$ from different years according to their maximum loading on the United States, United Kingdom, and China for plots (a), (b) and (c). Plot (d) contains the remaining factor for all the years. In such representation, plots (a),(b),(c) and (d) are considered together as top four types without ranking with respect to the accounted variance within them.

The factors in one heat map may ranked differently in terms of accounted variance at different times. But they correspond to the same interpretation at certain time periods.

Recall that each column in a heat map sums up to one. Thus, the value at each cell denotes a country's participation in a factor or type at a certain year. For example, the darkest cell corresponds to USA at year 1984 in plot (a) indicates that portion of trading taken by USA on latent type (a) is larger than those taken by all other countries. The changes of color intensity of the cells shows the evolution in a country's participation in the top four factors over 30 years.

The latent factor corresponding to Figure 3.3 (a) can be interpreted as representing the United States, as the loadings of the United States on this dimension dominate all other countries. From the plot, it is clear that the United States dominates the first dimension from 1984 to 2013. However, its participation in the first dimension gradually decreases since 2002 and reaches its minimal from year 2009 onwards, signaling the aftermath of the 2008 financial crisis. The decrease from United States is offset by increase from United Kingdom, Netherlands, Hong Kong, Japan, and Korea, which is manifested by the increasingly darker cells since 2002 for those countries.

The latent factor corresponding to Figure 3.3 (b) are aligned according to the maximum loading on United Kingdom, and not surprisingly, they are also heavily loaded on European countries such as France, Italy, Netherlands, Spain and Germany. Therefore, this dimension can be interpreted as representing European countries. From 1985 to 1989, Germany's trading was so distinctive from other European countries that it took a separate dimension as shown in Figure 3.3 (d). During this period, France, United Kingdom, Italy and Netherlands accounted for large portions of European's trading. After 1990, Germany, France, United Kingdom, and Italy took approximately equal portions. With the introduction of Euro in 2002, Netherlands, Spain, and United Kingdom's participations in trade increase. We should also note that the loading of some Asian economies, such as Hong Kong, Japan, Malaysia and Singapore, on this dimension is also significant in certain periods such as from 1992 to 1994, and from 2008 onwards. This suggests that, in these periods, the factor representing Asian economies explain more variance in the original data than the European factor and replace European

factor as one of the top four factor.

Figure 3.3: Latent factor loadings for trading level on $r = 4$ dimensions for a series of 30 rolling five-year periods indexed from 1984 to 2013.

The latent factor corresponding to Figure 3.3 (c) are factors that China Mainland has maximum loadings on. Before 1989, Japan loads more on this dimension than

China does. China's loading on this dimension keeps increasing all the time. Its value becomes larger than Japan's loading from the year 1989. It shows a clearer transition of trading centrality of large Asia economies.

The latent factor corresponding to Figure 3.3 (d) features sizable loadings on Canada, Mexico, Japan and Korea. Thus the fourth dimension of the latent factor matrix represents the group of large economies in North American and Asia except for the US and China. The evolution of the dimension (d) is striking. Before 1989, Germany's trading is so distinctive from the other European countries that it dominates this single dimension. After that, this dimension is dominated by NAFTA countries from 1990 to 2000 and from 2007 to 2012 and by APEC countries from 2001 to 2007.

Figure 3.4 plots the trading network among four latent types as well as the relationship between countries and latent types for four selected years. The trading network among latent types is plotted based on the average of 4×4 latent factor matrix $\boldsymbol{F}_t$ in the corresponding 5-year rolling window. The colored circles represent 4 latent dimensions. Note that the eigen-decomposition algorithm we used does not guarantee positive entries in $\boldsymbol{F}_t$. The negative values in $\boldsymbol{F}_t$ are interpreted as a change of trading direction. Non-negative matrix factorization proposed by Lee and Seung (2001) can be used to eliminate negative entries. The size of each circle conveys the trading volumes within each latent dimension, i.e., the values of the diagonal elements in the latent factor matrix. The width of the solid lines connecting circles conveys the trading volume between different latent dimensions, i.e., the values of the off-diagonal elements in the latent factor matrix. The direction of the flow is conveyed by the color of the line. Specifically, the color of the line is the same as its export dimension. For example, a blue line connecting a blue node and a red node represents the trade flow from the blue node to the red node. Note that the widths of the solid lines across different network plots are not comparable because they are scaled to fit each individual plot, otherwise the lines in the 2015 plot will overwhelm the whole plot because the trading volume is much larger in 2015 than in 1985.

The relationships between countries and $r = 4$ latent dimensions, shown as the dotted lines, are plotted using a simplified version of the estimated loading matrix $\widehat{\boldsymbol{A}}$

to provide an uncluttered view that only captures the prominent relations. Specifically, we generate a base matrix by rounding $10\widehat{\boldsymbol{A}}$. We set all non-dominating entries to zero for each row (country) of the base matrix, and then re-weight the non-zero entries such the sum of row is 1. We alternate between the eliminating and re-weighting steps until no changes occur. The non-dominating entries for each row are defined as values that are more than 0.5 smaller than the maximum entry of the row. The countries with zero loadings in the resulting matrix are not plotted. The size of the dotted line conveys the strength of connection between a country and a latent dimension.

Clearly shown in the network plot, the United States (node #2) and Germany (node #4) stands out as two single dimensions in 1985. Latent dimension node #3 is composed of European countries such as Spain, Netherlands, France, Sweden, United Kingdom and Italy. Latent dimension node #3 is composed of Japan, Korea, China and Canada. As shown by the thick orange lines, node #2, representing the U.S., exports mostly to node #1, which load mostly on large Asian countries and Canada. The thick pink and purple lines connecting nodes #3 and #4 implies that Germany trades a lot with other European countries even through itself stands out from the European countries.

In 1995, European countries become closer and they form a single dimension node #3, which reflects the effects the foundation of European Union in 1993. The within group trading is largest in European countries. The year of 1995 also celebrates developments of Asian countries when they dominate two latent types, namely node #1 and #2. This can be explained by the fast development of these Asian countries to emulate the developed economies in North American and European economies during the late 80's and early 90's. There are large amount of exporting from Asian countries to the United States and European countries as indicated by the thick pink and green lines to node #2 and node #3. Also, the trading among Asian countries is also large as shown by the think lines connection green node #1 and pink node #2.

In 2003, factor #1 is composed of Canada, China and Mexico. It represents the latent type that exports a lot to factor #2 (Hong Kong and United States). Factor #3 that composed of Netherlands, France, Italy, Spain, United Kingdom stays the same as the European type in 1995. Factor #4 can be interpreted as APEC type because it is

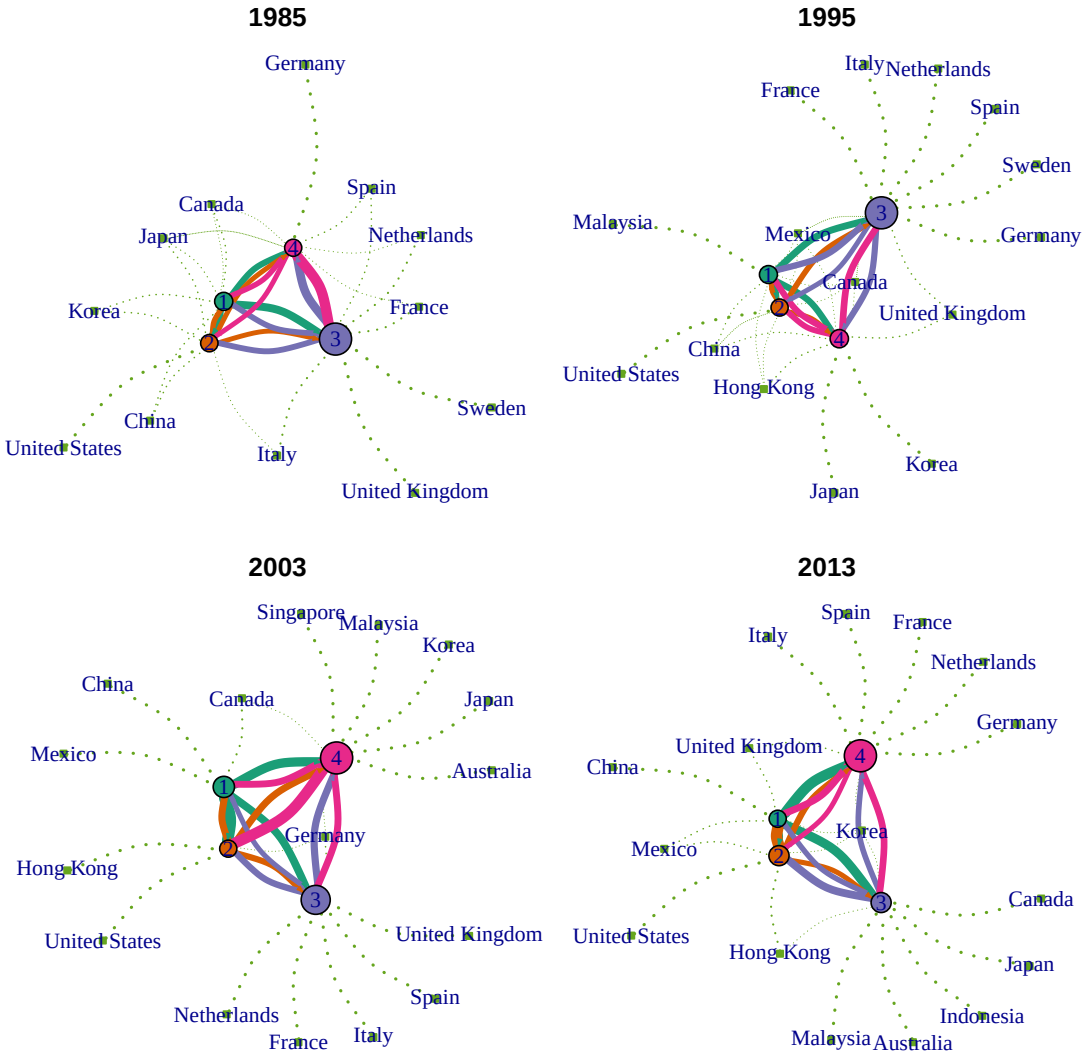


Figure 3.4: Trading level network plot of latent dimensions and relationship between countries and the latent dimensions. Thickness of the solid line represents the volume of trades among latent dimensions. Thickness of the dotted lines represents the level of connection between latent dimensions and countries. Note that a country can be related to multiple latent dimensions.

composed of Australia, Japan, Korea, Malaysia and Singapore. The United States still loaded completely on factor #2. However, Hong Kong also load heavily on this factor, indicating that these two countries share some similar import/export pattern as that of US. For example, Hong Kong trades a lot with the Canada, China and Mexico (the thick orange and green lines between nodes #1 and #2) and it also imports a large volume from the APEC type #4. The exporting volumes from factor #1 (Canada, China and Mexico) to factor #2 and from APEC type node #4 to factor #2 are among the largest trading volumes in this period.

In 2013, China dominates a single factor #1, indicating China's growing importance in international trade in the 2010's. The European dimension – factor #4 – does not change from 2003. However, the within Europe trading volume (the size of pink node #4) increase a lot compared with that in 2003. The United States still loads completely on factor #2. But it shares this dimension with Mexico and Hong Kong which also trade heavily with China.

Figure 3.5 shows the clustering of countries based on their loadings on first four latent dimensions over years. The rectangles denotes clusters that divide countries into six groups. It offers a new perspective to inspect the dynamics of countries' trading behaviors. The United States accounts for a single factor for all years because of its large trading volumes with other countries. European countries fall into the same group for most of the time while Germany stands out differently some time from other European countries. China's weight in the global trade over the years has been gradually increasing: in 1985 China's trading behavior is more like economies such as Korea. However, from 1990's to 2010's, as China's trade becomes more active, its trading behavior becomes more similar to that of the United States and it makes up single cluster. Again, these patterns echo with some of the observations from Figures 3.3 and 3.4.

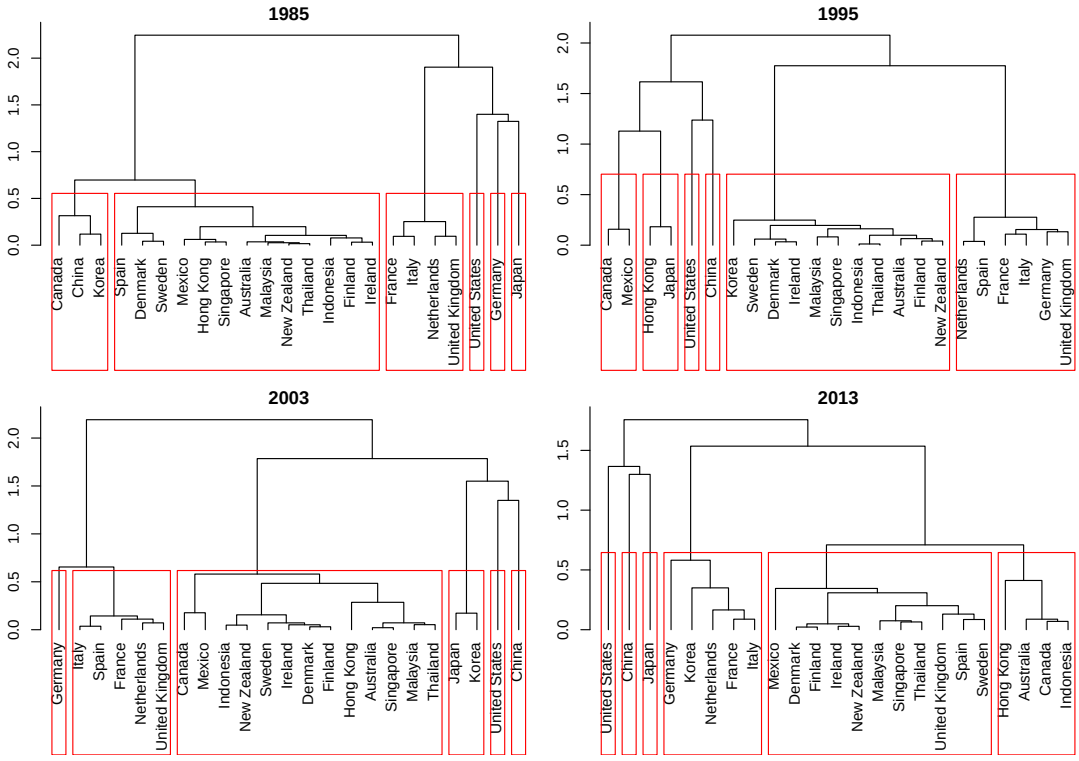


Figure 3.5: Clustering of countries based on their trading level latent dimension representations.

Chapter 4

Factor Models for Multivariate Spatial-Temporal Process

Multivariate spatio-temporal data arise more and more frequently in a wide range of applications; however, there are relatively few general statistical methods that can readily use that incorporate spatial, temporal and variable dependencies simultaneously. In this paper, we propose a new approach to represent non-parametrically the linear dependence structure of a multivariate spatio-temporal process in terms of latent common factors. The matrix structure of observations from the multivariate spatio-temporal process is well reserved through the matrix factor model configuration. The spatial loading functions are estimated non-parametrically by sieve approximation and the variable loading matrix is estimated via an eigen-analysis of a symmetric non-negative definite matrix. Though factor decomposition along the space mode is similar to the low-rank approximation methods in spatial statistics, the fundamental difference is that the low-dimensional structure is completely unknown in our setting. Additionally, our method accommodates non-stationarity over space. The estimated loading functions facilitate spatial prediction. For temporal forecasting, we preserve the matrix structure of observations at each time point by utilizing the matrix autoregressive model of order one MAR(1). Asymptotic properties of the proposed methods are established. Performance of the proposed method is investigated on both synthetic and real datasets.

The remainder of the chapter is outlined as follows. Section 4.1 introduces the model settings. Section 4.2 discusses estimation procedures for loading matrix and loading functions. Section 4.3 discuss the procedures for kriging and forecasting over space and time, respectively. Section 4.4 presents the asymptotic properties of the estimators. Section 4.5 illustrates the proposed model and estimation scheme on a synthetic dataset; And finally Section 4.6 applies the proposed method to a real dataset.

Technique proofs are relegated to the Appendix.

4.1 The Model

Consider a p-dimension multivariate spatio-temporal process $\boldsymbol{y}_t(\boldsymbol{s}) = (y_{t,1}(\boldsymbol{s}), \ldots, y_{t,p}(\boldsymbol{s}))'$

$$\boldsymbol{y}_t(\boldsymbol{s}) = \boldsymbol{C}'(s)\boldsymbol{z}_t(\boldsymbol{s}) + \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\epsilon}_t(\boldsymbol{s}), \quad t = 0, \pm 1, \pm 2, \cdots, \ \boldsymbol{s} \in \mathcal{S} \subset \mathcal{R}^2, \tag{4.1}$$

where $\boldsymbol{z}_t(\boldsymbol{s})$ is an $m \times 1$ observable covariate vector, $\boldsymbol{C}(s)$ is a $m \times p$ unknown parameter matrix, the additive error vector $\boldsymbol{\epsilon}_t(\boldsymbol{s})$ is unobservable and constitutes the nugget effect over space in the sense that

$$\mathrm{E}\{\boldsymbol{\epsilon}_t(\boldsymbol{s})\} = \boldsymbol{0}, \quad \mathrm{Var}\{\boldsymbol{\epsilon}_t(\boldsymbol{s})\} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{s}), \quad \mathrm{Cov}\{\boldsymbol{\epsilon}_{t_1}(\boldsymbol{u}), \boldsymbol{\epsilon}_{t_2}(\boldsymbol{v})\} = \boldsymbol{0} \ \forall \ (t_1, \boldsymbol{u}) \neq (t_2, \boldsymbol{v}), \tag{4.2}$$

$\boldsymbol{\xi}_t(\boldsymbol{s})$ is a p-dimension latent spatio-temporal vector process satisfying the condtions

$$\mathrm{E}\{\boldsymbol{\xi}_t(\boldsymbol{s})\} = \boldsymbol{0}, \quad \mathrm{Cov}\{\boldsymbol{\xi}_{t_1}(\boldsymbol{u}), \boldsymbol{\xi}_{t_2}(\boldsymbol{v})\} = \boldsymbol{\Sigma}_{|t_1 - t_2|}(\boldsymbol{u}, \boldsymbol{v}). \tag{4.3}$$

Under the above condtions, $\boldsymbol{y}_t(\boldsymbol{s}) - \boldsymbol{C}'(s)\boldsymbol{z}_t(\boldsymbol{s})$ is seond order stationary in time t,

$$\begin{gathered}
\mathrm{E}\{\boldsymbol{y}_t(\boldsymbol{s}) - \boldsymbol{C}'(s)\boldsymbol{z}_t(\boldsymbol{s})\} = \boldsymbol{0}, \\
\mathrm{Cov}\{\boldsymbol{y}_{t_1}(\boldsymbol{u}) - \boldsymbol{C}'(\boldsymbol{u})\boldsymbol{z}_{t_1}(\boldsymbol{u}), \boldsymbol{y}_{t_2}(\boldsymbol{v}) - \boldsymbol{C}'(\boldsymbol{v})\boldsymbol{z}_{t_2}(\boldsymbol{v})\} \\
= \boldsymbol{\Sigma}_{|t_1 - t_2|}(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(\boldsymbol{u}) \cdot \mathrm{I}\{(\mathrm{t}_1, \boldsymbol{u}) = (\mathrm{t}_2, \boldsymbol{v})\}.
\end{gathered}$$

Finally, we assume that $\boldsymbol{\Sigma}_t(\boldsymbol{u}, \boldsymbol{v})$ is continuous in $\boldsymbol{u}$ and $\boldsymbol{v}$. Note that model (4.1) does not impose any stationary conditions over space, though it requires that $\boldsymbol{y}_t(\boldsymbol{s})$ is second order stationary in time t.

We assume that the latent spatial-temporal vector process are driven by a lower-dimention latent spatial-temporal factor process, that is

$$\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{f}_t(\boldsymbol{s}), \tag{4.4}$$

where $\boldsymbol{f}_t(\boldsymbol{s})$ is the r-dimensional latent factor process ($r \ll p$) and $\boldsymbol{B}$ is the $p \times r$ loading matrix.

Further, we assume that the latent $r \times 1$ factor process $\boldsymbol{f}_t(\boldsymbol{s})$ admits a finite functional structure,

$$\boldsymbol{f}_t(\boldsymbol{s}) = \sum_{j=1}^{d} a_j(\boldsymbol{s})\boldsymbol{x}_{tj}, \tag{4.5}$$

where $a_1(\cdot), \cdots, a_d(\cdot)$ are deterministic and linear independent functions (i.e. none of them can be written as a linear combination of the others) in the Hilbert space $L_2(\mathcal{S})$, and $\boldsymbol{x}_{tj} = (\boldsymbol{x}_{tj,1}, \ldots, \boldsymbol{x}_{tj,r})$ is a $r \times 1$ random vector. Combining (4.4) and (4.5), we have

$$\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B}\sum_{j=1}^{d} a_j(\boldsymbol{s})\boldsymbol{x}_{tj} = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}), \tag{4.6}$$

where $\boldsymbol{X}_t = (\boldsymbol{x}_{t1}, \cdots, \boldsymbol{x}_{td})'$ and $\boldsymbol{a}(\boldsymbol{s}) = (a_1(\boldsymbol{s}), \cdots, a_d(\boldsymbol{s}))'$.

Stacking $\boldsymbol{\xi}_t(\boldsymbol{s})$ from n locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ together as rows, we have a $n \times p$ matrix of p signals from n locations $\boldsymbol{\Xi}_t = (\boldsymbol{\xi}_t(\boldsymbol{s}_1), \cdots, \boldsymbol{\xi}_t(\boldsymbol{s}_n))'$. It follows from (4.6) that

$$\boldsymbol{\Xi}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}', \tag{4.7}$$

where $\boldsymbol{A} = [A_{ij}] = [a_j(\boldsymbol{s}_i)]$, $i = 1, \ldots, n$ and $j = 1, \ldots, d$.

Obviously $a_1(\cdot), \cdots, a_d(\cdot)$ are not uniquely defined by (4.5) and $\boldsymbol{B}$ is not uniquely defined by (4.4). We assume that $a_1(\cdot), \cdots, a_d(\cdot)$ are orthonormal in the sense that $\langle a_j\,, a_k\rangle = \mathrm{I}\{\mathrm{j} = \mathrm{k}\}$ and $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_r$. Thus, the kernel reproducing Hilbert space (KRHS) spanned by $a_1(\cdot), \cdots, a_d(\cdot)$ and the vector space spanned by columns of $\boldsymbol{B}$ (i.e. $\mathcal{M}(\boldsymbol{B})$) are uniquely defined. We estimate the KRHS and $\mathcal{M}(\boldsymbol{B})$ in this article.

4.2 Estimation

Let $\{(\boldsymbol{y}_t(\boldsymbol{s}_i), \boldsymbol{z}_t(\boldsymbol{s}_i))\,, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T\}$ be the available observations over space and time, where $\boldsymbol{y}_t(\boldsymbol{s}_i)$ is a vector of p variables and $\boldsymbol{z}_t(\boldsymbol{s}_i)$ is a vector of m covariates observed at location $\boldsymbol{s}_i$ at time t. In this article, we restrict attention to the isotopic case where all variables have been measured at the same sample locations $\boldsymbol{s}_i$, $i = 1, \ldots, n$.

To simplify the notation, we first consider a special case where $\boldsymbol{C}(\boldsymbol{s}) \equiv \boldsymbol{0}$ in (4.1). Now the observations are from the process

$$\boldsymbol{y}_t(\boldsymbol{s}) = \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\epsilon}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}) + \boldsymbol{\epsilon}_t(\boldsymbol{s}). \tag{4.8}$$

Stacking $\boldsymbol{y}_t(\boldsymbol{s}_i)$, $i = 1, \ldots, n$ together as rows, we have

$$\boldsymbol{Y}_t = \boldsymbol{\Xi}_t + \boldsymbol{E}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}' + \boldsymbol{E}_t, \tag{4.9}$$

where $\boldsymbol{Y}_t = (\boldsymbol{y}_t(\boldsymbol{s}_1), \cdots, \boldsymbol{y}_t(\boldsymbol{s}_n))$ and $\boldsymbol{E}_t = (\boldsymbol{\epsilon}_t(\boldsymbol{s}_1), \cdots, \boldsymbol{\epsilon}_t(\boldsymbol{s}_n))'$.

4.2.1 Estimation of the Partitioned Spatial Loading Matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$

To exclude nugget effect in our estimation, we divide n locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ into two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ with n_1 and n_2 elements respectively. Let $\boldsymbol{Y}_{lt}$ be a matrix consisting of $\boldsymbol{y}_t(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S}_l$, $l = 1, 2$ as rows. Then $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$ are two matrices of dimention $n_1 \times p$ and $n_2 \times p$ respectively. It follows from (4.8) that

$$\boldsymbol{Y}_{1t} = \boldsymbol{\Xi}_{1t} + \boldsymbol{E}_{1t} = \boldsymbol{A}_1\boldsymbol{X}_t\boldsymbol{B}' + \boldsymbol{E}_{1t}, \qquad \boldsymbol{Y}_{2t} = \boldsymbol{\Xi}_{2t} + \boldsymbol{E}_{2t} = \boldsymbol{A}_2\boldsymbol{X}_t\boldsymbol{B}' + \boldsymbol{E}_{2t}, \tag{4.10}$$

where $\boldsymbol{A}_l$ is a $n_l \times d$ matrix, its rows are $(a_1(\boldsymbol{s}), \ldots, a_d(\boldsymbol{s}))$ at diffent locations $\boldsymbol{s} \in \mathcal{S}_l$ and $\boldsymbol{E}_{t,l}$ consists of $\boldsymbol{\epsilon}_t(\boldsymbol{s})$ as rows with $\boldsymbol{s} \in \mathcal{S}_l$, $l = 1, 2$.

For model identification, we assume $\boldsymbol{A}_1'\boldsymbol{A}_1 = \boldsymbol{I}_d$ and $\boldsymbol{A}_2'\boldsymbol{A}_2 = \boldsymbol{I}_d$, which however implies that $\boldsymbol{X}_t$ in the second equation in (4.10) will be different from that in the first eqaution. Thus, we may rewrite (4.10) as

$$\boldsymbol{Y}_{1t} = \boldsymbol{\Xi}_{1t} + \boldsymbol{E}_{1t} = \boldsymbol{A}_1\boldsymbol{X}_t\boldsymbol{B}' + \boldsymbol{E}_{1t}, \qquad \boldsymbol{Y}_{2t} = \boldsymbol{\Xi}_{2t} + \boldsymbol{E}_{2t} = \boldsymbol{A}_2\boldsymbol{X}_t^*\boldsymbol{B}' + \boldsymbol{E}_{2t}, \tag{4.11}$$

where $\boldsymbol{X}_t^* = \boldsymbol{Q}\boldsymbol{X}_t^*$ and $\boldsymbol{Q}$ is an invertible $d \times d$ matrix. Under this assumption, $\mathcal{M}(\boldsymbol{A}_1)$ and $\mathcal{M}(\boldsymbol{A}_2)$, which are the column spaces of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, are uniquely defined.

Let $Y_{lt,\cdot j}$ be the j-th column of $\boldsymbol{Y}_{lt}$, $E_{lt,\cdot j}$ be the j-th column of $\boldsymbol{E}_{lt}$ and $B_{j\cdot}$ be the j-th row of $\boldsymbol{B}$, $l = 1, 2$ and $j = 1, \ldots, p$. Define spatial-cross-covariance matrix between the i-th and j-th variables as

$$\begin{aligned}
\boldsymbol{\Omega}_{A,ij} &= \mathrm{Cov}\{Y_{1t,\cdot i}, Y_{2t,\cdot j}\} \\
&= \mathrm{Cov}\{\boldsymbol{A}_1\boldsymbol{X}_t B_{i\cdot} + E_{1t,i}, \boldsymbol{A}_2\boldsymbol{X}_t^* B_{j\cdot} + E_{2t,j}\} \\
&= \boldsymbol{A}_1 \mathrm{Cov}\{\boldsymbol{X}_t B_{i\cdot}, \boldsymbol{X}_t^* B_{j\cdot}\}\boldsymbol{A}_2
\end{aligned} \tag{4.12}$$

When $n \ll d$, it is reasonable to assume that

$$\text{rank}\left(\boldsymbol{\Omega}_{A,ij}\right) = \text{rank}\left(\text{Cov}\{\boldsymbol{X}_t B_{i\cdot}, \boldsymbol{X}_t^* B_{j\cdot}\}\right) = d.$$

Define

$$\begin{aligned}
\boldsymbol{M}_{A_1} &= \sum_{i=1}^{p}\sum_{j=1}^{p} \boldsymbol{\Omega}_{A,ij}\boldsymbol{\Omega}'_{A,ij} \\
&= \boldsymbol{A}_1 \left\{ \sum_{i=1}^{p}\sum_{j=1}^{p} \text{Cov}\{\boldsymbol{X}_t B_{i\cdot}, \boldsymbol{X}_t^* B_{j\cdot}\}\text{Cov}\{\boldsymbol{X}_t^* B_{j\cdot}, \boldsymbol{X}_t B_{i\cdot}\} \right\} \boldsymbol{A}_1', \quad (4.13) \\
\boldsymbol{M}_{A_2} &= \sum_{i=1}^{p}\sum_{j=1}^{p} \boldsymbol{\Omega}'_{A,ij}\boldsymbol{\Omega}_{A,ij} \\
&= \boldsymbol{A}_2 \left\{ \sum_{i=1}^{p}\sum_{j=1}^{p} \text{Cov}\{\boldsymbol{X}_t^* B_{j\cdot}, \boldsymbol{X}_t B_{i\cdot}\}\text{Cov}\{\boldsymbol{X}_t B_{i\cdot}, \boldsymbol{X}_t^* B_{j\cdot}\} \right\} \boldsymbol{A}_2', \quad (4.14)
\end{aligned}$$

$\boldsymbol{M}_{A_1}$ and $\boldsymbol{M}_{A_2}$ share the same d positive eigenvalues and $\boldsymbol{M}_{A_l}\boldsymbol{q} = \boldsymbol{0}$ for any vector $\boldsymbol{q}$ perpendicular to $\mathcal{M}(\boldsymbol{A}_l)$, $l = 1, 2$. Therefore, the columns of $\mathcal{M}(\boldsymbol{A}_l)$, $l = 1, 2$, can be estimated as the d orthonormal eigenvectors of matrix $\boldsymbol{M}_{A_l}$ correspond to d positive eigenvalues and the columns are arranged such that the corresponding eigenvalues are in the descending order.

Now we define the sample version of these quantities and introduce the estimation procedure. Suppose we have centered our observations $\boldsymbol{Y}_{1t}$ and $\boldsymbol{Y}_{2t}$, let $\widehat{\boldsymbol{\Omega}}_{A,ij}$ be the sample cross-space covariance of i-th and j-th variables and $\widehat{\boldsymbol{M}}_{A_l}$ be the sample version of $\boldsymbol{M}_{A_l}$, $l = 1, 2$, that is

$$\widehat{\boldsymbol{\Omega}}_{A,ij} = \frac{1}{T}\sum_{t=1}^{T} Y_{1t,\cdot i} Y'_{2t,\cdot j}, \quad \widehat{\boldsymbol{M}}_{A_1} = \sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\boldsymbol{\Omega}}_{A,ij}\widehat{\boldsymbol{\Omega}}'_{A,ij}, \quad \widehat{\boldsymbol{M}}_{A_2} = \sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\boldsymbol{\Omega}}'_{A,ij}\widehat{\boldsymbol{\Omega}}_{A,ij}. \quad (4.15)$$

A natural estimator for $\boldsymbol{A}_l$ is defined as $\widehat{\boldsymbol{A}}_l = \{\widehat{\boldsymbol{a}}_{l1}, \cdots, \widehat{\boldsymbol{a}}_{ld}\}$, $l = 1, 2$, where $\widehat{\boldsymbol{a}}_{lj}$ is the eigenvector of $\widehat{\boldsymbol{M}}_{A_l}$ corresponding to its j-th largest eigenvalue. However such an estimator ignores the fact that $\boldsymbol{\xi}_t(\boldsymbol{s})$ is continuous over the set $\mathcal{S}$.

4.2.2 Estimation of the Variable Loading Matrix $\boldsymbol{B}$

To estimate the $p \times r$ variable loading matrix $\boldsymbol{B}$, we follow closely the method proposed by Wang et al. (2017) and work with discrete observations of (4.8) at n sampling sites.

Let the vector observed at site $\boldsymbol{s}_i$ at time t be $\boldsymbol{y}_t(\boldsymbol{s}_i)$. The temporal-cross-covariance between observations from site $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ for lag $h \geq 1$ is

$$\boldsymbol{\Omega}_{B,ij}(h) = \mathrm{Cov}\{\boldsymbol{y}_t(\boldsymbol{s}_i),\ \boldsymbol{y}_{t+h}(\boldsymbol{s}_j)\} = \boldsymbol{B}\,\mathrm{Cov}\{\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}_i),\ \boldsymbol{a}'(\boldsymbol{s}_j)\boldsymbol{X}_t\}\,\boldsymbol{B}'. \tag{4.16}$$

The last equation results form the assumption that $\boldsymbol{X}_t$ is uncorrelated with $\boldsymbol{E}_t$ at all leads and lags and $\boldsymbol{E}_t$ is white noise. For a pre-determined maximum lag h_0, define

$$\boldsymbol{M}_B = \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\Omega}_{B,ij}(h)\boldsymbol{\Omega}_{B,ij}'(h). \tag{4.17}$$

By (4.16) and (4.17), it follows that

$$\boldsymbol{M}_B = \boldsymbol{B}\left(\sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\{\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}_i),\ \boldsymbol{a}'(\boldsymbol{s}_j)\boldsymbol{X}_t\}\mathrm{Cov}\{\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}_j),\ \boldsymbol{a}'(\boldsymbol{s}_i)\boldsymbol{X}_t\}\right)\boldsymbol{B}'. \tag{4.18}$$

$\boldsymbol{M}_B$ shares the same r positive eigenvalues and $\boldsymbol{M}_B\boldsymbol{q} = \boldsymbol{0}$ for any vector $\boldsymbol{q}$ perpendicular to $\mathcal{M}(\boldsymbol{B})$. Therefore, the columns of $\mathcal{M}(\boldsymbol{B})$ can be estimated as the r orthonormal eigenvectors of matrix $\boldsymbol{M}_B$ correspond to r positive eigenvalues and the columns are arranged such that the corresponding eigenvalues are in the descending order.

Define the sample version of $\boldsymbol{\Omega}_{B,ij}(h)$ and $\boldsymbol{M}_B$ for centered observation $\boldsymbol{Y}_t$ as

$$\widehat{\boldsymbol{\Omega}}_{B,ij} = \frac{1}{T-h}\sum_{t=1}^{T-h}Y_{1t,\cdot i}Y_{2\,t+h,\cdot j}', \quad \widehat{\boldsymbol{M}}_B = \sum_{h=1}^{h_0}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{\boldsymbol{\Omega}}_{B,ij}\widehat{\boldsymbol{\Omega}}_{B,ij}'. \tag{4.19}$$

A natural estimator for $\boldsymbol{B}$ can be obtained as $\widehat{\boldsymbol{B}} = \{\widehat{\boldsymbol{b}}_1, \cdots, \widehat{\boldsymbol{b}}_r\}$, where $\widehat{\boldsymbol{b}}_i$ is the eigenvector of $\widehat{\boldsymbol{M}}_B$ corresponding to its i-th largest eigenvalue.

4.2.3 Estimation of the Latent Factor Matrix $\boldsymbol{X}_t$ and Signal Matrix $\boldsymbol{\Xi}_t$

By (4.10), the estimators of two representations of the latent matrix factor $\boldsymbol{X}_t$ are defined as

$$\widehat{\boldsymbol{X}}_t = \widehat{\boldsymbol{A}}_1'\boldsymbol{Y}_{1t}\widehat{\boldsymbol{B}}, \qquad \widehat{\boldsymbol{X}}_t^* = \widehat{\boldsymbol{A}}_2'\boldsymbol{Y}_{2t}\widehat{\boldsymbol{B}}. \tag{4.20}$$

The latent signal process are estimated by

$$\widehat{\boldsymbol{\Xi}}_t = \begin{bmatrix}\widehat{\boldsymbol{\Xi}}_{1t}\\ \widehat{\boldsymbol{\Xi}}_{2t}\end{bmatrix}, \tag{4.21}$$

where

$$\widehat{\boldsymbol{\Xi}}_{1t} = \widehat{\boldsymbol{A}}_1 \widehat{\boldsymbol{X}}_t \widehat{\boldsymbol{B}}' = \widehat{\boldsymbol{A}}_1 \widehat{\boldsymbol{A}}_1' \boldsymbol{Y}_{1t} \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}', \qquad \widehat{\boldsymbol{\Xi}}_{2t} = \widehat{\boldsymbol{A}}_2 \widehat{\boldsymbol{X}}_t^* \widehat{\boldsymbol{B}}' = \widehat{\boldsymbol{A}}_2 \widehat{\boldsymbol{A}}_2' \boldsymbol{Y}_{2t} \widehat{\boldsymbol{B}} \widehat{\boldsymbol{B}}'. \tag{4.22}$$

4.2.4 Estimation of the Spatial Loading Matrix $\boldsymbol{A}$ and Loading Function $\boldsymbol{A}(\boldsymbol{s})$

Note that now we only have estimated spatial loading matrices $\widehat{\boldsymbol{A}}_1$ and $\widehat{\boldsymbol{A}}_2$ on two partitioned set of sampling locations under the constraint that $\boldsymbol{A}_1'\boldsymbol{A}_1 = \boldsymbol{A}_2'\boldsymbol{A}_2 = \boldsymbol{I}_d$. Estimate loading functions from $\widehat{\boldsymbol{A}}_1$ and $\widehat{\boldsymbol{A}}_2$ separately will result in inefficient use of sampling locations. Also, the constraint that $\boldsymbol{A}_1'\boldsymbol{A}_1 = \boldsymbol{A}_2'\boldsymbol{A}_2 = \boldsymbol{I}_d$ complicates the estimation of the loading functions $\boldsymbol{a}_j(\boldsymbol{s})$. In addition, (4.20) gives estimators for two different representations of the latent matrix factor $\boldsymbol{X}_t$. To get estimators of spatial loading matrix $\boldsymbol{A}$ for all sampling locations and $\boldsymbol{X}_t$, we use the estimated $\widehat{\boldsymbol{\Xi}}_t$ to re-estimate $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{X}}_t$.

The population signals process is $\boldsymbol{\xi}_t(\boldsymbol{s}) = \boldsymbol{B}\sum_{j=1}^{d} a_j(\boldsymbol{s})\boldsymbol{x}_{tj} = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s})$. The $n \times p$ matrix $\boldsymbol{\Xi}_t = \boldsymbol{A}\boldsymbol{X}_t\boldsymbol{B}$ is the signal matrix at discretized sampling locations at each time t. To reduce dimension, we consider the $n \times r$ variable-factor matrix $\boldsymbol{\Psi}_t = \boldsymbol{\Xi}_t\boldsymbol{B}' = \boldsymbol{A}\boldsymbol{X}_t$. Let $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_T \end{pmatrix}$ and $\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_1 & \cdots & \boldsymbol{\Psi}_T \end{pmatrix} = \boldsymbol{A}\boldsymbol{X}$, then

$$\frac{1}{nprT}\boldsymbol{\Psi}'\boldsymbol{\Psi} = \frac{1}{nprT}\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{X}.$$

Let the rows of $\frac{1}{\sqrt{rT}}\boldsymbol{W}$ be the eigenvectors of $\frac{1}{nprT}\boldsymbol{\Psi}'\boldsymbol{\Psi}$ corresponding to its d non-zero egienvalues. The column space of $\boldsymbol{X}'$ can be estimated as that of $\boldsymbol{W}'$. And $\boldsymbol{A}^* = \frac{1}{rT}\boldsymbol{\Psi}\boldsymbol{W}'$ is the loading function values at discretized sampling site corresponding to $\boldsymbol{W}$.

However, true $\boldsymbol{\Xi}_t$'s or $\boldsymbol{\Psi}_t$'s are not observable and only the estimated values $\widehat{\boldsymbol{\Xi}}_t$ and $\widehat{\Psi} = \widehat{\boldsymbol{\Xi}}_t\widehat{\boldsymbol{B}}$ are available. Thus, we estimate $\frac{1}{\sqrt{rT}}\widehat{\boldsymbol{W}}$ whose columns are the eigenvectors of $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}}$ corresponding to its d non-zero egienvalues and $\widehat{\boldsymbol{A}} = \frac{1}{rT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}'$. The reason that $\widehat{\boldsymbol{\Psi}}$ is choosen over $\widehat{\boldsymbol{\Xi}}$ is that $\widehat{\boldsymbol{\Psi}}$ has the same esimaton error bound but is of lower dimension.

Once $\widehat{\boldsymbol{A}}$ is estimated, we estimate loading functions $a_j(\boldsymbol{s})$ from the estimated n

observations in column $\widehat{A}_{\cdot j}$ by the sieve approximation. Any set of bivariate basis functions can be chosen. In our procedure, we consider the tensor product linear sieve space Θ_n, which is constructed as a tensor product space of some commonly used univariate linear approximating spaces, such as B-spline, orthogonal wavelets and polynomial series. Then for each $j \leq d$,

$$a_j(\boldsymbol{s}) = \sum_{i=1}^{J_n} \beta_{i,j} u_i(\boldsymbol{s}) + r_j(\boldsymbol{s}).$$

Here $\beta_{i,j}$'s are the sieve coefficients of i basis function $u_i(\boldsymbol{s})$ corresponding to the j-th factor loading function; $r_j(\boldsymbol{s})$ is the sieve approximation error; J_n represents the number of sieve terms which grows slowly as n goes to infinity. We estimate $\widehat{\beta}_{i,j}$'s and the loading functions are approximated by $\widehat{a}_j(\boldsymbol{s}) = \sum_{i=1}^{J_n} \widehat{\beta}_{i,j} u_i(\boldsymbol{s})$.

4.3 Prediction

4.3.1 Spatial Prediction

A major focus of spatio-temporal data analysis is the prediction of variable of interest over new locations. For some new location $\boldsymbol{s}_0 \in \mathcal{S}$ and $\boldsymbol{s}_0 \neq \boldsymbol{s}_i$ for $i = 1, \ldots, n$, we aim to predict the unobserved value $\boldsymbol{y}_t(\boldsymbol{s}_0)$, $t = 1, \ldots, T$, based on observations $\boldsymbol{Y}_t$. By (4.8), we have $\boldsymbol{y}_t(\boldsymbol{s}_0) = \boldsymbol{\xi}_t(\boldsymbol{s}_0) + \boldsymbol{\epsilon}_t(\boldsymbol{s}_0) = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}_0) + \boldsymbol{\epsilon}_t(\boldsymbol{s}_0)$. As recommended by Cressie and Wikle (2015), we predict $\boldsymbol{\xi}_t(\boldsymbol{s}_0) = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}_0)$ instead of $\boldsymbol{y}_t(\boldsymbol{s}_0)$ directly. Thus, a natural estimator is

$$\widehat{\boldsymbol{\xi}}_t(\boldsymbol{s}_0) = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{X}}_t'\widehat{\boldsymbol{a}}(\boldsymbol{s}_0), \tag{4.23}$$

where $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{X}}$ and $\widehat{\boldsymbol{a}}(\boldsymbol{s})$ are estimated following procedures in Section 4.2.

4.3.2 Temporal Prediction

Temporal prediction focuses on predict the future values $\boldsymbol{y}_{t+h}(\boldsymbol{s}_1), \ldots, \boldsymbol{y}_{t+h}(\boldsymbol{s}_n)$ for some $h \geq 1$. By (4.8), we have $\boldsymbol{y}_{t+h}(\boldsymbol{s}) = \boldsymbol{\xi}_{t+h}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{t+h}(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_{t+h}'\boldsymbol{a}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{t+h}(\boldsymbol{s})$. Since $\boldsymbol{\epsilon}_{t+h}(\boldsymbol{s})$ is unpredictable white noise, the ideal predictor for $\boldsymbol{y}_{t+h}(\boldsymbol{s})$ is that for $\boldsymbol{\xi}_{t+h}(\boldsymbol{s})$. Thus, we focus on predict $\boldsymbol{\xi}_{t+h}(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_{t+h}'\boldsymbol{a}(\boldsymbol{s})$. The temporal dynamics of the $\boldsymbol{\xi}_{t+h}(\boldsymbol{s})$ present in a lower dimensional matrix factor $\boldsymbol{X}_{t+h}'$, thus a more effective

approach is to predict $\boldsymbol{X}'_{t+h}$ based on $\boldsymbol{X}'_{t-l}, \ldots, \boldsymbol{X}'_t$ where l is a prescribed integer. The rows and columns of $\boldsymbol{X}_t$ represents the spatial factors and the variable factor, respectively. To preserve the matrix structure intrinsic to $\boldsymbol{X}_t$, we model $\{\boldsymbol{X}_t\}_{1:T}$ as the matrix autoregressive model of order one. Mathematically,

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_R \, \boldsymbol{X}_{t-1} \, \boldsymbol{\Phi}_C + \boldsymbol{U}_t, \tag{4.24}$$

where Φ_R and Φ_C are row and column coefficient matrices, respectively. The covariance structure of the matrix white noise $\boldsymbol{U}_t$ is not restricted. Thus, $vec\boldsymbol{U}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_U)$ where $\boldsymbol{\Sigma}_U$ is an arbitrary covariance matrix. Matrix Φ_R captures the auto-correlations between the spatial latent factors and Φ_C captures the auto-correlations between the variable latent factors.

Following the generalized iterative method proposed in Yang et al. (2017), we have estimators $\widehat{\boldsymbol{\Phi}}_R$ and $\widehat{\boldsymbol{\Phi}}_C$. The prediction for $\boldsymbol{y}_{t+h}(\boldsymbol{s})$ is best approximate by

$$\widehat{\boldsymbol{\xi}}_{t+h}(\boldsymbol{s}) = \widehat{\boldsymbol{B}} \, \widehat{\boldsymbol{X}}'_{t+h} \, \widehat{\boldsymbol{a}}(\boldsymbol{s}) = \widehat{\boldsymbol{B}} \, \widehat{\boldsymbol{\Phi}}_R^h \, \widehat{\boldsymbol{X}}_t \, \widehat{\boldsymbol{\Phi}}_C^h \, \widehat{\boldsymbol{a}}(\boldsymbol{s}), \tag{4.25}$$

where $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{X}}$ and $\widehat{\boldsymbol{a}}(\boldsymbol{s})$ are estimated following procedures in Section 4.2 and $\widehat{\boldsymbol{\Phi}}_R^h$ and $\widehat{\boldsymbol{a}}(\boldsymbol{s})$ is estimated from MAR(1) model.

4.4 Asymptotic properties

In this section, we investigate the rates of convergence for the estimators under the setting that n, p and T all go to infinity while d and r are fixed and the factor structure does not change over time. In what follows, let $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}'\boldsymbol{A})}$ and $\|\boldsymbol{A}\|_F = \sqrt{tr(\boldsymbol{A}'\boldsymbol{A})}$ denote the spectral and Frobenius norms of the matrix $\boldsymbol{A}$, respectively. $\|\boldsymbol{A}\|_{min}$ denotes the positive square root of the minimal eigenvalue of $\boldsymbol{A}'\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}'$, whichever is a smaller matrix. When $\boldsymbol{A}$ is a square matrix, we denote by $tr(\boldsymbol{A})$, $\lambda_{max}(\boldsymbol{A})$ and $\lambda_{max}(\boldsymbol{A})$ the trace, maximum and minimum eigenvalues of the matrix $\boldsymbol{A}$, respectively. For two sequences a_N and b_N, we write $a_N \asymp b_N$ if $a_N = O(b_N)$ and $b_N = O(a_N)$. The following regularity conditions are imposed before we derive the asymptotics of the estimators.

Condition 1. Alpha-mixing. $\{\mathrm{vec}\,(\boldsymbol{X}_t)\,, t = 0, \pm 1, \pm 2, \cdots\}$ is strictly stationary and α-mixing. Specifically, for some $\gamma > 2$, the mixing coefficients satisfy the condition that $\sum_{h=1}^{\infty} \alpha(h)^{1-2/\gamma} < \infty$, where $\alpha(h) = \sup_{\tau} \sup_{A \in \mathcal{F}_{-\infty}^{\tau}, B \in \mathcal{F}_{\tau+h}^{\infty}} |P(A \cap B) - P(A)P(B)|$ and $\mathcal{F}_{\tau}^{s}$ is the σ-field generated by $\{vec(\boldsymbol{X}_t) : \tau \leq t \leq s\}$.

Condition 2. Let $X_{t,ij}$ be the ij-th entry of $\boldsymbol{X}_t$. Then, $E(|X_{t,ij}|^{2\gamma}) \leq C$ for any $i = 1, \ldots, d$, $j = 1, \ldots, r$ and $t = 1, \ldots, T$, where C is a positive constant and γ is given in Condition 1. In addition, there exists an integer h satisfying $1 \leq h \leq h_0$ such that $\boldsymbol{\Sigma}_f(h)$ is of rank $k = \max(d, r)$ and $\|\boldsymbol{\Sigma}_f(h)\|_2 \asymp O(1) \asymp \sigma_k(\boldsymbol{\Sigma}_f(h))$. For $i = 1, \ldots, d$ and $j = 1, \ldots, r$, $\frac{1}{T-h} \sum_{t=1}^{T-h} Cov(X_{t,i\cdot}, X_{t+h,i\cdot}) \neq \mathbf{0}$ and $\frac{1}{T-h} \sum_{t=1}^{T-h} Cov(X_{t,\cdot j}, X_{t+h,\cdot j}) \neq \mathbf{0}$.

Condition 3. Spacial factor strength. For any partition $\{\mathcal{S}_1, \mathcal{S}_2\}$ of locations $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, there exists a constant $\delta \in [0, 1]$ such that $\|\boldsymbol{A}_1\|_{min}^2 \asymp n_1^{1-\delta} \asymp \|\boldsymbol{A}_1\|_2^2$ and $\|\boldsymbol{A}_2\|_{min}^2 \asymp n_2^{1-\delta} \asymp \|\boldsymbol{A}_2\|_2^2$, where n_1 and n_2 are number of locations in sets $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively, and $n_1 + n_2 = n$.

Condition 4. Variable factor strength. There exists a constant $\gamma \in [0, 1]$ such that $\|\boldsymbol{B}\|_{min}^2 \asymp p^{1-\gamma} \asymp \|\boldsymbol{B}\|_2^2$ as p goes to infinity and r is fixed.

Condition 5. Loading functions belongs to Hölder class. For $j = 1, \ldots, d$, the loading functions $\boldsymbol{a}_j(\boldsymbol{s})$, $\boldsymbol{s} \in \mathcal{S} \in \mathbb{R}^2$ belongs to a Hölder class $\mathcal{A}_c^{\kappa}(\mathcal{S})$ (κ-smooth) defined by

$$\mathcal{A}_c^{\kappa}(\mathcal{S}) = \left\{ a \in \mathcal{C}^m(\mathcal{S}) : \sup_{[\eta] \leq m} \sup_{\boldsymbol{s} \in \mathcal{S}} |D^{\eta}\, a(\boldsymbol{s})| \leq c, \text{ and } \sup_{[\eta]=m} \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}} \frac{|D^{\eta}\, a(\boldsymbol{u}) - D^{\eta}\, a(\boldsymbol{v})|}{\|\boldsymbol{u} - \boldsymbol{v}\|_2^{\alpha}} \leq c \right\},$$

for some positive number c. Here, $\mathcal{C}^m(\mathcal{S})$ is the space of all m-times continuously differentiable real-value functions on $\mathcal{S}$. The differential operator D^{η} is defined as $D^{\eta} = \frac{\partial^{[\eta]}}{\partial s_1^{\eta_1} \partial s_2^{\eta_2}}$ and $[\eta] = \eta_1 + \eta_2$ for nonnegative integers η_1 and η_2.

Theorem 5 presents the error bound for estimated loading matrix $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

Theorem 5. *Under Condition 1-4 and $n^{\delta} p^{\gamma} T^{-1/2} = o(1)$, we have*

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = O_p((n_1 n_2^{\delta-1} p^{\gamma} + n_1^{\delta-1} n_2 p^{\gamma} + n_1^{\delta} n_2^{\delta} p^{2\gamma}) T^{-1})^{1/2}. \tag{4.26}$$

If $n_1 \asymp n_2 \asymp n$, we have

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = O_p(n^{\delta} p^{\gamma} T^{-1/2}). \tag{4.27}$$

Theorem 6 presents the error bound for estimated signal $\widehat{\boldsymbol{\Xi}}_{it}$ and $\widehat{\boldsymbol{\Xi}}_{t}$.

Theorem 6. *This proposition considers the error bound of signal estimator as in (4.22) for each partition. Under $n^{\delta}p^{\gamma}T^{-1} = o_p(1)$, if $n_1 \asymp n2 \asymp n$, then*

$$n^{-1/2}p^{-1/2}\|\widehat{\boldsymbol{\Xi}}_{it} - \boldsymbol{\Xi}_{it}\|_2 = O_p(n^{\delta/2}p^{\gamma/2}T^{-1/2} + n^{-1/2}p^{-1/2}), \tag{4.28}$$

for $i = 1, 2$, and

$$n^{-1}p^{-1}\|\widehat{\boldsymbol{\Xi}}_{t} - \boldsymbol{\Xi}_{t}\|_2^2 = O_p(n^{\delta}p^{\gamma}T^{-1} + n^{-1/2+\delta/2}p^{-1/2+\gamma/2}T^{-1/2} + n^{-1}p^{-1}) \tag{4.29}$$

Let $\Delta_{npT} = n^{\delta}p^{\gamma}T^{-1} + n^{-1/2+\delta/2}p^{-1/2+\gamma/2}T^{-1/2} + n^{-1}p^{-1}$. Theorem 7 presents the error bound for re-estimated latent factor $\frac{1}{rT}\boldsymbol{W}_t$ whose columns are assume to be the eigenvectors of $\frac{1}{rT}\boldsymbol{\Psi}'\boldsymbol{\Psi}$. And Proposition 1 presents the error bound for re-estimated whole loading matrix $\boldsymbol{A}$ corresponding to estimated $\boldsymbol{W}$.

Theorem 7.

$$\frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\|_F^2 = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right)$$

Proposition 1 presents the error bond for estimated spatial loading matrix $\widehat{\boldsymbol{A}}$.

Proposition 1.

$$\frac{1}{np}\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_F^2 = O_p\left(\Delta_{npT}\right).$$

Theorem 8 presents the space kriging error bound based on sieve approximated function $\widehat{\boldsymbol{A}}(\boldsymbol{s})$.

Theorem 8.

$$\frac{1}{pT}\|\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0) - \boldsymbol{\xi}(\boldsymbol{s}_0)\|_2^2 = O_p(J_n^{-2\kappa}n^{-\delta}p^{-\gamma} + \Delta_{npT} + 1/T) \tag{4.30}$$

4.5 Simulation

In this section we study the numerical performance of the proposed method on synthetic datasets. We let $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$ be drawn randomly from the uniform distribution on $[-1, 1]^2$ and the observed data $\boldsymbol{y}_t(\boldsymbol{s})$ be generated according to model (4.8),

$$\boldsymbol{y}_t(\boldsymbol{s}) = \boldsymbol{\xi}_t(\boldsymbol{s}) + \boldsymbol{\epsilon}_t(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{X}_t'\boldsymbol{a}(\boldsymbol{s}) + \boldsymbol{\epsilon}_t(\boldsymbol{s}).$$

The dimensions of $\boldsymbol{X}_t$ are chosen to be $d = 3$, $r = 2$, and are fixed in all simulations. The latent factor $\boldsymbol{X}_t$ is generated from the Gaussian matrix time series (4.24)

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_R \, \boldsymbol{X}_{t-1} \, \boldsymbol{\Phi}_C + \boldsymbol{U}_t,$$

where $\boldsymbol{\Phi}_R = diag(0.7, 0.8, 0.9)$, $\boldsymbol{\Phi}_C = diag(0.8, 0.6)$ and the entries of $\boldsymbol{U}_t$ are white noise Gaussian process with mean $\mathbf{0}$ and covariance structure such that

$$\boldsymbol{\Sigma}_U = \mathrm{Cov}\{vec(\boldsymbol{U}_t)\} :$$

- Model I: $\boldsymbol{\Sigma}_U = \boldsymbol{I}_{dr}$. (Used in our simulation.)
- Model II: Kronecker product covariance structure $\boldsymbol{\Sigma}_U = \boldsymbol{\Sigma}_C \otimes \boldsymbol{\Sigma}_R$, where $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\Sigma}_C$ are of sizes $d \times d$ and $r \times r$, respectively. Both $\boldsymbol{\Sigma}_R$ and $\boldsymbol{\Sigma}_C$ have values 1 on the diagonal entries and 0.2 on the off-diagonal entries.
- Model III: Arbitrary covariance matrix $\boldsymbol{\Sigma}_U$.

The entries of $\boldsymbol{B}$ is independently sampled from the uniform distribution $\mathcal{U}(-1, 1) \cdot p^{\gamma/2}$. The nugget process $\boldsymbol{\epsilon}_t(\boldsymbol{s})$ are independent and normal with mean $\mathbf{0}$ and the covariance $(1 + s_1^2 + s_2^2)/2\sqrt{3} \cdot \boldsymbol{I}_p$. The basis functions $a_j(\boldsymbol{s})$'s are designed to be

$$a_1(\boldsymbol{s}) = (s_1 - s_2)/2, \quad a_2(\boldsymbol{s}) = \cos\left(\pi\sqrt{2(s_1^2 + s_2^2)}\right), \quad a_3(\boldsymbol{s}) = 1.5 s_1 s_2.$$

With the above generating model setting, the signal-noise-ratio of p-dimensional variable, which is defined as

$$SNR \equiv \frac{\int_{\boldsymbol{s}\in[-1,1]^2} Trace\left[Cov\left(\boldsymbol{\xi}_t(\boldsymbol{s})\right)\right] d\boldsymbol{s}}{\int_{\boldsymbol{s}\in[-1,1]^2} Trace\left[Cov\left(\boldsymbol{\epsilon}_t(\boldsymbol{s})\right)\right] d\boldsymbol{s}} \approx 2.58.$$

We run 200 simulations for each combination of $n = 50, 100, 200, 400$, $p = 10, 20, 40$, and $T = 60, 120, 240$. With each simulation, we calculate $\widehat{d}$, $\widehat{r}$, $\widehat{\boldsymbol{A}}_1$, $\widehat{\boldsymbol{A}_2}$, $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Xi}}_t$, reestimate $\widehat{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{\Xi}}_t$, then use $\widehat{\boldsymbol{A}}$ to get approximated $\widehat{a}_j(\boldsymbol{s})$ following the estimation procedure described in Section 4.2.

Table 4.1 presents the relative frequencies of estimated rank pairs over 200 simulations. The columns corresponding to the true rank pair $(3, 2)$ is highlighted.

The performance of correctly estimating the loading spaces are measured by the space distance between the estimated and true loading matrices $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$, which is defined as

$$\mathcal{D}(\mathcal{M}(\widehat{\boldsymbol{A}}), \mathcal{M}(\boldsymbol{A})) = \left(1 - \frac{1}{\max(d, \widehat{d})} tr\left(\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{A}}'\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}' \cdot \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\right)\right)^{\frac{1}{2}}.$$

It can be shown that $\mathcal{D}(\mathcal{M}(\widehat{\boldsymbol{A}}), \mathcal{M}(\boldsymbol{A}))$ takes its value in $[0, 1]$, it equals to 0 if and only if $\mathcal{M}(\widehat{\boldsymbol{A}}) = \mathcal{M}(\boldsymbol{A})$, and equals to 1 if and only if $\mathcal{M}(\widehat{\boldsymbol{A}}) \perp \mathcal{M}(\boldsymbol{A})$.

Table 4.1: Relative frequency of estimated rank pair $(\widehat{d}, \widehat{r})$ over 200 simulations. The columns correspond to the true value pair $(3, 2)$ are highlighted. Blank cell represents zero value.

$(\widehat{d}, \widehat{r})$			$\gamma = 0$					$\gamma = 0.5$					
T	p	n	(3,2)	(3,1)	(2,2)	(1,2)	(1,1)	(3,2)	(3,1)	(2,2)	(2,1)	(1,2)	(1,1)
60	10	50	0.74	0.04	0.04	0.18	0.02	0.11	0.01	0.13	0.01	0.61	0.14
120	10	50	0.93	0.07			0.01	0.37	0.05	0.06	0.02	0.42	0.09
240	10	50	0.95	0.06				0.82	0.10	0.01		0.07	0.02
60	20	50	0.86		0.02	0.13		0.02		0.10		0.88	0.01
120	20	50	1.00					0.08		0.04		0.88	
240	20	50	1.00					0.49		0.01		0.50	
60	40	50	0.96		0.01	0.04		0.03		0.09		0.89	
120	40	50	1.00					0.02		0.07		0.91	
240	40	50	1.00					0.32		0.01		0.68	
60	10	100	0.94	0.04	0.02			0.64	0.11	0.20	0.02	0.03	0.01
120	10	100	0.96	0.05				0.93	0.07		0.01		
240	10	100	0.97	0.03				0.94	0.06				
60	20	100	1.00					0.73		0.22		0.06	
120	20	100	1.00					0.97		0.04			
240	20	100	1.00					1.00					
60	40	100	1.00					0.72		0.24		0.05	
120	40	100	1.00					0.96		0.04			
240	40	100	1.00					1.00					
60	10	200	0.98	0.03				0.84	0.11	0.03		0.03	0.01
120	10	200	0.97	0.04				0.94	0.07				
240	10	200	0.97	0.03				0.95	0.05				
60	20	200	1.00					0.94		0.02		0.04	
120	20	200	1.00					1.00					
240	20	200	1.00					1.00					
60	40	200	1.00					0.97		0.01		0.03	
120	40	200	1.00					1.00					
240	40	200	1.00					1.00					
60	10	400	0.98	0.02				0.90	0.09			0.02	0.01
120	10	400	0.97	0.03				0.93	0.08				
240	10	400	0.97	0.03				0.96	0.04				
60	20	400	1.00					1.00				0.01	
120	20	400	1.00					1.00					
240	20	400	1.00					1.00					
60	40	400	1.00					1.00				0.01	
120	40	400	1.00					1.00					
240	40	400	1.00					1.00					

Figure 4.1 presents the box plot of the average space distance

$$\frac{1}{2}\left(\mathcal{D}(\mathcal{M}(\widehat{\boldsymbol{A}}_1), \mathcal{M}(\boldsymbol{A}_1)) + \mathcal{D}(\mathcal{M}(\widehat{\boldsymbol{A}}_2), \mathcal{M}(\boldsymbol{A}_2))\right)$$

and compare it with the box plot of space distance between re-estimated $\widehat{\boldsymbol{A}}$ and the truth $\boldsymbol{A}$.

Figure 4.2 presents the box plot of the space distance between $\widehat{\boldsymbol{B}}$ and the truth $\boldsymbol{B}$.

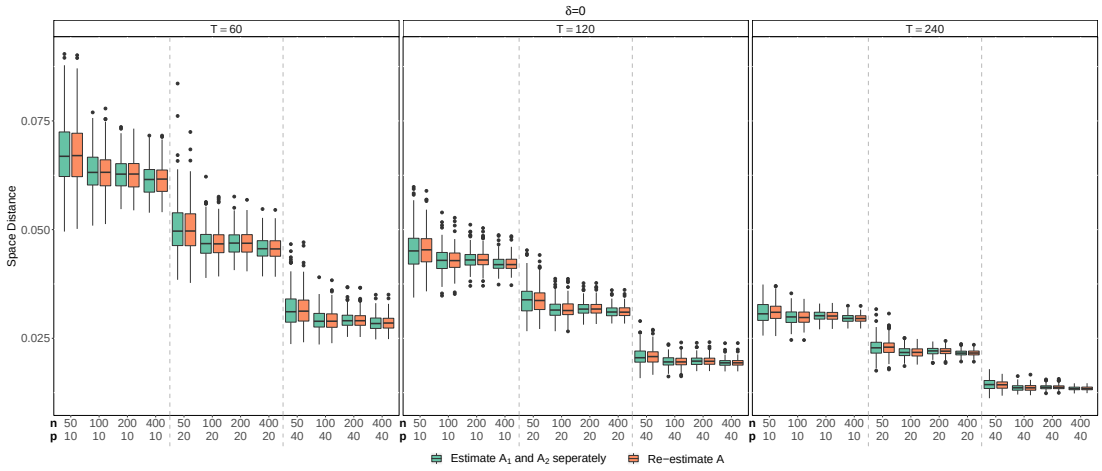


Figure 4.1: Box-plots of the estimation accuracy measured by $\mathcal{D}(\widehat{\boldsymbol{A}}, \boldsymbol{A})$ for the case of orthogonal constraints. Gray boxes represent the average of $\mathcal{D}(\widehat{\boldsymbol{A}}_1, \boldsymbol{A}_1)$ and $\mathcal{D}(\widehat{\boldsymbol{A}}_2, \boldsymbol{A}_2)$. The results are based on 200 iterations. See Table 4.3 in Appendix 4.8 for mean and standard deviations of the spatial distance.

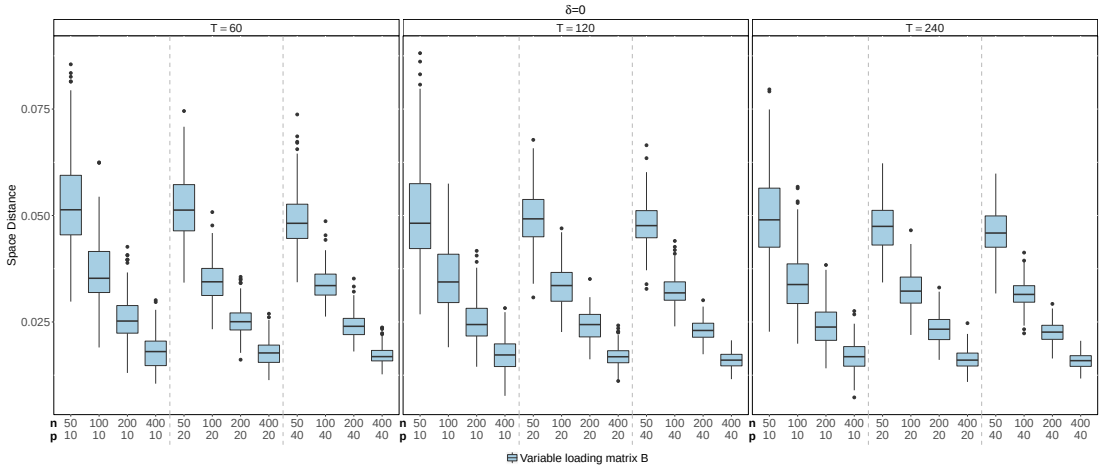


Figure 4.2: Box-plots of the estimation accuracy of variable loading matrix measured by $\mathcal{D}(\widehat{\boldsymbol{B}}, \boldsymbol{B})$. The results are based on 200 iterations. See Table 4.3 in Appendix 4.8 for mean and standard deviations of the spatial distance.

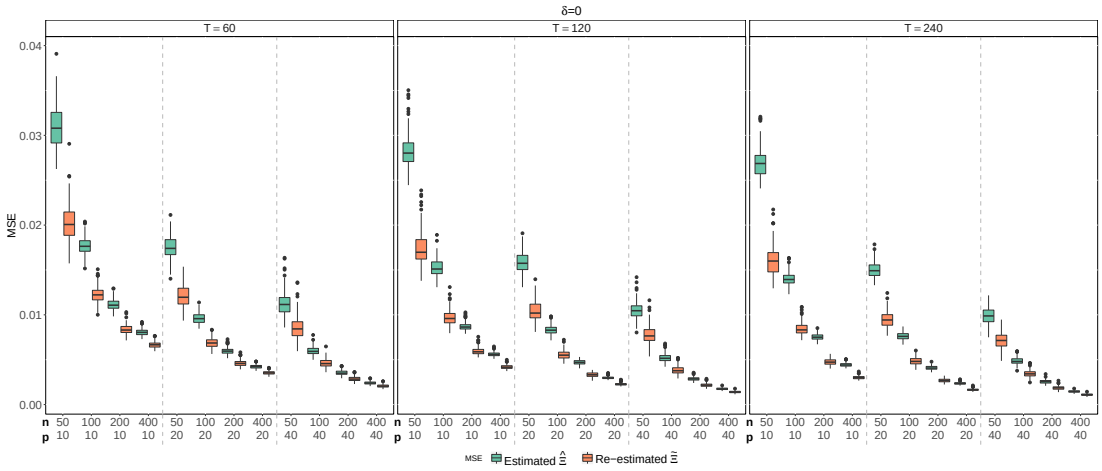


Figure 4.3: Box-plots of the estimation of signals MSE. Gray boxes represent the our procedure. The results are based on 200 iterations. See Table 4.3 in Appendix 4.8 for mean and standard deviations of the MSE.

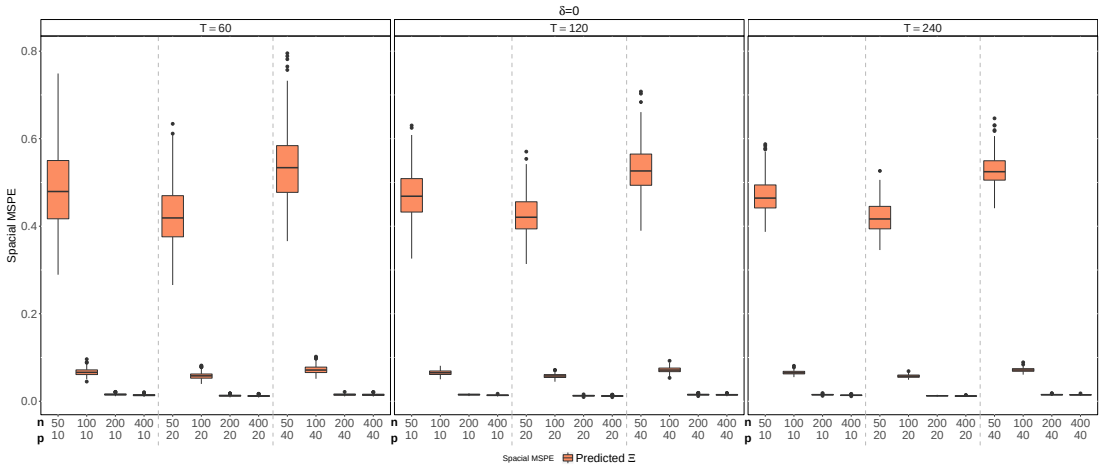


Figure 4.4: Box-plots of the spatial prediction measured by average MSPE for 50 new locations. Colored boxes represent the our model. The results are based on 200 iterations. See Table 4.4 in Appendix 4.8 for mean and standard deviations of the MSPE.

Define the mean squared error of estimated signals $\widehat{\boldsymbol{\xi}}$ as

$$MSE(\widehat{\boldsymbol{\xi}}) = \frac{1}{npT}\sum_{t=1}^{T}\sum_{i=1}^{n}\|\widehat{\boldsymbol{\xi}}_t(\boldsymbol{s}_i) - \boldsymbol{\xi}_t(\boldsymbol{s}_i)\|_2^2.$$

We compare the mean square error between first estimated $\widehat{\boldsymbol{\Xi}}_t$ defined in (4.21) and re-estimated $\widetilde{\boldsymbol{\Xi}}_t$ defined as

$$\widetilde{\boldsymbol{\Xi}} = \left[\widetilde{\boldsymbol{\Xi}}_1, \cdots, \widetilde{\boldsymbol{\Xi}}_T\right] = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{B}}'.$$

The box plots of $MSE(\widehat{\boldsymbol{\xi}})$ and $MSE(\widetilde{\boldsymbol{\xi}})$ are in Figure 4.4. Re-estimated provides much

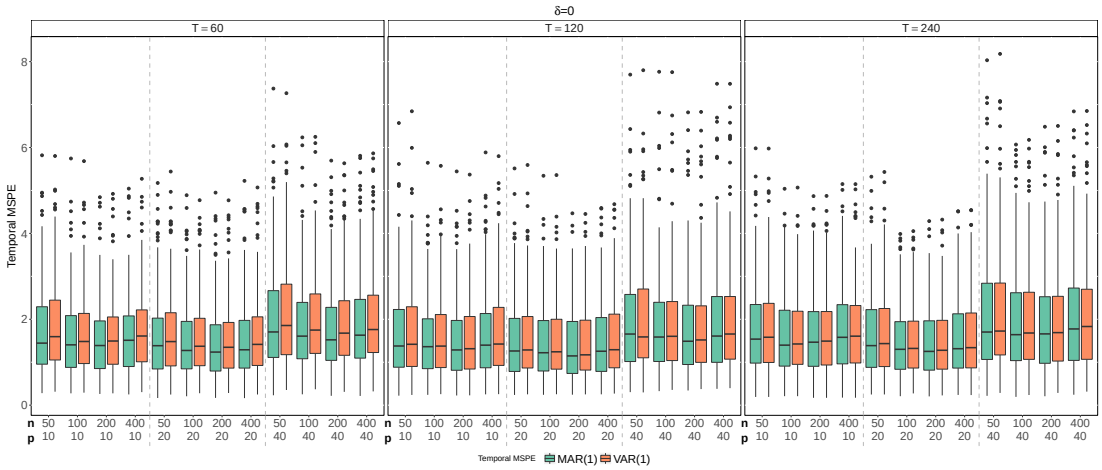


Figure 4.5: Box-plots of the one step ahead forecasting accuracy measured by MSPE. Gray boxes represent the MAR(1) model. The results are based on 200 iterations. See Table 4.4 in Appendix 4.8 for mean and standard deviations of the MSPE.

more accurate estimate for $\boldsymbol{\xi}_t(\boldsymbol{s}_j)$ than $\widetilde{\boldsymbol{\xi}}_t(\boldsymbol{s}_j)$ does.

To demonstrate the performance of spatial prediction, we generate data at a set $\mathcal{S}_0$ of 50 new locations randomly sampled from $\mathcal{U}[-1,1]^2$. For each $t = 1, \ldots, T$, we calculate the spatial prediction $\widehat{\boldsymbol{y}}_t(\cdot) = \widehat{\boldsymbol{\xi}}_t(\cdot)$ defined in (4.23) for each location in $\mathcal{S}_0$. The mean squared spatial prediction error is calculated as

$$MSPE(\widehat{\boldsymbol{y}}) = \frac{1}{50pT} \sum_{t=1}^{T} \sum_{s_0 \in \mathcal{S}_0} \|\widehat{\boldsymbol{y}}_t(\boldsymbol{s}_0) - \boldsymbol{\xi}_t(\boldsymbol{s}_0)\|_2^2.$$

To demonstrate the performance of temporal forecasting, we generate $\boldsymbol{X}_{T+h}$ according to the matrix time series (4.24) for $h = 1, 2$ and compute both the one-step-ahead and two-step-ahead predictions at time T. The mean square temporal prediction error is computed as +

$$MSPE(\widehat{\boldsymbol{y}}_{T+h}) = \frac{1}{np} \sum_{j=1}^{n} \|\widehat{\boldsymbol{y}}_{T+h}(\boldsymbol{s}_j) - \boldsymbol{\xi}_{T+h}(\cdot)\|_2^2.$$

Figure 4.4 presents box-plots of the spatial prediction measured by average MSPE for 50 new locations. The results are based on 200 iterations. Figure 4.5 compares the MSPEs using matrix time series MAR(1) and vectorized time series VAR(1) estimates.

The means and standard errors of the MSPEs from 200 simulations for each model setting are reported in Table 4.4 in Appendix 4.8. It also reports the means and

standard errors of the MSPEs using matrix time series MAR(1) and vectorized time series VAR(1) estimates.

4.6 Real Data Application

In this section, we apply the proposed method to the Comprehensive Climate Dataset (CCDS) – a collection of climate records of North America. The dataset was compiled from five federal agencies sources by Lozano et al. (2009). It contains monthly observations of 17 climate variables spanning from 1990 to 2001 on a 2.5×2.5 degree grid for latitudes in $(30.475, 50.475)$, and longitudes in $(-119.75, -79.75)$. The total number of observation locations is 125 and the length of the whole time series is 156. Table 4.2 lists the variables used in our analysis. Detailed information about data pre-processing is given in Lozano et al. (2009).

Table 4.2: Variables and data sources in the Comprehensive Climate Dataset (CCDS)

Variables (Short name)	Variable group	Type	Source
Methane (CH4)	CH_4	Greenhouse Gases	NOAA
Carbon-Dioxide (CO2)	CO_2		
Hydrogen (H2)	H_2		
Carbon-Monoxide (CO)	CO		
Temperature (TMP)	TMP	Climate	CRU
Temp Min (TMN)	TMP		
Temp Max (TMX)	TMP		
Precipitation (PRE)	PRE		
Vapor (VAP)	VAP		
Cloud Cover (CLD)	CLD		
Wet Days (WET)	WET		
Frost Days (FRS)	FRS		
Global Horizontal (GLO)	SOL	Solar Radiation	NCDC
Direct Normal (DIR)	SOL		
Global Extraterrestrial (ETR)	SOL		
Direct Extraterrestrial (ETRN)	SOL		
Utra Violet (UV)	AER	Aerosol Index	NASA

We first remove the trend and annually seasonal component by taking difference between observations from the same month in consecutive years. Then we normalized this data set by removing the trend and dividing it by the standards deviation for each variable across space. We randomly select 10% of locations and predict the value of all variables over the whole time span for these locations. We repeat the procedure 100 times and the average spatial MSPE is 0.4812.

4.7 Proofs

4.7.1 Factor loadings

Lemma 8. *Let $X_{t,ij}$ denote the ij-th entry of $\boldsymbol{X}_t$. Under Condition 1 and 2, for any $i,k = 1,\ldots,d$ and $j,l = 1,\cdots,r$, we have*

$$\left|\frac{1}{T}\sum_{t=1}^{T}\left(X_{t,ij}X_{t,kl} - Cov(X_{t,ij}X_{t,kl})\right)\right| = O_p(T^{-1/2}). \tag{4.31}$$

Lemma 9. *Under Conditions 1-6, it holds that*

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{s_1s_2,ij} - \boldsymbol{\Omega}_{s_1s_2,ij}\|_2^2 &= O_p((n_1n_2)^{1-\delta}p^{2-2\gamma}T^{-1}),, &(4.32)\\
\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{s_1e_2,ij} - \boldsymbol{\Omega}_{s_1e_2,ij}\|_2^2 &= O_p(n_1^{2-\delta}p^{2-\gamma}T^{-1}),, &(4.33)\\
\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{e_1s_2,ij} - \boldsymbol{\Omega}_{e_1s_2,ij}\|_2^2 &= O_p(n_2^{2-\delta}p^{2-\gamma}T^{-1}),, &(4.34)\\
\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{e_1e_2,ij} - \boldsymbol{\Omega}_{e_1e_2,ij}\|_2^2 &= O_p(n_1n_2p^2T^{-1}). &(4.35)
\end{aligned}$$

Lemma 10. *Under Conditions 1-6, it holds that*

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2^2 == O_p\left(n_1^{2-\delta}p^{2-\gamma}T^{-1} + n_2^{2-\delta}p^{2-\gamma}T^{-1} + n_1n_2p^2T^{-1}\right). \tag{4.36}$$

Proof.

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}}_{ij} &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Y}_{1t,\cdot i}\boldsymbol{Y}'_{2t,\cdot j}\\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{A}_1\boldsymbol{X}_tB_{i\cdot} + E_{t,\cdot i}\right)\left(\boldsymbol{A}_2\boldsymbol{X}_tB_{j\cdot} + E_{t,\cdot j}\right)'\\
&= \widehat{\boldsymbol{\Omega}}_{s,ij} + \widehat{\boldsymbol{\Omega}}_{se,ij} + \widehat{\boldsymbol{\Omega}}_{es,ij} + \widehat{\boldsymbol{\Omega}}_{e,ij}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2^2\\
&\le 4\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\|\widehat{\boldsymbol{\Omega}}_{s_1s_2,ij} - \boldsymbol{\Omega}_{s_1s_2,ij}\|_2^2 + \|\widehat{\boldsymbol{\Omega}}_{s_1e_2,ij} - \boldsymbol{\Omega}_{s_1e_2,ij}\|_2^2 + \|\widehat{\boldsymbol{\Omega}}_{e_1s_2,ij} - \boldsymbol{\Omega}_{e_1s_2,ij}\|_2^2 + \|\widehat{\boldsymbol{\Omega}}_{e_1e_2,ij} - \boldsymbol{\Omega}_{e_1e_2,ij}\|_2^2\right)\\
&= O_p(n_1^{2-\delta}p^{2-\gamma}T^{-1} + n_2^{2-\delta}p^{2-\gamma}T^{-1} + n_1n_2p^2T^{-1})
\end{aligned}$$

□

Lemma 11. *Under Conditions 1-6 and $m_1 p_1^{-1+\delta_1} m_2 p_2^{-1+\delta_2} T^{-1/2} = o_p(1)$, it holds that*

$$\|\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1\|_2 = O_p\left(n^{2-\delta} p^{2-\gamma} T^{-1/2}\right). \tag{4.37}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{\Omega}_{ij}\|_2^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{A}_1\frac{1}{T}\sum_{t=1}^{T}\mathrm{Cov}\{\boldsymbol{X}_t B_{i\cdot}, \boldsymbol{X}_t B_{j\cdot}\}\boldsymbol{A}_2'\|_2^2 \\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t B_{i\cdot} B_{j\cdot}'\boldsymbol{X}_t'\right\}\|_2^2 \\
&\le \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\sum_{i=1}^{p}\sum_{j=1}^{p}\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\mathrm{vec}\left(B_{i\cdot}B_{j\cdot}'\right)\|_2^2 \\
&\le \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\sum_{i=1}^{p}\sum_{j=1}^{p}\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\|_2^2\|\mathrm{vec}\left(B_{i\cdot}B_{j\cdot}'\right)\|_2^2 \\
&= \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\sum_{i=1}^{p}\sum_{j=1}^{p}\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\|_2^2\|B_{i\cdot}B_{j\cdot}'\|_F^2 \\
&\le \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\|_2^2\sum_{i=1}^{p}\sum_{j=1}^{p}\|B_{i\cdot}\|_2^2\|B_{j\cdot}'\|_2^2 \\
&= \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\|_2^2\|B\|_F^4 \\
&\le \|\boldsymbol{A}_1\|_2^2\|\boldsymbol{A}_2\|_2^2\|\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left\{\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right\}\|_2^2\cdot r^2\cdot\|B\|_2^4 \\
&= O_p\left((n_1 n_2)^{1-\delta}p^{2-2\gamma}\right)
\end{aligned}$$

Then,

$$\begin{aligned}
\|\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1\|_2 &= \left\|\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\widehat{\boldsymbol{\Omega}}_{ij}\widehat{\boldsymbol{\Omega}}_{ij}' - \boldsymbol{\Omega}_{ij}\boldsymbol{\Omega}_{ij}'\right)\right\|_2 \\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{\Omega}_{ij}\|_2\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2 \\
&\le \sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2^2 + 2\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\|\boldsymbol{\Omega}_{ij}\|_2^2\cdot\sum_{i=1}^{p}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\Omega}}_{ij} - \boldsymbol{\Omega}_{ij}\|_2^2\right)^{1/2} \\
&= O_p((n_1^{2-\delta}p^{2-\gamma} + n_2^{2-\delta}p^{2-\gamma} + n_1 n_2 p^2)T^{-1}) \\
&\quad + O_p\left(((n_1^{3-2\delta}n_2^{1-\delta}p^{4-3\gamma} + n_1^{1-\delta}n_2^{3-2\delta}p^{4-3\gamma} + n_1^{2-\delta}n_2^{2-\delta}p^{4-2\gamma})T^{-1})^{1/2}\right).
\end{aligned}$$

□

Lemma 12. *Under Condition 2.6 and 2, we have*

$$\lambda_i(\boldsymbol{M}_1) \asymp (n_1 n_2)^{1-\delta} p^{2-2\gamma}, \qquad i = 1, 2, \ldots, k_1,$$

where $\lambda_i(\boldsymbol{M}_1)$ denotes the i-th largest singular value of $\boldsymbol{M}_1$.

Theorem 9. *Under Condition 1-4 and $n^\delta p^\gamma T^{-1/2} = o(1)$, we have*

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = O_p((n_1 n_2^{\delta-1} p^\gamma + n_1^{\delta-1} n_2 p^\gamma + n_1^\delta n_2^\delta p^{2\gamma}) T^{-1})^{1/2}. \tag{4.38}$$

If $n_1 \asymp n_2 \asymp n$, we have

$$\mathcal{D}\left(\mathcal{M}(\widehat{\boldsymbol{A}}_i), \mathcal{M}(\boldsymbol{A}_i)\right) = O_p(n^\delta p^\gamma T^{-1/2}). \tag{4.39}$$

Proof. By Perturbation Theorem,

$$\begin{aligned}
\|\widehat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\|_2 &\leq \frac{8}{\lambda_{min}(\boldsymbol{M}_1)} \|\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1\|_2 \\
&= O_p((n_1 n_2^{\delta-1} p^\gamma + n_1^{\delta-1} n_2 p^\gamma + n_1^\delta n_2^\delta p^{2\gamma}) T^{-1}) \\
&\quad + O_p((n_1 n_2^{\delta-1} p^\gamma + n_1^{\delta-1} n_2 p^\gamma + n_1^\delta n_2^\delta p^{2\gamma}) T^{-1})^{1/2} \\
&= O_p((n_1 n_2^{\delta-1} p^\gamma + n_1^{\delta-1} n_2 p^\gamma + n_1^\delta n_2^\delta p^{2\gamma}) T^{-1})^{1/2}.
\end{aligned}$$

If $n_1 \asymp n_2 \asymp n/2$, we have $\|\widehat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\|_2 = O_p(n^\delta p^\gamma T^{-1/2})$.

If set $n_2 = c$ fixed and $n_1 = n - c$, we have $\|\widehat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\|_2 = O_p((np^{-\gamma} + n^\delta)^{1/2} p^\gamma T^{-1/2})$.

We have the same result for $\|\widehat{\boldsymbol{A}}_2 - \boldsymbol{A}_2\|_2$. □

Theorem 10. *This proposition considers the error bound of signal estimator as in (4.22) for each partition. Under $n^\delta p^\gamma T^{-1} = o_p(1)$, if $n_1 \asymp n2 \asymp n$, then*

$$n^{-1/2} p^{-1/2} \|\widehat{\boldsymbol{\Xi}}_{it} - \boldsymbol{\Xi}_{it}\|_2 = O_p(n^{\delta/2} p^{\gamma/2} T^{-1/2} + n^{-1/2} p^{-1/2}), \tag{4.40}$$

for $i = 1, 2$, and

$$n^{-1} p^{-1} \|\widehat{\boldsymbol{\Xi}}_t - \boldsymbol{\Xi}_t\|_2^2 = O_p(n^\delta p^\gamma T^{-1} + n^{-1/2+\delta/2} p^{-1/2+\gamma/2} T^{-1/2} + n^{-1} p^{-1}) \tag{4.41}$$

Proof.

$$\begin{aligned}
\|\widehat{\boldsymbol{\Xi}}_{1t} - \boldsymbol{\Xi}_{1t}\|_2 &= \left\|\widehat{\boldsymbol{A}}_1^{(j)}\widehat{\boldsymbol{A}}_1^{(j)'}\left(\boldsymbol{A}_1^{(j)}\boldsymbol{X}_t\boldsymbol{B}' + \boldsymbol{E}_t^{(j)}\right)\widehat{\boldsymbol{B}}^{(j)}\widehat{\boldsymbol{B}}^{(j)'} - \boldsymbol{A}_1^{(j)}\boldsymbol{X}_t\boldsymbol{B}'\right\|_2 \\
&\le \left\|\widehat{\boldsymbol{A}}_1^{(j)}\widehat{\boldsymbol{A}}_1^{(j)'}\boldsymbol{A}_1^{(j)}\boldsymbol{X}_t\boldsymbol{B}'\left(\widehat{\boldsymbol{B}}^{(j)}\widehat{\boldsymbol{B}}^{(j)'} - \boldsymbol{B}\boldsymbol{B}'\right)\right\|_2 \\
&\quad + \left\|\left(\widehat{\boldsymbol{A}}_1^{(j)}\widehat{\boldsymbol{A}}_1^{(j)'} - \boldsymbol{A}_1^{(j)}\boldsymbol{A}_1^{(j)'}\right)\boldsymbol{A}_1^{(j)}\boldsymbol{X}_t\boldsymbol{B}'\right\|_2 \\
&\quad + \left\|\widehat{\boldsymbol{A}}_1^{(j)}\widehat{\boldsymbol{A}}_1^{(j)'}\boldsymbol{E}_t^{(j)}\widehat{\boldsymbol{B}}^{(j)}\widehat{\boldsymbol{B}}^{(j)'}\right\|_2 \\
&= \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3.
\end{aligned}$$

$$\begin{aligned}
\|\boldsymbol{I}_1\|_2 &\le 2\|\boldsymbol{X}_t\|_2\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}^{(j)}\|_2 = O_p(n_1^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}\|_2) \\
&= O_p(n_1^{1/2-\delta/2}n^{\delta}p^{1/2+\gamma/2}T^{-1/2}) \\
\|\boldsymbol{I}_2\|_2 &\le 2\|\widehat{\boldsymbol{A}}_1^{(j)} - \boldsymbol{A}_1^{(j)}\|_2\|\boldsymbol{X}_t\|_2 = O_p(n_1^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{A}}_1^{(j)} - \boldsymbol{A}_1^{(j)}\|_2) \\
&= O_p((n_1n_2^{\delta-1}p^{\gamma} + n_1^{\delta-1}n_2p^{\gamma} + n_1^{\delta}n_2^{\delta}p^{2\gamma})T^{-1})^{1/2}n_1^{1/2-\delta/2}p^{1/2-\gamma/2}) \\
&= O_p((n_1^{2-\delta}n_2^{\delta-1} + n_2 + n_1n_2^{\delta}p^{\gamma})pT^{-1})^{1/2} \\
\|\boldsymbol{I}_3\|_2 &\le \|\widehat{\boldsymbol{A}}_1^{(j)'}\boldsymbol{E}_t^{(j)}\widehat{\boldsymbol{B}}^{(j)}\| = \|(\widehat{\boldsymbol{B}}^{(j)'} \otimes \widehat{\boldsymbol{A}}_1^{(j)'})\mathrm{vec}\left(\boldsymbol{E}_t^{(j)}\right)\|_2 \le dr\|\boldsymbol{\Sigma}_e\|_2 = O_p(1).
\end{aligned}$$

Thus,

$$\|\widehat{\boldsymbol{\Xi}}_{1t} - \boldsymbol{\Xi}_{1t}\|_2 = O_p(n_1^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{A}}_1^{(j)} - \boldsymbol{A}_1^{(j)}\|_2) + O_p(n_1^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}\|_2) + O_p(1).$$

$$n_1^{-1/2}p^{-1/2}\|\widehat{\boldsymbol{\Xi}}_{1t} - \boldsymbol{\Xi}_{1t}\|_2 = O_p(n_1^{-\delta/2}p^{-\gamma/2}\|\widehat{\boldsymbol{A}}_1^{(j)} - \boldsymbol{A}_1^{(j)}\|_2) + O_p(n_1^{-\delta/2}p^{-\gamma/2}\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}\|_2) + O_p(n_1^{-1/2}p^{-1/2}).$$

Similarly for $\boldsymbol{\Xi}_{2t}$, we have

$$\|\widehat{\boldsymbol{\Xi}}_{2t} - \boldsymbol{\Xi}_{2t}\|_2 = O_p(n_2^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{A}}_2^{(j)} - \boldsymbol{A}_2^{(j)}\|_2) + O_p(n_2^{1/2-\delta/2}p^{1/2-\gamma/2}\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}\|_2) + O_p(1).$$

$$n_2^{-1/2}p^{-1/2}\|\widehat{\boldsymbol{\Xi}}_{2t} - \boldsymbol{\Xi}_{2t}\|_2 = O_p(n_2^{-\delta/2}p^{-\gamma/2}\|\widehat{\boldsymbol{A}}_1^{(j)} - \boldsymbol{A}_1^{(j)}\|_2) + O_p(n_2^{-\delta/2}p^{-\gamma/2}\|\widehat{\boldsymbol{B}}^{(j)} - \boldsymbol{B}\|_2) + O_p(n_2^{-1/2}p^{-1/2}).$$

If $n_1 \asymp n_2 \asymp n$, then

$$\|\widehat{\boldsymbol{\Xi}}_{it} - \boldsymbol{\Xi}_{it}\|_2 = O_p(n^{1/2+\delta/2}p^{1/2+\gamma/2}T^{-1/2}) + O_p(1), \qquad i = 1, 2. \tag{4.42}$$

Now we find the L_2-norm bounds for

$$\|\widehat{\boldsymbol{\Xi}}_t - \boldsymbol{\Xi}_t\|_2^2 = \left\|\begin{pmatrix}\widehat{\boldsymbol{\Xi}}_{1t} - \boldsymbol{\Xi}_{1t} \\ \widehat{\boldsymbol{\Xi}}_{2t} - \boldsymbol{\Xi}_{2t}\end{pmatrix}\right\|_2^2.$$

Let $\boldsymbol{M} = \widehat{\boldsymbol{\Xi}}_t - \boldsymbol{\Xi}_t = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix}$, the above problem is equivelent to finding $\lambda_{max}(\boldsymbol{M}'\boldsymbol{M})$ from $\lambda_{max}(\boldsymbol{M}_1'\boldsymbol{M}_1)$ and $\lambda_{max}(\boldsymbol{M}_2'\boldsymbol{M}_2)$.

Since

$$\lambda_{max}(\boldsymbol{M}'\boldsymbol{M}) = \lambda_{max}(\boldsymbol{M}_1'\boldsymbol{M}_1 + \boldsymbol{M}_2'\boldsymbol{M}_2) \leq \lambda_{max}(\boldsymbol{M}_1'\boldsymbol{M}_1) + \lambda_{max}(\boldsymbol{M}_2'\boldsymbol{M}_2),$$

We have

$$\begin{aligned} \|\widehat{\boldsymbol{\Xi}}_t - \boldsymbol{\Xi}_t\|_2^2 &\leq \|\widehat{\boldsymbol{\Xi}}_{1t} - \boldsymbol{\Xi}_{1t}\|_2^2 + \|\widehat{\boldsymbol{\Xi}}_{2t} - \boldsymbol{\Xi}_{2t}\|_2^2 \\ &= O_p(n^{1+\delta}p^{1+\gamma}T^{-1}) + O_p(n^{1/2+\delta/2}p^{1/2+\gamma/2}T^{-1/2}) + O_p(1). \end{aligned}$$

$n^{-1}p^{-1}\|\widehat{\boldsymbol{\Xi}}_t - \boldsymbol{\Xi}_t\|_2^2 = O_p(n^{\delta}p^{\gamma}T^{-1} + n^{-1/2+\delta/2}p^{-1/2+\gamma/2}T^{-1/2} + n^{-1}p^{-1})$. □

4.7.2 Space factor loading matrix re-estimation

Lemma 13. *If $n_1 \asymp n2 \asymp n$, then*

$$n^{-1/2}p^{-1/2}\|\widehat{\boldsymbol{\Psi}}_{it} - \boldsymbol{\Psi}_{it}\|_2 = O_p(n^{\delta/2}p^{\gamma/2}T^{-1/2}) + O_p(n^{-1/2}p^{-1/2}), \tag{4.43}$$

for $i = 1, 2$, and

$$n^{-1}p^{-1}\|\widehat{\boldsymbol{\Psi}}_t - \boldsymbol{\Psi}_t\|_2^2 = O_p(n^{\delta}p^{\gamma}T^{-1} + n^{-1/2+\delta/2}p^{-1/2+\gamma/2}T^{-1/2} + n^{-1}p^{-1}) \tag{4.44}$$

Proof. We have

$$\begin{aligned} \|\boldsymbol{\Psi}_{it} - \boldsymbol{\Psi}_{it}\|_2 &= \left\|\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Z}}_t - \boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\right\|_2 = \left\|\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}'(\boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\boldsymbol{Q}_B' + \boldsymbol{E}_t)\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\right\|_2 \\ &= \left\|\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}'\boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\boldsymbol{Q}_B'(\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B) + (\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}' - \boldsymbol{Q}_{A_i}\boldsymbol{Q}_{A_i}')\boldsymbol{Q}_{A_i}\boldsymbol{Z}_t + \widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}'\boldsymbol{E}_t\widehat{\boldsymbol{Q}}_B\right\|_2 \\ &\leq \left\|\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}'\boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\boldsymbol{Q}_B'(\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B)\right\|_2 + \left\|(\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}' - \boldsymbol{Q}_{A_i}\boldsymbol{Q}_{A_i}')\boldsymbol{Q}_{A_i}\boldsymbol{Z}_t\right\|_2 + \left\|\widehat{\boldsymbol{Q}}_{A_i}\widehat{\boldsymbol{Q}}_{A_i}'\boldsymbol{E}_t\widehat{\boldsymbol{Q}}_B\right\|_2 \end{aligned}$$

Then, similar to the proof of Theorem 10, we have the desired results. □

Let $\boldsymbol{U}_t = \widehat{\boldsymbol{\Psi}}_t - \boldsymbol{\Psi}_t$ and $\Delta_{npT} = n^{\delta}p^{\gamma}T^{-1} + n^{-1/2+\delta/2}p^{-1/2+\gamma/2}T^{-1/2} + n^{-1}p^{-1}$. Then Δ_{npT} is the convergence rate of $n^{-1}p^{-1}\|\boldsymbol{U}_t\|_2^2$. Since $\|\boldsymbol{U}_t\|_2^2 \leq \|\boldsymbol{U}_t\|_F^2 \leq r\|\boldsymbol{U}_t\|_2^2$ where r is fixed, we have $n^{-1}p^{-1}\|\boldsymbol{U}_t\|_F^2 = O_p(\Delta_{npT})$.

Define $\boldsymbol{W}_t = \boldsymbol{X}_t\boldsymbol{R}_B'$, $\boldsymbol{W} = (\boldsymbol{W}_1 \cdots \boldsymbol{W}_T)$, $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1 \cdots \boldsymbol{\Psi}_T) = \boldsymbol{A}\boldsymbol{W}$. Assume $\frac{1}{rT}\boldsymbol{W}\boldsymbol{W}' = \boldsymbol{I}_d$. The the columns of $\boldsymbol{W}$ compose of the eigenvectors of $\frac{1}{nprT}\boldsymbol{\Psi}'\boldsymbol{\Psi} =$

$\frac{1}{nprT}\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}$ corresponding to the d nonzero eigenvalues. However, we only have the estimate of $\widehat{\boldsymbol{\Psi}} = (\widehat{\boldsymbol{\Psi}}_1 \cdots \widehat{\boldsymbol{\Psi}}_T)$. Thus, $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{A}}$ can be estimated from $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}} = \frac{1}{nprT}(\boldsymbol{\Psi}+\boldsymbol{U})'(\boldsymbol{\Psi}+\boldsymbol{U})$, where $\boldsymbol{U} = (\boldsymbol{U}_1 \cdots \boldsymbol{U}_T)$ is the approximation error from the previous steps.

Let $\boldsymbol{V}_{npT}$ be the $d \times d$ diagonal matrix of the first d largest eigenvalues of $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}}$ in decreasing order. By definition of eigenvectors and eigenvalues, we have $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}' = \widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}$ or $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1} = \widehat{\boldsymbol{W}}'$.

Define $\boldsymbol{H} = \frac{1}{nprT}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1}$, then

$$\begin{aligned}
\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H} &= \frac{1}{nprT}\widehat{\boldsymbol{\Psi}}'\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1} - \frac{1}{nprT}\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1} \\
&= \left(\frac{1}{nprT}\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{U}\widehat{\boldsymbol{W}}' + \frac{1}{npT}\boldsymbol{U}'\boldsymbol{A}\boldsymbol{W}\widehat{\boldsymbol{W}}' + \frac{1}{nprT}\boldsymbol{U}'\boldsymbol{U}\widehat{\boldsymbol{W}}'\right)\boldsymbol{V}_{npT}^{-1} \\
&= (\boldsymbol{N}_1 + \boldsymbol{N}_2 + \boldsymbol{N}_3)\,\boldsymbol{V}_{npT}^{-1}.
\end{aligned}$$

Lemma 14. $\frac{1}{rT}\|\boldsymbol{N}_1\|_F^2 = \frac{1}{rT}\|\boldsymbol{N}_2\|_F^2 = O_p(n^{-\delta}p^{-\gamma}\Delta_{npT})$ *and* $\frac{1}{rT}\|\boldsymbol{N}_3\|_F^2 = O_p(\Delta_{npT}^2)$.

Proof. Note that $\|\boldsymbol{U}\|_F^2 = \|\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_F^2 = \left\|\sum_{t=1}^T(\widehat{\boldsymbol{\Psi}}_t - \boldsymbol{\Psi}_t)\right\|_F^2 \le T \max_{1\le t\le T}\|\widehat{\boldsymbol{\Psi}}_t - \boldsymbol{\Psi}_t\|_F^2 = O_p(npT\Delta_{npT})$ and $\|\boldsymbol{W}\|_F^2 = \|\widehat{\boldsymbol{W}}\|_F^2 = O_p(rT)$ and r is fixed. In addition, we have $\|\boldsymbol{A}\|_F^2 \asymp \|\boldsymbol{A}\|_2^2 = O_p(n^{1-\delta}p^{1-\gamma})$.

Thus,

$$\begin{aligned}
\frac{1}{rT}\|\boldsymbol{N}_1\|_F^2 &\le \frac{1}{n^2p^2r^3T^3}\|\boldsymbol{W}\|_F^2\|\boldsymbol{A}\|_F^2\|\boldsymbol{U}\|_F^2\|\widehat{\boldsymbol{W}}\|_F^2 = O_p(n^{-\delta}p^{-\gamma}\Delta_{npT}) \\
\frac{1}{rT}\|\boldsymbol{N}_2\|_F^2 &\le \frac{1}{n^2p^2r^3T^3}\|\boldsymbol{U}\|_F^2\|\boldsymbol{A}\|_F^2\|\boldsymbol{W}\|_F^2\|\widehat{\boldsymbol{W}}\|_F^2 = O_p(n^{-\delta}p^{-\gamma}\Delta_{npT}) \\
\frac{1}{rT}\|\boldsymbol{N}_3\|_F^2 &\le \frac{1}{n^2p^2r^3T^3}\|\boldsymbol{U}\|_2^4\|\widehat{\boldsymbol{W}}\|_F^2 = O_p(\Delta_{npT}^2)
\end{aligned}$$

□

Lemma 15. *(i)* $\|\boldsymbol{V}_{npT}\|_2 = O_p(n^{-\delta}p^{-\gamma})$, $\|\boldsymbol{V}_{npT}^{-1}\|_2 = O_p(n^{\delta}p^{\gamma})$.
(ii) $\|\boldsymbol{H}\|_2 = O_p(1)$.

Proof. The d eigenvalues of $\boldsymbol{V}_{npT}$ are the same as those of $\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}' = \frac{1}{np}\boldsymbol{A}\boldsymbol{A}' + \frac{1}{nprT}\boldsymbol{A}\boldsymbol{W}\boldsymbol{U}' + \frac{1}{nprT}\boldsymbol{U}\boldsymbol{W}'\boldsymbol{A}' + \frac{1}{nprT}\boldsymbol{U}\boldsymbol{U}'$, which follows from $\widehat{\boldsymbol{\Psi}} = \boldsymbol{A}\boldsymbol{W} + \boldsymbol{U}$ and $\boldsymbol{W}\boldsymbol{W}'/rT = I_d$. Thus

$$\|\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}' - \frac{1}{np}\boldsymbol{A}\boldsymbol{A}'\|_2 \leq \frac{1}{nprT}\|\boldsymbol{A}\boldsymbol{W}\boldsymbol{U}_t'\|_2 + \frac{1}{nprT}\|\boldsymbol{U}\boldsymbol{W}'\boldsymbol{A}'\|_2 + \frac{1}{nprT}\|\boldsymbol{U}\boldsymbol{U}_t'\| = o_p(1).$$

Using the inequality for the kth eigenvalue, $|\lambda_k(\boldsymbol{W}) - \lambda_k(\boldsymbol{W}_1)| \leq |\boldsymbol{W} - \boldsymbol{W}_1|$, we have $|\lambda_k(\frac{1}{nprT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}') - \lambda(\frac{1}{np}\boldsymbol{A}\boldsymbol{A}')| = o_p(1)$. $\lambda_k(\frac{1}{np}\boldsymbol{A}\boldsymbol{A}') \asymp n^{-\delta}p^{-\gamma}$, $k = 1, \ldots, d$. Thus, $\|\boldsymbol{V}_{npT}\|_{min} \asymp n^{-\delta}p^{-\gamma} \asymp \|\boldsymbol{V}_{npT}\|_2$, $\|\boldsymbol{V}_{npT}^{-1}\|_{min} \asymp n^{\delta}p^{\gamma} \asymp \|\boldsymbol{V}_{npT}^{-1}\|_2$, and $\|\boldsymbol{H}\|_2 = O_p(1)$.

□

Lemma 16.

$$\frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H}\|_F^2 = O_p(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2)$$

Proof. Follow from Lemma 6, 7 and 8. □

Lemma 17.

$$\|\boldsymbol{H} - \boldsymbol{I}_d\|_F = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right) + O_p\left(\Delta_{npT}T^{-1} + n^{\delta}p^{\gamma}\Delta_{npT}^2T^{-1}\right)^{1/2}.$$

Proof. $\boldsymbol{H} = \frac{1}{nprT}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1}$

$$\begin{aligned}
\left\|\boldsymbol{I}_d - \frac{1}{rT}\widehat{\boldsymbol{W}}\boldsymbol{W}'\boldsymbol{H}\right\|_F &= \left\|\frac{1}{rT}\widehat{\boldsymbol{W}}(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H})\right\|_F \\
&\leq \frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H}\|_F^2 + \frac{1}{rT}\|\boldsymbol{W}(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H})\|_F \\
&= O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right) + O_p\left(\Delta_{npT}T^{-1} + n^{\delta}p^{\gamma}\Delta_{npT}^2T^{-1}\right)^{1/2}
\end{aligned}$$

$$\left\|\tfrac{1}{rT}\widehat{\boldsymbol{W}}\boldsymbol{W}'\boldsymbol{H} - \boldsymbol{H}'\boldsymbol{H}\right\|_F = \left\|\tfrac{1}{rT}(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'H)'\boldsymbol{W}'\boldsymbol{H}\right\|_F = O_p\left(\Delta_{npT}T^{-1} + n^{\delta}p^{\gamma}\Delta_{npT}^2T^{-1}\right)^{1/2}$$

Thus,

$$\left\|\boldsymbol{I}_d - \boldsymbol{H}'\boldsymbol{H}\right\|_F = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right) + O_p\left(\Delta_{npT}T^{-1} + n^{\delta}p^{\gamma}\Delta_{npT}^2T^{-1}\right)^{1/2}$$

In addition, by the definition of $\boldsymbol{H} = \frac{1}{nprT}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}\widehat{\boldsymbol{W}}'\boldsymbol{V}_{npT}^{-1}$, we have

$$\|\boldsymbol{H}\boldsymbol{V}_{npT} - \frac{1}{np}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{H}\|_F = \frac{1}{nprT}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{W}(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H}) = O_p\left(n^{-\delta}p^{-\gamma}(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2)^{-1/2}\right).$$

With the same argument of Proposition C.3 in Fan et al. (2016), we have

$$\|\boldsymbol{H} - \boldsymbol{I}_d\|_F = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right) + O_p\left(\Delta_{npT}T^{-1} + n^{\delta}p^{\gamma}\Delta_{npT}^2T^{-1}\right)^{1/2}.$$

□

Theorem 11.

$$\frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\|_F^2 = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right)$$

Proof.

$$\frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\|_F^2 \leq \frac{2}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\boldsymbol{H}\|_F^2 + 2\|\boldsymbol{H} - \boldsymbol{I}_d\|_F^2 = O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right)$$

□

Proposition 2.

$$\frac{1}{np}\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_F^2 = O_p\left(\Delta_{npT}\right).$$

Proof.

$$\widehat{\boldsymbol{A}} = \frac{1}{rT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}'$$

$$\frac{1}{rT}\|\boldsymbol{\Psi}\|_2^2 = \|\boldsymbol{\Psi}\boldsymbol{\Psi}'\|_2 = \left\|\frac{1}{rT}\sum_{t=1}^{T}\boldsymbol{\Psi}_t\boldsymbol{\Psi}_t'\right\|_2 \leq \max_{1\leq t\leq T}\|\boldsymbol{\Psi}_t\boldsymbol{\Psi}_t'\|_2/r = O_p(n^{1-\delta}p^{1-\gamma})$$

$$\begin{aligned}
\frac{1}{np}\left\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\right\|_F^2 &= \frac{1}{np}\left\|\frac{1}{rT}\widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{W}}' - \frac{1}{rT}\boldsymbol{\Psi}\boldsymbol{W}'\right\|_F^2 = \frac{1}{np}\left\|\frac{1}{rT}(\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi})\widehat{\boldsymbol{W}}' + \frac{1}{rT}\boldsymbol{\Psi}(\widehat{\boldsymbol{W}}' - \boldsymbol{W}')\right\|_F^2 \\
&\leq 2\frac{\|\boldsymbol{U}\|_F^2}{nprT}\cdot\frac{1}{rT}\|\widehat{\boldsymbol{W}}'\|_F^2 + 2\frac{\|\boldsymbol{\Psi}\|_F^2}{nprT}\cdot\frac{1}{rT}\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\|_F^2 \\
&= O_p\left(\Delta_{npT} + n^{-\delta}p^{-\gamma}(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2)\right) \\
&= O_p(\Delta_{npT})
\end{aligned}$$

□

4.7.3 Sieve approximation of space loading function

$\boldsymbol{A}(\boldsymbol{s}) = (a_1(\boldsymbol{s}), \cdots, a_d(\boldsymbol{s}))$, now we want to approximate $a_j(\boldsymbol{s})$ with linear combination of basis functions, the approximating functions are $\widehat{g}_j(\boldsymbol{s})$. We estimate $\widehat{g}_j(\boldsymbol{s})$ based on estimated value $\widehat{A}_{\cdot j}$'s. $\widehat{A}_{\cdot j} = A_{\cdot j} + E_{A,\cdot j}$. Since for $n \times d$ matrix $\boldsymbol{A}$ with fixed column dimension d, $\|\boldsymbol{A}\|_2^2 \leq \|\boldsymbol{A}\|_F^2 \leq d\|\boldsymbol{A}\|_2^2$. we have $\|\widehat{A}_{\cdot j} - A_{\cdot j}\|_2^2 = O_p(np\Delta_{npT})$, $j = 1, \ldots, d$.

$A_{\cdot j} = a_j(\boldsymbol{s})$, then $\widehat{A}_{\cdot j} = \widehat{a}_j(\boldsymbol{s}) = a_j(\boldsymbol{s}) + e_{a,j}(\boldsymbol{s})$.

Lemma 18. *If Hölder class, then* $|a_j(\boldsymbol{s})|_\infty^2 \asymp n^{-\delta}p^{1-\gamma}$, $|e_{a,j}(\boldsymbol{s})|_\infty^2 = O_p(p\Delta_{npT})$.

Proof.

$$\lambda_{max}(\boldsymbol{A}\boldsymbol{A}') = \lambda_{max}(\sum_{j=1}^{d} A_{\cdot j}A_{\cdot j}') \geq \lambda_{min}(\sum_{j=1}^{d} A_{\cdot j}A_{\cdot j}') \quad \geq \quad \sum_{j=1}^{d}\lambda_{min}(A_{\cdot j}'A_{\cdot j}) = \sum_{j=1}^{d}\sum_{i=1}^{n} A_{ij}^2$$

$$\lambda_{min}(\boldsymbol{A}\boldsymbol{A}') = \lambda_{min}(\sum_{j=1}^{d} A_{\cdot j}A_{\cdot j}') \leq \lambda_{max}(\sum_{j=1}^{d} A_{\cdot j}A_{\cdot j}') \quad \leq \quad \sum_{j=1}^{d}\lambda_{max}(A_{\cdot j}'A_{\cdot j}) = \sum_{j=1}^{d}\sum_{i=1}^{n} A_{ij}^2$$

Since $\|\boldsymbol{A}\|_{min}^2 \asymp \|\boldsymbol{A}\|_{max}^2 \asymp n^{1-\delta}p^{1-\gamma}$, then $\|A_{\cdot j}\|^2 \asymp n^{1-\delta}p^{1-\gamma}$.

If Hölder class, then $|a_j(\boldsymbol{s})|_\infty^2 \asymp n^{-\delta}p^{1-\gamma}$ by multivariate Taylor expansion and Sandwich Theorem.

□

Lemma 19. $\|\widehat{g}_j(\boldsymbol{s}) - a_j(\boldsymbol{s})\|_\infty = O_p(J_n^{-\kappa}n^{-\delta/2}p^{1/2-\gamma/2}) + O_p(\sqrt{p\Delta_{npT}})$.

Proof. Following Theorem 12.6, 12.7 and 12.8 in Schumaker (2007), we have $\|\widehat{g}_j(\boldsymbol{s}) - a_j(\boldsymbol{s})\|_\infty = \|\boldsymbol{P}\widehat{a}_j(\boldsymbol{s}) - a_j(\boldsymbol{s})\| \leq \|\boldsymbol{P}a(\boldsymbol{s}) - a_j(\boldsymbol{s})\| + \|\boldsymbol{P}e_{a,j}(\boldsymbol{s}) - e_{a,j}(\boldsymbol{s})\| + \|e_{a,j}(\boldsymbol{s})\| = O_p(J_n^{-\kappa}n^{-\delta/2}p^{1/2-\gamma/2}) + O_p(\sqrt{p\Delta_{npT}})$. □

Theorem 12.

$$\frac{1}{pT}\|\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0) - \boldsymbol{\xi}(\boldsymbol{s}_0)\|_2^2 = O_p(J_n^{-2\kappa}n^{-\delta}p^{-\gamma} + \Delta_{npT} + 1/T) \tag{4.45}$$

Proof. Let $\boldsymbol{\xi}_t'(\boldsymbol{s}_0) = \boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{X}_t\boldsymbol{B} = \boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{X}_t\boldsymbol{R}_B'\boldsymbol{Q}_B'$

$\boldsymbol{\xi}'(\boldsymbol{s}_0) = (\boldsymbol{\xi}_1'(\boldsymbol{s}_0)\cdots\boldsymbol{\xi}_T'(\boldsymbol{s}_0)) = (\boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{X}_1\boldsymbol{R}_B'\boldsymbol{Q}_B'\cdots\boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{X}_T\boldsymbol{R}_B'\boldsymbol{Q}_B') = \boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{W}(\boldsymbol{I}_T\otimes\boldsymbol{Q}_B')$

$\widehat{\boldsymbol{\xi}}'(\boldsymbol{s}_0) = \widehat{\boldsymbol{g}}'(\boldsymbol{s}_0)\widehat{\boldsymbol{W}}(\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B')$.

$\widehat{\boldsymbol{\xi}}'(\boldsymbol{s}_0) - \boldsymbol{\xi}'(\boldsymbol{s}_0) = \widehat{\boldsymbol{g}}'(\boldsymbol{s}_0)\widehat{\boldsymbol{W}}'(\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B') - \boldsymbol{a}'(\boldsymbol{s}_0)\boldsymbol{W}'(\boldsymbol{I}_T\otimes\boldsymbol{Q}_B')$.

$$\begin{aligned}
\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0) - \boldsymbol{\xi}(\boldsymbol{s}_0) &= (\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B)\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{g}}(\boldsymbol{s}_0) - (\boldsymbol{I}_T\otimes\boldsymbol{Q}_B)\boldsymbol{Q}_W\boldsymbol{a}(\boldsymbol{s}_0)\\
&= (\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B)\left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{g}}(\boldsymbol{s}_0) - \boldsymbol{Q}_W\boldsymbol{a}(\boldsymbol{s}_0)\right) + \left(\boldsymbol{I}_T\otimes(\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B)\right)\boldsymbol{Q}_W\boldsymbol{a}(\boldsymbol{s}_0)\\
&= (\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B)\widehat{\boldsymbol{W}}'\left(\widehat{\boldsymbol{g}}(\boldsymbol{s}_0) - \boldsymbol{a}(\boldsymbol{s}_0)\right) + (\boldsymbol{I}_T\otimes\widehat{\boldsymbol{Q}}_B)\left(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\right)\boldsymbol{a}(\boldsymbol{s}_0) + \left(\boldsymbol{I}_T\otimes(\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B)\right)\boldsymbol{W}'\boldsymbol{a}(\boldsymbol{s}_0)
\end{aligned}$$

$$
\begin{aligned}
\frac{1}{\sqrt{T}}\|(\boldsymbol{I}_T \otimes \widehat{\boldsymbol{Q}}_B)\widehat{\boldsymbol{W}}'\left(\widehat{\boldsymbol{g}}(\boldsymbol{s}_0) - \boldsymbol{a}(\boldsymbol{s}_0)\right)\|_2 &\leq \frac{1}{\sqrt{T}}\|\boldsymbol{I}_T \otimes \widehat{\boldsymbol{Q}}_B\|_2\|\widehat{\boldsymbol{W}}'\|_2\|\widehat{\boldsymbol{g}}(\boldsymbol{s}_0) - \boldsymbol{a}(\boldsymbol{s}_0)\|_2 \\
&= O_p(J_n^{-\kappa}n^{-\delta/2}p^{1/2-\gamma/2} + \sqrt{p\Delta_{npT}}) \\
\frac{1}{T}\|(\boldsymbol{I}_T \otimes \widehat{\boldsymbol{Q}}_B)\left(\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\right)\boldsymbol{a}(\boldsymbol{s}_0)\|_2^2 &\leq \frac{1}{T}\|\boldsymbol{I}_T \otimes \widehat{\boldsymbol{Q}}_B\|_2^2\|\widehat{\boldsymbol{W}}' - \boldsymbol{W}'\|_2\|\boldsymbol{a}(\boldsymbol{s}_0)\|_2^2 \\
&= O_p\left(\Delta_{npT} + n^{\delta}p^{\gamma}\Delta_{npT}^2\right)O_p(n^{-\delta}p^{1-\gamma}) \\
&= O_p(n^{-\delta}p^{1-\gamma}\Delta_{npT} + p\Delta_{npT}^2) \\
\frac{1}{\sqrt{T}}\|\Big(\boldsymbol{I}_T \otimes (\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B)\Big)\boldsymbol{W}'\boldsymbol{a}(\boldsymbol{s}_0)\|_2 &\leq \frac{1}{\sqrt{T}}\|\boldsymbol{I}_T \otimes (\widehat{\boldsymbol{Q}}_B - \boldsymbol{Q}_B)\|_2\|\boldsymbol{W}'\|_2\|\boldsymbol{a}(\boldsymbol{s}_0)\|_2 \\
&= O_p(n^{\delta/2}p^{\gamma/2}T^{-1/2})O_p(n^{-\delta/2}p^{1/2-\gamma/2}) \\
&= O_p(\sqrt{p/T})
\end{aligned}
$$

Thus,

$$
\frac{1}{pT}\|\widehat{\boldsymbol{\xi}}(\boldsymbol{s}_0) - \boldsymbol{\xi}(\boldsymbol{s}_0)\|_2^2 = O_p(J_n^{-2\kappa}n^{-\delta}p^{-\gamma} + \Delta_{npT} + 1/T). \tag{4.46}
$$

□

4.8 Appendix

Table 4.3: Mean and standard deviations (in parentheses) of the estimated accuracy measured by $\mathcal{D}(\widehat{\cdot}, \cdot)$ for spatial and variable loading matrices. All numbers in the table are 10 times the true numbers for clear representation. The results are based on 200 simulations.

			$\gamma = 0$					$\gamma = 0.5$				
T	p	n	$\mathcal{D}(\widehat{\boldsymbol{A}}_1, \boldsymbol{A}_1)$	$\mathcal{D}(\widehat{\boldsymbol{A}}_2, \boldsymbol{A}_2)$	Average	$\mathcal{D}(\widehat{\boldsymbol{A}}, \boldsymbol{A})$	$\mathcal{D}(\widehat{\boldsymbol{B}}, \boldsymbol{B})$	$\mathcal{D}(\widehat{\boldsymbol{A}}_1, \boldsymbol{A}_1)$	$\mathcal{D}(\widehat{\boldsymbol{A}}_2, \boldsymbol{A}_2)$	Average	$\mathcal{D}(\widehat{\boldsymbol{A}}, \boldsymbol{A})$	$\mathcal{D}(\widehat{\boldsymbol{B}}, \boldsymbol{B})$
60	10	50	0.68(0.1)	0.67(0.1)	0.68(0.08)	0.67(0.07)	0.53(0.11)	1.27(0.19)	1.25(0.21)	1.26(0.16)	1.25(0.15)	0.69(0.14)
120	10	50	0.45(0.06)	0.46(0.06)	0.45(0.05)	0.45(0.04)	0.5(0.12)	0.83(0.12)	0.84(0.12)	0.84(0.09)	0.84(0.08)	0.63(0.13)
240	10	50	0.31(0.04)	0.31(0.04)	0.31(0.03)	0.31(0.02)	0.49(0.11)	0.57(0.07)	0.57(0.08)	0.57(0.05)	0.57(0.04)	0.6(0.13)
60	20	50	0.5(0.07)	0.5(0.09)	0.5(0.06)	0.5(0.06)	0.52(0.08)	1.18(0.21)	1.18(0.24)	1.18(0.17)	1.17(0.15)	0.69(0.1)
120	20	50	0.34(0.05)	0.34(0.05)	0.34(0.03)	0.34(0.03)	0.5(0.07)	0.79(0.12)	0.79(0.12)	0.79(0.09)	0.78(0.08)	0.6(0.08)
240	20	50	0.23(0.03)	0.23(0.03)	0.23(0.02)	0.23(0.02)	0.47(0.06)	0.52(0.07)	0.52(0.07)	0.52(0.05)	0.52(0.05)	0.54(0.06)
60	40	50	0.32(0.06)	0.32(0.05)	0.32(0.04)	0.32(0.04)	0.49(0.07)	0.98(0.21)	0.95(0.19)	0.96(0.15)	0.95(0.13)	0.67(0.07)
120	40	50	0.21(0.03)	0.21(0.03)	0.21(0.02)	0.21(0.02)	0.48(0.05)	0.63(0.1)	0.62(0.1)	0.63(0.08)	0.62(0.07)	0.58(0.06)
240	40	50	0.15(0.02)	0.14(0.02)	0.14(0.01)	0.14(0.01)	0.46(0.05)	0.42(0.06)	0.41(0.06)	0.41(0.04)	0.41(0.03)	0.53(0.06)
60	10	100	0.63(0.06)	0.63(0.07)	0.63(0.05)	0.63(0.05)	0.36(0.07)	1.13(0.12)	1.13(0.13)	1.13(0.1)	1.13(0.09)	0.48(0.09)
120	10	100	0.43(0.04)	0.43(0.04)	0.43(0.03)	0.43(0.03)	0.35(0.07)	0.77(0.08)	0.77(0.07)	0.77(0.05)	0.77(0.05)	0.44(0.08)
240	10	100	0.3(0.03)	0.3(0.03)	0.3(0.02)	0.3(0.02)	0.34(0.07)	0.54(0.05)	0.53(0.05)	0.54(0.03)	0.54(0.03)	0.41(0.08)
60	20	100	0.47(0.05)	0.47(0.05)	0.47(0.04)	0.47(0.04)	0.35(0.05)	1.01(0.11)	1.02(0.11)	1.01(0.08)	1.01(0.08)	0.47(0.06)
120	20	100	0.32(0.03)	0.32(0.03)	0.32(0.02)	0.32(0.02)	0.34(0.05)	0.68(0.07)	0.68(0.07)	0.68(0.05)	0.68(0.05)	0.41(0.05)
240	20	100	0.22(0.02)	0.22(0.02)	0.22(0.01)	0.22(0.01)	0.32(0.05)	0.47(0.04)	0.47(0.04)	0.47(0.03)	0.47(0.03)	0.37(0.05)
60	40	100	0.29(0.03)	0.29(0.03)	0.29(0.02)	0.29(0.02)	0.34(0.04)	0.77(0.1)	0.77(0.1)	0.77(0.07)	0.77(0.07)	0.47(0.04)
120	40	100	0.2(0.02)	0.2(0.02)	0.2(0.01)	0.2(0.01)	0.32(0.04)	0.52(0.05)	0.51(0.05)	0.52(0.04)	0.52(0.04)	0.4(0.04)
240	40	100	0.14(0.01)	0.14(0.01)	0.14(0.01)	0.14(0.01)	0.32(0.03)	0.35(0.03)	0.36(0.03)	0.35(0.02)	0.35(0.02)	0.35(0.04)
60	10	200	0.63(0.05)	0.62(0.05)	0.63(0.04)	0.63(0.04)	0.26(0.06)	1.11(0.08)	1.1(0.08)	1.1(0.07)	1.1(0.07)	0.33(0.07)
120	10	200	0.43(0.03)	0.43(0.03)	0.43(0.02)	0.43(0.02)	0.25(0.05)	0.77(0.05)	0.76(0.05)	0.77(0.04)	0.77(0.04)	0.31(0.06)
240	10	200	0.3(0.02)	0.3(0.02)	0.3(0.01)	0.3(0.01)	0.24(0.05)	0.54(0.03)	0.54(0.03)	0.54(0.02)	0.54(0.02)	0.29(0.06)
60	20	200	0.47(0.04)	0.47(0.04)	0.47(0.03)	0.47(0.03)	0.25(0.03)	0.99(0.07)	0.98(0.07)	0.98(0.06)	0.98(0.06)	0.34(0.05)
120	20	200	0.32(0.02)	0.32(0.02)	0.32(0.02)	0.32(0.02)	0.24(0.04)	0.68(0.05)	0.67(0.04)	0.67(0.04)	0.67(0.03)	0.29(0.04)
240	20	200	0.22(0.01)	0.22(0.01)	0.22(0.01)	0.22(0.01)	0.23(0.03)	0.47(0.03)	0.47(0.03)	0.47(0.02)	0.47(0.02)	0.26(0.04)
60	40	200	0.29(0.03)	0.29(0.02)	0.29(0.02)	0.29(0.02)	0.24(0.03)	0.73(0.06)	0.73(0.05)	0.73(0.05)	0.73(0.05)	0.33(0.04)
120	40	200	0.2(0.01)	0.2(0.01)	0.2(0.01)	0.2(0.01)	0.23(0.02)	0.5(0.03)	0.5(0.03)	0.5(0.03)	0.5(0.03)	0.28(0.03)
240	40	200	0.14(0.01)	0.14(0.01)	0.14(0.01)	0.14(0.01)	0.22(0.02)	0.35(0.02)	0.35(0.02)	0.35(0.01)	0.35(0.01)	0.25(0.03)
60	10	400	0.61(0.04)	0.61(0.04)	0.61(0.04)	0.61(0.04)	0.18(0.04)	1.08(0.07)	1.08(0.07)	1.08(0.06)	1.08(0.06)	0.24(0.05)
120	10	400	0.42(0.02)	0.42(0.02)	0.42(0.02)	0.42(0.02)	0.17(0.04)	0.75(0.04)	0.75(0.04)	0.75(0.03)	0.75(0.03)	0.22(0.05)
240	10	400	0.3(0.01)	0.3(0.01)	0.3(0.01)	0.3(0.01)	0.17(0.04)	0.52(0.02)	0.53(0.02)	0.53(0.02)	0.53(0.02)	0.2(0.04)
60	20	400	0.46(0.03)	0.46(0.03)	0.46(0.03)	0.46(0.03)	0.18(0.03)	0.95(0.05)	0.95(0.06)	0.95(0.05)	0.95(0.05)	0.24(0.04)
120	20	400	0.31(0.02)	0.31(0.02)	0.31(0.01)	0.31(0.01)	0.17(0.02)	0.65(0.04)	0.65(0.03)	0.65(0.03)	0.65(0.03)	0.2(0.03)
240	20	400	0.22(0.01)	0.22(0.01)	0.22(0.01)	0.22(0.01)	0.16(0.02)	0.46(0.02)	0.46(0.02)	0.46(0.01)	0.46(0.01)	0.18(0.03)
60	40	400	0.29(0.02)	0.29(0.02)	0.29(0.02)	0.29(0.02)	0.17(0.02)	0.7(0.04)	0.7(0.05)	0.7(0.04)	0.7(0.04)	0.24(0.02)
120	40	400	0.19(0.01)	0.19(0.01)	0.19(0.01)	0.19(0.01)	0.16(0.02)	0.49(0.02)	0.48(0.02)	0.48(0.02)	0.48(0.02)	0.2(0.02)
240	40	400	0.13(0.01)	0.13(0.01)	0.13(0)	0.13(0)	0.16(0.02)	0.34(0.02)	0.34(0.01)	0.34(0.01)	0.34(0.01)	0.18(0.02)

Table 4.4: Mean and standard deviations (in parentheses) of the mean squared prediction errors (MSPE).

			Spatial	Temporal MAR(1)		Temporal VAR(1)	
T	p	n	$MSPE(\widehat{\boldsymbol{y}}_t(\boldsymbol{s}_0)))$	$MSPE(\widehat{\boldsymbol{y}}_{t+1}(\boldsymbol{s})))$	$MSPE(\widehat{\boldsymbol{y}}_{t+2}(\boldsymbol{s})))$	$MSPE(\widehat{\boldsymbol{y}}_{t+1}(\boldsymbol{s})))$	$MSPE(\widehat{\boldsymbol{y}}_{t+2}(\boldsymbol{s})))$
60	10	50	0.486(0.089)	1.716(1.064)	1.823(1.201)	1.825(1.075)	2.019(1.257)
120	10	50	0.471(0.06)	1.658(1.121)	1.634(1.116)	1.705(1.133)	1.732(1.144)
240	10	50	0.47(0.041)	1.78(1.079)	1.588(1.244)	1.802(1.076)	1.624(1.229)
60	20	50	0.424(0.069)	1.592(1.004)	1.657(1.033)	1.69(1.032)	1.819(1.061)
120	20	50	0.424(0.048)	1.535(0.972)	1.547(1.111)	1.575(0.983)	1.634(1.128)
240	20	50	0.419(0.036)	1.619(0.985)	1.426(1.05)	1.64(0.988)	1.463(1.047)
60	40	50	0.537(0.085)	2.001(1.237)	2.101(1.353)	2.13(1.276)	2.308(1.39)
120	40	50	0.534(0.055)	2.006(1.345)	1.94(1.286)	2.065(1.36)	2.051(1.296)
240	40	50	0.53(0.037)	2.141(1.434)	1.834(1.237)	2.162(1.432)	1.877(1.23)
60	10	100	0.067(0.009)	1.597(0.966)	1.647(1.006)	1.685(0.969)	1.82(1.03)
120	10	100	0.066(0.006)	1.564(0.984)	1.502(0.95)	1.608(0.997)	1.593(0.973)
240	10	100	0.065(0.004)	1.631(0.92)	1.476(1.02)	1.65(0.915)	1.514(1.015)
60	20	100	0.058(0.008)	1.466(0.876)	1.508(0.901)	1.557(0.891)	1.663(0.926)
120	20	100	0.058(0.005)	1.45(0.883)	1.403(0.915)	1.489(0.891)	1.478(0.922)
240	20	100	0.058(0.004)	1.491(0.856)	1.317(0.864)	1.51(0.854)	1.353(0.859)
60	40	100	0.072(0.01)	1.845(1.075)	1.893(1.105)	1.975(1.113)	2.085(1.126)
120	40	100	0.072(0.006)	1.889(1.229)	1.765(1.076)	1.939(1.247)	1.859(1.077)
240	40	100	0.072(0.005)	1.961(1.223)	1.707(1.074)	1.984(1.22)	1.754(1.068)
60	10	200	0.015(0.002)	1.542(0.922)	1.597(0.972)	1.629(0.921)	1.766(1)
120	10	200	0.015(0.001)	1.515(0.976)	1.454(0.913)	1.557(0.982)	1.538(0.934)
240	10	200	0.015(0.001)	1.599(0.915)	1.42(0.988)	1.619(0.912)	1.458(0.988)
60	20	200	0.013(0.002)	1.419(0.86)	1.461(0.88)	1.51(0.88)	1.61(0.897)
120	20	200	0.013(0.001)	1.401(0.853)	1.358(0.88)	1.44(0.861)	1.429(0.883)
240	20	200	0.013(0.001)	1.464(0.859)	1.276(0.84)	1.481(0.86)	1.308(0.838)
60	40	200	0.015(0.002)	1.786(1.04)	1.836(1.099)	1.906(1.066)	2.02(1.122)
120	40	200	0.015(0.001)	1.828(1.211)	1.714(1.042)	1.875(1.22)	1.808(1.049)
240	40	200	0.015(0.001)	1.92(1.214)	1.652(1.031)	1.941(1.213)	1.698(1.027)
60	10	400	0.014(0.002)	1.63(0.965)	1.714(1.033)	1.727(0.965)	1.893(1.059)
120	10	400	0.014(0.001)	1.63(1.058)	1.556(0.975)	1.676(1.069)	1.647(1.009)
240	10	400	0.014(0.001)	1.711(0.985)	1.527(1.077)	1.728(0.983)	1.568(1.075)
60	20	400	0.012(0.002)	1.511(0.914)	1.561(0.926)	1.611(0.936)	1.719(0.949)
120	20	400	0.012(0.001)	1.502(0.923)	1.452(0.934)	1.543(0.931)	1.534(0.945)
240	20	400	0.012(0.001)	1.569(0.929)	1.373(0.915)	1.589(0.931)	1.407(0.912)
60	40	400	0.015(0.002)	1.907(1.108)	1.964(1.166)	2.033(1.14)	2.159(1.181)
120	40	400	0.015(0.001)	1.967(1.319)	1.831(1.107)	2.021(1.334)	1.937(1.117)
240	40	400	0.015(0.001)	2.062(1.314)	1.775(1.118)	2.086(1.31)	1.823(1.111)

Bibliography

Aggarwal, C. and K. Subbian (2014). Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR) 47*(1), 10.

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research 9*(Sep), 1981–2014.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*, Volume 14. Wiley New York.

Apanasovich, T. V. and M. G. Genton (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika 97*(1), 15–30.

Apanasovich, T. V., M. G. Genton, and Y. Sun (2012). A valid matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association 107*(497), 180–193.

Bader, B. W., R. A. Harshman, and T. G. Kolda (2007). Temporal analysis of semantic graphs using asalsan. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pp. 33–42. IEEE.

Bahadori, M. T., Q. R. Yu, and Y. Liu (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in neural information processing systems*, pp. 3491–3499.

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

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

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

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

Bai, J. and S. Ng (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics 25*(1), 52–60.

Bansal, N., R. A. Connolly, and C. Stivers (2014). The stock-bond return relation, the term structures slope, and asset-class risk dynamics. *Journal of Financial and Quantitative Analysis 49*(3), 699–724.

Barrat, A., M. Barthelemy, and A. Vespignani (2008). *Dynamical processes on complex networks*. Cambridge university press.

Bornn, L., G. Shaddick, and J. V. Zidek (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association 107*(497), 281–289.

Bourgault, G. and D. Marcotte (1991). Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology 23*(7), 899–928.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bradley, J. R., S. H. Holan, C. K. Wikle, et al. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics 9*(4), 1761–1791.

Brockwell, P. J. and R. A. Davis (2013). *Time series: theory and methods*. Springer Science & Business Media.

Carlin, B. P., S. Banerjee, et al. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics 7*, 45–63.

Chamberlain, G. (1983, September). Funds, Factors, and Diversification in Arbitrage Pricing Models. *Econometrica 51*(5), 1305–23.

Chamberlain, G. and M. Rothschild (1983, September). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica 51*(5), 1281–304.

Chang, J., B. Guo, and Q. Yao (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics 189*(2), 297–312.

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

Chen, Y. (2017). Multivariate kriging on latent low-dimensional structures. *Unpublished technical report*.

Christensen, W. F. and Y. Amemiya (2001). Generalized shifted-factor analysis method for multivariate geo-referenced data. *Mathematical Geosciences 33*(7), 801.

Christensen, W. F. and Y. Amemiya (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association 97*(457), 302–317.

Christensen, W. F. and Y. Amemiya (2003). Modeling and prediction for multivariate spatial factor analysis. *Journal of statistical planning and inference 115*(2), 543–564.

Congdon, P. (2004). A multivariate model for spatio-temporal health outcomes with an application to suicide mortality. *Geographical Analysis 36*(3), 234–258.

Cook, D., N. Cressie, J. Majure, and J. Symanzik (1994). Some dynamic graphics for spatial data (with multiple attributes) in a gis. In *Compstat*, pp. 105–119. Springer.

Crane, H. et al. (2016). Dynamic random networks and their graph limits. *The Annals of Applied Probability 26*(2), 691–721.

Cranmer, S. J., P. Leifeld, S. D. McClurg, and M. Rolfe (2016). Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Cressie, N., T. Shi, and E. L. Kang (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics 19*(3), 724–745.

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Daniels, M. J., Z. Zhou, and H. Zou (2006). Conditionally specified space-time models for multivariate processes. *Journal of Computational and Graphical Statistics 15*(1), 157–177.

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and computing 18*(2), 173–183.

De Iaco, S., D. Myers, M. Palma, and D. Posa (2013). Using simultaneous diagonalization to identify a space–time linear coregionalization model. *Mathematical Geosciences 45*(1), 69–86.

De Iaco, S., M. Palma, and D. Posa (2013). Prediction of particle pollution through spatio-temporal multivariate geostatistical analysis: spatial special issue. *AStA Advances in Statistical Analysis 97*(2), 133–150.

Diebold, F. X. and C. Li (2006). Forecasting the term structure of government bond yields. *Journal of econometrics 130*(2), 337–364.

Diebold, F. X., C. Li, and V. Z. Yue (2008). Global yield curve dynamics and interactions: a dynamic nelson–siegel approach. *Journal of Econometrics 146*(2), 351–363.

Diebold, F. X., M. Piazzesi, and G. Rudebusch (2005). Modeling bond yields in finance and macroeconomics. Technical report, National Bureau of Economic Research.

Diebold, F. X., G. D. Rudebusch, and S. B. Aruoba (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of econometrics 131*(1), 309–338.

Draief, M. and L. Massouli (2010). *Epidemics and rumours in complex networks*. Cambridge University Press.

Duijn, M. A., T. A. Snijders, and B. J. Zijlstra (2004). p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica 58*(2), 234–254.

Durand, D. E. (1953). Country classification. In R. G. D. Allen and E. J. Ely (Eds.), *International Trade Statistics*, pp. 117–129. Wiley.

Erdős, P. and A. Rényi (1959). On random graphs i. *Publ. Math. Debrecen 6*, 290–297.

Erdős, P. and A. Rényi (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5*(1), 17–60.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of statistics 44*(1), 219.

Fan, J. and Q. Yao (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. Springer.

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

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004). The generalized dynamic factor model consistency and rates. *Journal of Econometrics 119*(2), 231–255.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Frank, O. and D. Strauss (1986). Markov graphs. *Journal of the american Statistical association 81*(395), 832–842.

French, K. R. (2017). 100 Portfolios Formed on Size and Book-to-Markete. `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_100_port_sz.html`. [Online; Accessed 01-Jan-2017].

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika 89*(1), 197–210.

Gaspari, G. and S. E. Cohn (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society 125*(554), 723–757.

Gaspari, G., S. E. Cohn, J. Guo, and S. Pawson (2006). Construction and application of covariance functions with variable length-fields. *Quarterly Journal of the Royal Meteorological Society 132*(619), 1815–1838.

Gelfand, A. E., A. M. Schmidt, S. Banerjee, and C. Sirmans (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test 13*(2), 263–312.

Genton, M. G., W. Kleiber, et al. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science 30*(2), 147–163.

Geyer, C. J. (1991). Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association 86*(415), 717–724.

Gneiting, T., W. Kleiber, and M. Schlather (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association 105*(491), 1167–1177.

Goulard, M. and M. Voltz (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology 24*(3), 269–286.

Grzebyk, M. and H. Wackernagel (1994). Multivariate analysis and spatial/temporal scales: real and complex models. In *Proceedings of the XVIIth International Biometrics Conference*, Volume 1, pp. 19–33.

Gupta, A. K. and D. K. Nagar (1999). *Matrix variate distributions*, Volume 104. CRC Press.

Hall, P., N. I. Fisher, and B. Hoffmann (1994). On the nonparametric estimation of covariance functions. *The Annals of Statistics*, 2115–2134.

Hanneke, S., W. Fu, E. P. Xing, et al. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics 4*, 585–605.

Harshman, R. A. (1978). Models for analysis of asymmetrical relationships among n objects or stimuli. In *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, McMaster University, Hamilton, Ontario*, Volume 5.

Harshman, R. A., P. E. Green, Y. Wind, and M. E. Lundy (1982). A model for the analysis of asymmetric data in marketing research. *Marketing Science 1*(2), 205–242.

Harshman, R. A. and M. E. Lundy (1996). Uniqueness proof for a family of models sharing features of tucker's three-mode factor analysis and parafac/candecomp. *Psychometrika 61*(1), 133–154.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*, Volume 1.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 795–800.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics 5*(2), 173–190.

Higdon, D., J. Swall, and J. Kern (1999). Non-stationary spatial modeling. *Bayesian statistics 6*(1), 761–768.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association 100*(469), 286–295.

Hoff, P. D. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pp. 657–664.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory 15*(4), 261.

Hoff, P. D. (2015a). Dyadic data analysis with amen. *arXiv preprint arXiv:1506.08237*.

Hoff, P. D. (2015b). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics 9*(3), 1169.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association 97*(460), 1090–1098.

Holland, D. M., N. Saltzman, L. H. Cox, and D. Nychka (1998). Spatial prediction of sulfur dioxide in the eastern united states. In *geoENV IIGeostatistics for Environmental Applications*, pp. 65–76. Springer.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association 76*(373), 33–50.

Huang, D., Q. Yao, and R. Zhang (2016). Krigings over space and time based on latent low-dimensional structures. *arXiv preprint arXiv:1609.06789*.

Huisman, M. and T. A. Snijders (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological methods & research 32*(2), 253–287.

IMF (2017). Direction of trade statistics, international monetary fund.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(1), 016107.

Kiers, H. A. (1989). An alternating least squares algorithm for fitting the two-and three-way dedicom model and the idioscal model. *Psychometrika 54*(3), 515–521.

Kiers, H. A. (1993). An alternating least squares algorithm for parafac2 and three-way dedicom. *Computational Statistics & Data Analysis 16*(1), 103–118.

Kleiber, W. and D. Nychka (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis 112*, 76–91.

Kolaczyk, E. D. and G. Csárdi (2014). *Statistical analysis of network data with R*, Volume 65. Springer.

Kollo, T. and D. von Rosen (2006). *Advanced multivariate statistics with matrices*, Volume 579. Springer Science & Business Media.

Krivitsky, P. N. and M. S. Handcock (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 29–46.

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

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

Lam, C., Q. Yao, and N. Bathia (2011a). Estimation of latent factors for high-dimensional time series. *Biometrika 98*(4), 901–18.

Lam, C., Q. Yao, and N. Bathia (2011b). Estimation of latent factors for high-dimensional time series. *Biometrika 98*(4), 901–918.

Lee, D. D. and H. S. Seung (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562.

Leng, C. and C. Y. Tang (2012). Sparse matrix graphical models. *Journal of the American Statistical Association 107*(499), 1187–1200.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Linnemann, H. (1966). *An econometric study of international trade flows*, Volume 234. North-Holland Publishing Company Amsterdam.

Liu, X. and R. Chen (2016a). Regime-switching factor models for high-dimensional time series. *Statistica Sinica 26*, 1427–1451.

Liu, X. and R. Chen (2016b). Regime-switching factor models for high-dimensional time series. *Statistica Sinica 26*, 1427–1451.

Liu, Y., H. H. Zhang, C. Park, and J. Ahn (2007). Support vector machines with adaptive lq penalty. *Computational Statistics & Data Analysis 51*(12), 6380–6394.

Lopes, H. F., E. Salazar, D. Gamerman, et al. (2008). Spatial dynamic factor analysis. *Bayesian Analysis 3*(4), 759–792.

Lozano, A. C., H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe (2009). Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 587–596. ACM.

Lusher, D., J. Koskinen, and G. Robins (2012). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Majumdar, A. and A. E. Gelfand (2007). Multivariate spatial modeling for geostatistical data using convolved covariance functions. *Mathematical Geology 39*(2), 225–245.

Majumdar, A., D. Paul, and D. Bautista (2010). A generalized convolution model for multivariate nonstationary spatial processes. *Statistica Sinica*, 675–695.

Majure, J. J. and N. Cressie (1997). Dynamic graphics for exploring spatial dependence in multivariate spatial data. *Geographical Systems 4*(2), 131–158.

Mardia, K. V. and C. R. Goodall (1993). *Spatial-temporal statistical analysis of multivariate environmental monitoring data*, Volume 6 of *North-Holland Ser. Statist. Probab.*, pp. 347–386. North-Holland.

Matsuo, T., D. W. Nychka, and D. Paul (2011). Nonstationary covariance modeling for incomplete data: Monte carlo em approach. *Computational Statistics & Data Analysis 55*(6), 2059–2073.

Minhas, S., P. D. Hoff, and M. D. Ward (2016). Inferential approaches for network analyses: Amen for latent factor models. *arXiv preprint arXiv:1611.00460*.

Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of business*, 473–489.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*(455), 1077–1087.

Nychka, D., C. Wikle, and J. A. Royle (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling 2*(4), 315–331.

Obled, C. and J. Creutin (1986). Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *Journal of Climate and Applied meteorology 25*(9), 1189–1204.

OECD (2018). Consumer price index: Total all items for the united states [cpaltt01usm661s].

Paciorek, C. J. and M. J. Schervish (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics 17*(5), 483–506.

Pan, J. and Q. Yao (2008). Modelling multiple time series via common factors. *Biometrika*, 365–379.

Pena, D. and G. E. Box (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association 82*(399), 836–843.

Pettitt, A. N., I. S. Weir, and A. G. Hart (2002). A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12*(4), 353–367.

Porcu, E. and V. Zastavnyi (2011). Characterization theorems for some classes of covariance functions associated to vector valued random fields. *Journal of Multivariate Analysis 102*(9), 1293–1301.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks 29*(2), 173–191.

Rudebusch, G. D. and T. Wu (2008). A macro-finance model of the term structure, monetary policy and the economy. *The Economic Journal 118*(530), 906–926.

Sampson, P. D. (2010). *Constructions for nonstationary spatial processes*, pp. 119–130. CRC Press.

Sampson, P. D. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association 87*(417), 108–119.

Sang, H., M. Jun, and J. Z. Huang (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics*, 2519–2548.

Schmidt, A. M. and A. E. Gelfand (2003). A bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres 108*(D24).

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Snijders, T. (2006). *Statistical Methods for Network Dynamics*.

Snijders, T., C. Steglich, and M. Schweinberger (2007). *Modeling the coevolution of networks and behavior*, Chapter Chapter 3, pp. 41 – 72. Mahwah: Routledge Academic.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology 31*(1), 361–395.

Snijders, T. A. (2005). Models for longitudinal network data. *Models and methods in social network analysis 1*, 215–247.

Snijders, T. A., J. Koskinen, and M. Schweinberger (2010). Maximum likelihood estimation for social network dynamics. *The Annals of Applied Statistics 4*(2), 567.

Snijders, T. A., G. G. Van de Bunt, and C. E. Steglich (2010). Introduction to stochastic actor-based models for network dynamics. *Social networks 32*(1), 44–60.

Stein, M. L. (2005a). Nonstationary spatial covariance functions. *Unpublished technical report*.

Stein, M. L. (2005b). Space–time covariance functions. *Journal of the American Statistical Association 100*(469), 310–321.

Stephenson, J., C. Holmes, K. Gallagher, and A. Pintore (2005). A statistical technique for modelling non-stationary spatial processes. *Geostatistics Banff 2004*, 125–134.

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

Tsai, H. and R. S. Tsay (2010a). Constrained factor models. *Journal of the American Statistical Association 105*(492), 1593–1605.

Tsai, H. and R. S. Tsay (2010b). Constrained factor models. *Journal of the American Statistical Association 105*(492), 1593–1605.

Tsai, H., R. S. Tsay, E. M. Lin, and C.-W. Cheng (2016). Doubly constrained factor models with applications. *Statistica Sinica 26*, 1453–1478.

Tsay, R. S. (2013). *Multivariate Time Series Analysis: with R and financial applications*. John Wiley & Sons.

Tzala, E. and N. Best (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical methods in medical research 17*(1), 97–118.

Van Loan, C. and N. Pitsianis (1993). Approximation with kronecker products. In *Linear Algebra for Large Scale and Real-Time Applications*, pp. 293–314. Springer.

Vargas-Guzmán, J., A. Warrick, and D. Myers (2002). Coregionalization by linear combination of nonorthogonal components. *Mathematical Geology 34*(4), 405–419.

Ver Hoef, J. M. and R. P. Barry (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference 69*(2), 275–294.

Ver Hoef, J. M. and N. Cressie (1993). Multivariable spatial prediction. *Mathematical Geology 25*(2), 219–240.

Ver Hoef, J. M., N. Cressie, and R. P. Barry (2004). Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform (fft). *Journal of Computational and Graphical Statistics 13*(2), 265–282.

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma 62*(1-3), 83–92.

Wackernagel, H. (2006). *Multivariate Kriging*. John Wiley & Sons, Ltd.

Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Walden, A. and A. Serroukh (2002). Wavelet analysis of matrix–valued time–series. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 458, pp. 157–179. The Royal Society.

Wang, D., X. Liu, and R. Chen (2017). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*.

Wang, F. and M. M. Wall (2003). Generalized common spatial factor model. *Biostatistics 4*(4), 569–582.

Ward, M. D. and P. D. Hoff (2007). Persistent patterns of international commerce. *Journal of Peace Research 44*(2), 157–175.

Warner, R. M., D. A. Kenny, and M. Stoto (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology 37*(10), 1742.

Wasserman, S. and P. Pattison (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p^*. *Psychometrika 61*(3), 401–425.

Werner, K., M. Jansson, and P. Stoica (2008). On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing 56*(2), 478–491.

Westveld, A. H. and P. D. Hoff (2011). A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, 843–872.

Wikle, C. K. (2010). *Low-rank representations for spatial processes*, pp. 107–118. CRC Press.

Xing, E. P., W. Fu, L. Song, et al. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics 4*(2), 535–566.

Yang, D., X. Han, and R. Chen (2017). Autoregressive models for matrix-valued time series. *Working paper*.

Yin, J. and H. Li (2012). Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis 107*, 119–140.

Zhang, H. H. and W. Lu (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika 94*(3), 691–703.

Zhao, J. and C. Leng (2014). Structured lasso for regression with matrix covariates. *Statistica Sinica*, 799–814.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(2), 463–483.

Zhou, S. et al. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics 42*(2), 532–562.

Zhu, J., J. Eickhoff, and P. Yan (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics 61*(3), 674–683.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.