The London School of Economics and Political Science

# Three Essays on Time Series: Spatio-Temporal Modelling, Dimension Reduction and Change-Point Detection

Baojun Dou

Supervisor: Prof. Qiwei Yao

*A thesis submitted to the Department of Statistics of the London School of Economics for the degree of Doctor of Philosophy, London, September 2015*

# Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). I confirm that Section 1.5 was jointly co-authored with Dr. Maria Lucia Parrell from University of Salerno, Italy and Section 3.2 was jointly co-authored with Prof. Rongmao Zhang from Zhejiang University, China.

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

# Abstract

Modelling high dimensional time series and non-stationary time series are two import aspects in time series analysis nowadays. The main objective of this thesis is to deal with these two problems. The first two parts deal with high dimensionality and the third part considers a change point detection problem.

In the first part, we consider a class of spatio-temporal models which extend popular econometric spatial autoregressive panel data models by allowing the scalar coefficients for each location (or panel) different from each other. The model is of the following form:

$$\mathbf{y}_t = D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t + D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1} + D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{p,t})^T$ represents the observations from $p$ locations at time $t$, $D(\boldsymbol{\lambda}_k) = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kp})$ and $\lambda_{kj}$ is the unknown coefficient parameter for the $j$-th location, and $\mathbf{W}$ is the $p \times p$ spatial weight matrix which measures the dependence among different locations. All the elements on the main diagonal of $\mathbf{W}$ are zero. It is a common practice in spatial econometrics to assume $\mathbf{W}$ known. For example, we may let $w_{ij} = 1/(1 + d_{ij})$, for $i \neq j$, where $d_{ij} \geq 0$ is an appropriate distance between the $i$-th and the $j$-th location. It can simply be the geographical distance between the two locations or the distance reflecting the correlation or association between the variables at the two locations. In the above model, $D(\boldsymbol{\lambda}_0)$ captures the pure spatial effect, $D(\boldsymbol{\lambda}_1)$ captures the pure dynamic effect, and $D(\boldsymbol{\lambda}_2)$ captures the time-lagged spatial effect. We also assume that the error term $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{p,t})^T$ in (1) satisfies the condition $\text{Cov}(\mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t) = 0$. When $\lambda_{k1} = \cdots = \lambda_{kp}$ for all $k = 1, 2, 3$, (1) reduces to the model of Yu et al. (2008), in which there are only 3 unknown regressive coefficient parameters. In general the regression function in

(1) contains $3p$ unknown parameters. To overcome the innate endogeneity, we propose a generalized Yule-Walker estimation method which applies the least squares estimation to a Yule-Walker equation. The asymptotic theory is developed under the setting that both the sample size and the number of locations (or panels) tend to infinity under a general setting for stationary and $\alpha$-mixing processes, which includes spatial autoregressive panel data models driven by i.i.d. innovations as special cases. The proposed methods are illustrated using both simulated and real data.

In part 2, we consider a multivariate time series model which decomposes a vector process into a latent factor process and a white noise process. Let $\mathbf{y}_t = (y_{1,t}, \cdots, y_{p,t})^T$ be an observable $p \times 1$ vector time series process. The factor model decomposes $\mathbf{y}_t$ in the following form:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{2}$$

where $\mathbf{x}_t = (x_{1,t}, \cdots, x_{r,t})^T$ is a $r \times 1$ latent factor time series with unknown $r \leq p$ and $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r)$ is a $p \times r$ unknown constant matrix. $\boldsymbol{\varepsilon}_t$ is a white noise process with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. The first part of (2) is a dynamic part and the serial dependence of $\mathbf{y}_t$ is driven by $\mathbf{x}_t$. We will achieve dimension reduction once $r \ll p$ in the sense that the dynamics of $\mathbf{y}_t$ is driven by a much lower dimensional process $\mathbf{x}_t$. Motivated by practical needs and the characteristic of high dimensional data, the sparsity assumption on factor loading matrix is imposed. Different from Lam, Yao and Bathia (2011)'s method, which is equivalent to an eigenanalysis of a non negative definite matrix, we add a constraint to control the number of nonzero elements in each column of the factor loading matrix. Our proposed sparse estimator is then the solution of a constrained optimization problem. The asymptotic theory is developed under the setting that both the

sample size and the dimensionality tend to infinity. When the common factor is weak in the sense that $\delta > 1/2$ in Lam, Yao and Bathia (2011)'s paper, the new sparse estimator may have a faster convergence rate. Numerically, we employ the generalized deflation method (Mackey (2009)) and the GSLDA method (Moghaddam et al. (2006)) to approximate the estimator. The tuning parameter is chosen by cross validation. The proposed method is illustrated with both simulated and real data examples.

The third part is a change point detection problem. we consider the following covariance structural break detection problem:

$$\mathrm{Cov}(\mathbf{y}_t)I(t_{j-1} \leq t < t_j) = \mathbf{\Sigma}_{t_{j-1}},\ j = 1, \cdots, m+1,$$

where $\mathbf{y}_t$ is a $p \times 1$ vector time series, $\mathbf{\Sigma}_{t_{j-1}} \neq \mathbf{\Sigma}_{t_j}$ and $\{t_1, \ldots, t_m\}$ are change points, $1 = t_0 < t_1 < \cdots < t_{m+1} = n$. In the literature, the number of change points $m$ is usually assumed to be known and small, because a large $m$ would involve a huge amount of computational burden for parameters estimation. By reformulating the problem in a variable selection context, the group least absolute shrinkage and selection operator (LASSO) is proposed to estimate $m$ and the locations of the change points $\{t_1, \ldots, t_m\}$. Our method is model free, it can be extensively applied to multivariate time series, such as GARCH and stochastic volatility models. It is shown that both $m$ and the locations of the change points $\{t_1, \ldots, t_m\}$ can be consistently estimated from the data, and the computation can be efficiently performed. An improved practical version that incorporates group LASSO and the stepwise regression variable selection technique are discussed. Simulation studies are conducted to assess the finite sample performance.

# Acknowledgments

I would like to express my deepest gratitude and utmost respect to my supervisor Professor Qiwei Yao. His continued guidance, encouragement and tremendous support leads me to the completion of this work during my four years PhD and one year MSc study. He is not only a supervisor of my research but also of my life. It is my great honor in my life to be his student.

I would also like to thank my second supervisor, Dr. Clifford Lam, who helped me a lot and answered me a lot of tedious and time consuming technique problems in the literatures. I am grateful to Professor Oliver Linton and Dr. Barigozzi Matteo for willingly accepting to be part of the examination committee and to evaluate this thesis.

I am immensely indebted to my wife Duo Xu and my parents Wange Dou and Guiqin Hou for their love and support throughout all my life.

# Contents

# Chapter 1: Generalized Yule-Walker Estimation for Spatio-Temporal Models with Unknown Diagonal Coefficients

## 1.1 Introduction

The class of spatial autoregressive (SAR) models is introduced to model cross sectional dependence of different economic individuals at different locations (Cliff and Ord, 1973). More recent developments extend SAR models to spatial dynamic panel data (SDPD) models, i.e. adding time lagged terms to account for serial correlations across different locations. See, e.g. Lee and Yu (2010a). Baltagi et al. (2003) considers a static spatial panel model where the error term is a SAR model. Lin and Lee (2010) shows that in the presence of heteroskedastic disturbances, the maximum likelihood estimator for the SAR models without taking into account the heteroskedasticity is generally inconsistent and proposes an alternative GMM estimation method. Computationally the GMM methods

are more efficient than the QML estimation (Lee, 2001). Lee and Yu (2010a) classifies SDPD models into three categories: stable, spatial cointegration and explosive cases. As pointed out by Bai and Shi (2011), the cases with a large number of cross sectional units and a long history are rare. Hence it is pertinent to consider the setting with short time spans in order to include as many locations as possible. Both estimation method and asymptotic analysis need to be adapted under this new setting. Yu et al. (2008) and Yu et al. (2012) investigate the asymptotic properties when both the number of locations and the length of time series tend to infinity for both the stable case and spatial cointegration case, and show that QMLE is consistent.

Motivated by the evidence in some practical examples, we extend the model in Yu et al. (2008) and Yu et al. (2012) by allowing the scalar coefficients for each location (or panel) different from each other. This increase in model capacity comes with the cost of estimating substantially more parameters. In fact that the number of the parameters in this new setting is in the order of the number of locations. The model considered in this paper has four additive components: a pure spatial effect, a pure dynamic effect, a time-lagged spatial effect and a white noise. Due to the innate endogeneity, the conventional regression estimation methods such as the least squares method directly based on the model lead to inconsistent estimators. To overcome the difficulties caused by the endogeneity, we propose a generalized Yule-Walker type estimator for estimating the parameters in the model, which applies the least squares estimation to a Yule-Walker equation. The asymptotic normality of the proposed estimators is established under the setting that both the sample size $n$ and the number of locations (or panels) $p$ tend to infinity. Therefore the number of parameters to be estimated also diverges to infinity, which is a marked difference from, e.g., Yu et al.

(2012). We develop the asymptotic properties under a general setting for stationary and $\alpha$-mixing processes, which includes the spatial autoregressive panel data models driven by *i.i.d.* innovations as special cases.

The rest of the paper is organized as follows. Section 1.2 introduces the new model, its motivation and the generalized Yule-Walker estimation method. The asymptotic theory for the proposed estimation method is presented in Section 1.3. Simulation results and real data analysis are reported, respectively, in Section 1.4 and 1.5. All the technical proofs are relegated to an Appendix.

## 1.2 Model and Estimation Method

### 1.2.1 Models

The model considered in this paper is of the following form:

$$\mathbf{y}_t = D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t + D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1} + D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1.2.1}$$

where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{p,t})^T$ represents the observations from $p$ locations at time $t$, $D(\boldsymbol{\lambda}_k) = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kp})$ and $\lambda_{kj}$ is the unknown coefficient parameter for the $j$-th location, and $\mathbf{W}$ is the $p \times p$ spatial weight matrix which measures the dependence among different locations. All the main diagonal elements of $\mathbf{W}$ are zero. It is a common practice in spatial econometrics to assume $\mathbf{W}$ known. For example, we may let $w_{ij} = 1/(1 + d_{ij})$, for $i \neq j$, where $d_{ij} \geq 0$ is an appropriate distance between the $i$-th and the $j$-th location. It can simply be the geographical distance between the two locations or the distance reflecting the correlation or association between the variables at the two locations. In the above model, $D(\boldsymbol{\lambda}_0)$ captures the pure spatial effect, $D(\boldsymbol{\lambda}_1)$ captures the pure dynamic effect,

and $D(\boldsymbol{\lambda}_2)$ captures the time-lagged spatial effect. We also assume that the error term $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{p,t})^T$ in (1.2.1) satisfies the condition $\mathrm{Cov}\,(\mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t) = 0$. When $\lambda_{k1} = \cdots = \lambda_{kp}$ for $k = 0, 1, 2$, (1.2.1) reduces to the model of Yu et al. (2008), in which there are only 3 unknown regressive coefficient parameters. In general the regression function in (1.2.1) contains $3p$ unknown parameters.

The extension to use different scalar coefficients for different locations is motivated by practical needs. For example, we analyze the monthly change rates of the consumer price index (CPI) for the EU member states over the years 2003-2010. The detailed analysis for this data set will be presented in section 1.5. Figure 1.1 presents the scatter-plots of the observed data $y_{i,t}$ versus the spatial regressor $\mathbf{w}_i^T\mathbf{y}_t$ and $y_{i,t-1}$, for some of the EU member states, where $\mathbf{w}_i^T$ is the $i$-th row vector of the weight matrix $\mathbf{W}$ which is taken as the sample correlation matrix with all the elements on the main diagonal set to be 0. The superimposed straight lines are the simple regression lines estimated using the newly proposed method in Section 2.2 below. It is clear from Figure 1.1 that at least Greece and Belgium should have a different slope from those of France or Iceland.

### 1.2.2 Generalized Yule-Walker estimation

As $\mathbf{y}_t$ occurs on both sides of (1.2.1), $\mathbf{W}\mathbf{y}_t$ and $\boldsymbol{\varepsilon}_t$ are correlated with each other. Applying least squares method directly based on regressing $\mathbf{y}_t$ on $(\mathbf{W}\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{W}\mathbf{y}_{t-1})$ leads to inconsistent estimators. On the other hand, applying the maximum likelihood estimation requires to profile a $p \times p$ nuisance parameter matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathrm{Var}(\boldsymbol{\varepsilon}_t)$, which leads to a complex nonlinear optimization problem. Furthermore when $p$ is large in relation to $n$, the numerical stability is of concern.

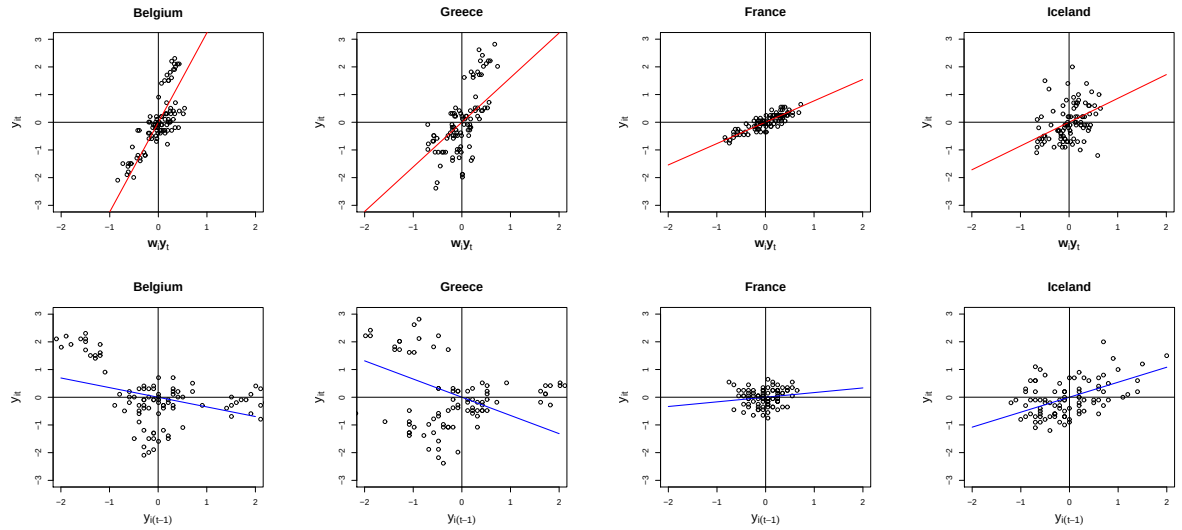


Figure 1.1: Plots of the monthly change rates $y_{i,t}$ of CPI against the spatial regressor $\mathbf{w}_i^T\mathbf{y}_t$ (on the top) and the dynamic regressor $y_{i,t-1}$ (on the bottom) for four EU member states in 2003-2010. The superimposed straight lines were estimated by the newly proposed method in Section 2.2.

We propose below a new estimation method which applies the least squares method to each individual row of a Yule-Walker equation. To this end, let $\mathbf{\Sigma}_k = \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t)$ for any $k \geq 0$. Note that we always assume that $\mathbf{y}_t$ is stationary, see condition A2 and Remark 1 in Section 1.3 below. Then the Yule-Walker equation below follows from (1.2.1) directly.

$$(\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W})\mathbf{\Sigma}_1 = (D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W})\mathbf{\Sigma}_0,$$

where $\mathbf{I}$ is a $p \times p$ identity matrix. The $i$-th row of the above equation is

$$(\mathbf{e}_i^T - \lambda_{0i}\mathbf{w}_i^T)\mathbf{\Sigma}_1 = (\lambda_{1i}\mathbf{e}_i^T + \lambda_{2i}\mathbf{w}_i^T)\mathbf{\Sigma}_0, \quad i = 1, \ldots, p, \tag{1.2.2}$$

where $\mathbf{w}_i$ is the $i$-th row vector of $\mathbf{W}$, and $\mathbf{e}_i$ is the unit vector with the $i$-th element equal to 1. Note that (1.2.2) is a system of $p$ linear equations with three unknown parameters $\lambda_{0i}$, $\lambda_{1i}$ and $\lambda_{2i}$. Since $E\mathbf{y}_t = 0$, we replace $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_0$ by the sample (auto)covariance

matrices

$$\widehat{\mathbf{\Sigma}}_1 = \frac{1}{n}\sum_{t=1}^{n-1}\mathbf{y}_{t+1}\mathbf{y}_t^T \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_0 = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t\mathbf{y}_t^T.$$

We estimate $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})^T$ by the least squares method, i.e. to solve the minimization problem

$$\min_{\lambda_{0i},\lambda_{1i},\lambda_{2i}} \|\widehat{\mathbf{\Sigma}}_1^T(\mathbf{e}_i - \lambda_{0i}\mathbf{w}_i) - \widehat{\mathbf{\Sigma}}_0(\lambda_{1i}\mathbf{e}_i + \lambda_{2i}\mathbf{w}_i)\|_2^2.$$

The resulting estimators are called generalized Yule-Walker estimators which admits the explicit expression:

$$(\widehat{\lambda}_{0i}, \widehat{\lambda}_{1i}, \widehat{\lambda}_{2i})^T = (\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i)^{-1}\widehat{\mathbf{X}}_i^T\widehat{\mathbf{Y}}_i, \tag{1.2.3}$$

where

$$\widehat{\mathbf{X}}_i = (\widehat{\mathbf{\Sigma}}_1^T\mathbf{w}_i, \widehat{\mathbf{\Sigma}}_0\mathbf{e}_i, \widehat{\mathbf{\Sigma}}_0\mathbf{w}_i) \quad \text{and} \quad \widehat{\mathbf{Y}}_i = \widehat{\mathbf{\Sigma}}_1^T\mathbf{e}_i.$$

More explicitly,

$$\widehat{\mathbf{X}}_i = \left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t), \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}y_{i,t-1}, \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_{t-1})\right), \quad \widehat{\mathbf{Y}}_i = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}y_{i,t}.$$

Then it holds that for $i = 1, \cdots, p$,

$$\begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} = (\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i)^{-1} \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t) \times \frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T y_{i,t-1} \times \frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1}) \times \frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} \end{pmatrix}.$$

### 1.2.3 A root-$n$ consistent estimator for large $p$

When $p/\sqrt{n} \to \infty$, the estimator (1.2.3) admits non-standard convergence rates (i.e. the rates different from $\sqrt{n}$); see Theorems 2 and 4 in Section 1.3 below. Note that there are $p$ equations with only 3 parameters in (1.2.2). Hence (1.2.3) can be viewed as a GMME for an over-determined scenario. The estimation may suffer when the number of estimation

equations increases. See, for example, a similar result in Theorem 1 of Chang, Chen and Chen (2015). A further compounding factor is that the estimation for the covariance matrices $\mathbf{\Sigma}_0$, $\mathbf{\Sigma}_1$ using their sample counterparts leads to non-negligible errors even when $n \to \infty$ (as long as $p$ is very large). Below we propose an alternative estimator which restricts the number of the estimation equations to be used in order to restore the $\sqrt{n}$-consistency and the asymptotic normality.

For $i = 1, \cdots, p$, put $\mathbf{X}_i = (\mathbf{\Sigma}_1^T \mathbf{w}_i, \mathbf{\Sigma}_0 \mathbf{e}_i, \mathbf{\Sigma}_0 \mathbf{w}_i)$. Note that the $k$-th row of $\mathbf{X}_i$ is $(\mathbf{e}_k^T \mathbf{\Sigma}_1^T \mathbf{w}_i,$

$\mathbf{e}_k^T \mathbf{\Sigma}_0 \mathbf{e}_i, \mathbf{e}_k^T \mathbf{\Sigma}_0 \mathbf{w}_i)$ which is the covariance between $y_{k,t-1}$ and $(\mathbf{w}_i^T \mathbf{y}_t,\ y_{i,t-1},\ \mathbf{w}_i^T \mathbf{y}_{t-1})$. Let

$$\rho_k^{(i)} = \left|\mathbf{e}_k^T \mathbf{\Sigma}_1^T \mathbf{w}_i\right| + \left|\mathbf{e}_k^T \mathbf{\Sigma}_0 \mathbf{e}_i\right| + \left|\mathbf{e}_k^T \mathbf{\Sigma}_0 \mathbf{w}_i\right|, \quad k = 1, \cdots, p. \tag{1.2.4}$$

Then $\rho_k^{(i)}$ may be viewed as a measure for the correlation between $y_{k,t-1}$ and $(\mathbf{w}_i^T \mathbf{y}_t, y_{i,t-1}, \mathbf{w}_i^T \mathbf{y}_{t-1})^T$. When $\rho_k^{(i)}$ is small, say, close to 0, the $k$-th equation in (1.2.2) carries little information on $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})$. Therefore as far as the estimation for $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})$ is concerned, we only keep the $k$-th equation in (1.2.2) for large $\rho_k^{(i)}$.

Let $\mathbf{z}_{t-1}^i$ be the $d_i \times 1$ vector consisting of those $y_{k,t-1}$ corresponding to the $d_i$ largest $\widehat{\rho}_k^{(i)}$ $(1 \le k \le p)$, where $\widehat{\rho}_k^{(i)}$ is defined as in (1.2.4) but with $(\mathbf{\Sigma}_1,\ \mathbf{\Sigma}_0)$ replaced by $(\widehat{\mathbf{\Sigma}}_1,\ \widehat{\mathbf{\Sigma}}_0)$. The new estimator is defined as

$$(\widetilde{\lambda}_{0i},\ \widetilde{\lambda}_{1i},\ \widetilde{\lambda}_{2i})^T = (\widehat{\mathbf{Z}}_i^T \widehat{\mathbf{Z}}_i)^{-1} \widehat{\mathbf{Z}}_i^T \widetilde{\mathbf{Y}}_i, \quad i = 1, \cdots, p. \tag{1.2.5}$$

where

$$\widehat{\mathbf{Z}}_i = \Big(\frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_{t-1}^i (\mathbf{w}_i^T \mathbf{y}_t),\ \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_{t-1}^i y_{i,t-1},\ \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_{t-1}^i (\mathbf{w}_i^T \mathbf{y}_{t-1})\Big), \tag{1.2.6}$$

and

$$\widetilde{\mathbf{Y}}_i = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_{t-1}^i y_{i,t}.$$

Now it holds that

$$\begin{pmatrix} \widetilde{\lambda}_{0i} \\ \widetilde{\lambda}_{1i} \\ \widetilde{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} = (\widehat{\mathbf{Z}}_i^T \widehat{\mathbf{Z}}_i)^{-1} \widehat{\mathbf{Z}}_i^T \begin{pmatrix} \frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{z}_{t-1}^i \\ \frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{z}_{t-1}^i \\ \frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{z}_{t-1}^i \end{pmatrix}.$$

Theorem 3 in Section 3 below shows the asymptotic normality of the above estimator provided that the number of estimation equations used satisfies condition $d_i = o(\sqrt{n})$.

## 1.3 Theoretical properties

We introduce some notations first. For a $p \times 1$ vector $\mathbf{v} = (v_1, \cdots, v_p)^T$, $\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$ is the Euclidean norm, $\|\mathbf{v}\|_1 = \sum_{i=1}^p |v_i|$ is the $L_1$ norm. For a matrix $\mathbf{H} = (h_{ij})$, $\|\mathbf{H}\|_F = \sqrt{tr(\mathbf{H}^T\mathbf{H})}$ is the Frobenius norm, $\|\mathbf{H}\|_2 = \sqrt{\lambda_{\max}(\mathbf{H}^T\mathbf{H})}$ is the operator norm, where $\lambda_{\max}(\cdot)$ is the largest eigenvalue of a matrix. We denote by $|\mathbf{H}|$ the matrix $(|h_{ij}|)$ which is a matrix of the same size as $\mathbf{H}$ but with the $(i,j)$-th element $h_{ij}$ replaced by $|h_{ij}|$. Note the determinant of $\mathbf{H}$ is denoted by $\det(\mathbf{H})$. A strictly stationary process $\{\mathbf{y}_t\}$ is $\alpha$-mixing if

$$\alpha(k) \equiv \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} \big|P(A)P(B) - P(AB)\big| \to 0, \quad \text{as } k \to \infty, \tag{1.3.7}$$

where $\mathcal{F}_i^j$ denotes the $\sigma$-algebra generated by $\{\mathbf{y}_t, i \le t \le j\}$. See, e.g., Section 2.6 of Fan and Yao (2003) for a compact review of $\alpha$-mixing processes.

Let $\mathbf{S}(\boldsymbol{\lambda}_0) \equiv \mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W}$ be invertible. It follows from (1.2.1) that

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{S}^{-1}(\boldsymbol{\lambda}_0)\boldsymbol{\varepsilon}_t,$$

where $\mathbf{A} = \mathbf{S}^{-1}(\boldsymbol{\lambda}_0)(D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W})$. Some regularity conditions are now in order.

A1. The spatial weight matrix $\mathbf{W}$ is known with zero main diagonal elements; $\mathbf{S}(\boldsymbol{\lambda}_0)$ is invertible.

A2. (a) The disturbance $\boldsymbol{\varepsilon}_t$ satisfies

$$\mathrm{Cov}(\mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t) = 0.$$

(b) The process $\{\mathbf{y}_t\}$ in model (1.2.1) is strictly stationary and $\alpha$-mixing with $\alpha(k)$, defined in (1.3.7), satisfying

$$\sum_{k=1}^{\infty} \alpha(k)^{\frac{\gamma}{4+\gamma}} < \infty,$$

for some constant $\gamma > 0$.

(c) For $\gamma > 0$ specified in (b) above,

$$\sup_p \mathrm{E}\left|\mathbf{w}_i^T \boldsymbol{\Sigma}_0 \mathbf{y}_t\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{w}_i^T \boldsymbol{\Sigma}_1 \mathbf{y}_t\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T \boldsymbol{\Sigma}_0 \mathbf{y}_t\right|^{4+\gamma} < \infty,$$

$$\sup_p \mathrm{E}\left|\mathbf{w}_i^T \mathbf{y}_t\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T \mathbf{y}_t\right|^{4+\gamma} < \infty,$$

where $\mathbf{w}_i$ denotes the $i$-th row of $\mathbf{W}$. The diagonal elements of $\mathbf{V}_i$ defined in (1.3.8) are bounded uniformly in $p$.

A3. The rank of matrix $(\boldsymbol{\Sigma}_1^T \mathbf{w}_i, \boldsymbol{\Sigma}_0 \mathbf{e}_i, \boldsymbol{\Sigma}_0 \mathbf{w}_i)$ is equal to 3.

**Remark 1**. Condition A1 is standard for spatial econometric models. Condition A3 ensures that $\lambda_{0i}, \lambda_{1i}$ and $\lambda_{2i}$ are identifiable in (1.2.2). Condition A2(c) limits the dependence across different spatial locations. It is implied by, for example, the conditions imposed in Yu et al. (2008). Lemma 2 in the Appendix shows that Condition A2 holds with $\gamma = 4$ under conditions A1 and B1 – B3 below. Note that conditions B1–B3 are often directly imposed in the spatial econometrics literature including, for example, Lee and Yu (2010a), and Yu et al. (2008).

B1. The errors $\varepsilon_{i,t}$ are $i.i.d$ across $i$ and $t$ with $\mathrm{E}(\varepsilon_{i,t}) = 0$, $\mathrm{Var}(\varepsilon_{i,t}) = \sigma_0^2$, and $\mathrm{E}\,|\varepsilon_{i,t}|^{4+\gamma} < \infty$. The density function of $\varepsilon_{i,t}$ exists.

B2. The row and column sums of $|\mathbf{W}|$ and $|\mathbf{S}^{-1}(\boldsymbol{\lambda}_0)|$ are bounded uniformly in $p$.

B3. The row and column sums of $\sum_{j=0}^{\infty} |\mathbf{A}^j|$ are bounded uniformly in $p$.

Now we are ready to present the asymptotic properties for $(\widehat{\lambda}_{0i}, \widehat{\lambda}_{1i}, \widehat{\lambda}_{2i})^T$, $i = 1, \ldots, p$, with fixed $p$ and $n \to \infty$ first, and then $p \to \infty$ and $n \to \infty$.

### 1.3.1 Asymptotics for fixed $p$

For $i = 1, \ldots, p$, let

$$\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}(j) = \mathrm{Cov}(\mathbf{y}_{t-1+j}\varepsilon_{i,t+j}, \mathbf{y}_{t-1}\varepsilon_{i,t}), \qquad j = 0, 1, 2, \cdots,$$

$$\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i} = \boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}(0) + \sum_{j=1}^{\infty} \left[\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}(j) + \boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}^T(j)\right],$$

$$\mathbf{V}_i = \begin{pmatrix} \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0\mathbf{w}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0\mathbf{w}_i \end{pmatrix}, \tag{1.3.8}$$

and

$$\mathbf{U}_i = \begin{pmatrix} \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_1^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{w}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\mathbf{y},\varepsilon_i}\boldsymbol{\Sigma}_0\mathbf{w}_i \end{pmatrix}. \tag{1.3.9}$$

**Theorem 1** *Let conditions A1 – A3 hold and $p \geq 1$ be fixed. Then as $n \to \infty$, it holds that*

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix}\right) \xrightarrow{d} N(0, \mathbf{V}_i^{-1}\mathbf{U}_i\mathbf{V}_i^{-1}), \qquad i = 1, \ldots, p,$$

*where* $\mathbf{V}_i$ *and* $\mathbf{U}_i$ *are given in* (1.3.8) *and* (1.3.9).

### 1.3.2 Asymptotics with diverging $p$

When $p$ diverges together with $n$, $\mathbf{U}_i, \mathbf{V}_i$ in (1.3.9) and (1.3.8) are no longer constant matrices. Let $\mathbf{U}_i^{-\frac{1}{2}}$ be a matrix such that $(\mathbf{U}_i^{-\frac{1}{2}})^2 = \mathbf{U}_i^{-1}$.

**Theorem 2** *Let condition A1 – A3 hold.*

*(i) As* $n \to \infty$, $p \to \infty$ *and* $p = o(\sqrt{n})$,

$$\sqrt{n}\mathbf{U}_i^{-\frac{1}{2}}\mathbf{V}_i \left( \begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right) \xrightarrow{d} N(0, \mathbf{I}_3), \qquad i = 1, \ldots, p.$$

*(ii) As* $n \to \infty$, $p \to \infty$, $\sqrt{n} = O(p)$ *and* $p = o(n)$,

$$\left\| \begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\|_2 = O_p\left(\frac{p}{n}\right), \qquad i = 1, \ldots, p.$$

Intuitively, condition A2(c) reflects the spatial dependence, that is the structures of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. It includes the case that $y_{ti}$ and $y_{tj}$ are asymptotically uncorrelated given $i$ and $j$ are far enough. Hence for $y_{ti}$, as $p \to \infty$, the correlation of $y_{ti}$ and the far enough elements of IV $\mathbf{y}_{t-1}$ are asymptotically 0. This means more such IV's does not add more information to the estimation. At the same time, adding one more IV means we have one more estimation equation in GMM, noise then accumulates. This can explain what Theorem 2 says: given condition A2(c), if $p$ is sufficient small such that $p = o(\sqrt{n})$, using more IV does not improve the estimation, and the total noise accumulation is dominated by $1/\sqrt{n}$, hence the

effect of $p$ can not be seen anymore; When $p$ increases such that $\sqrt{n} << p << n$, using more IV still does not improve the estimation, however now the total noise accumulation reaches the extent such that $p/n$ dominates; When $p$ go on increasing such that $p \geq Cn$, the estimator is even inconsistent due to the noise accumulation.

Theorem 2 indicates that the standard root-$n$ convergence rate prevails as long as $p = o(\sqrt{n})$. However the convergence rate may be slower when $p$ is of higher orders than $\sqrt{n}$. Theorem 2 presents the convergence rates for the $L_2$ norm of the estimation errors. The rates also hold for the $L_1$ norm of the errors as well. Corollary 1 consider the estimation errors over $p$ locations together, for which we have established the result for $L_1$ norm only.

**Corollary 1** *Let condition A1 hold, and condition A2 and A3 hold for all $i = 1, \cdots, p$. Then as $n \to \infty$ and $p \to \infty$, it holds that*

$$\frac{1}{p}\sum_{i=1}^{p}\left\|\begin{pmatrix}\widehat{\lambda}_{0i}\\ \widehat{\lambda}_{1i}\\ \widehat{\lambda}_{2i}\end{pmatrix}-\begin{pmatrix}\lambda_{0i}\\ \lambda_{1i}\\ \lambda_{2i}\end{pmatrix}\right\|_1 = \begin{cases} O_p(\frac{1}{\sqrt{n}}) & \text{if } \frac{p}{\sqrt{n}} = O(1), \\ O_p(\frac{p}{n}) & \text{if } \frac{p}{\sqrt{n}} \to \infty \text{ and } \frac{p}{n} = o(1). \end{cases}$$

To derive the asymptotic properties of the estimators defined in (1.2.5), we introduce some new notation. For $i = 1, \ldots, p$, let

$$\mathbf{\Sigma}_0^i = \mathrm{Cov}(\mathbf{y}_t, \mathbf{z}_t^i), \quad \mathbf{\Sigma}_1^i = \mathrm{Cov}(\mathbf{y}_t, \mathbf{z}_{t-1}^i),$$

$$\mathbf{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(j) = \mathrm{Cov}(\mathbf{z}_{t-1+j}^i \varepsilon_{i,t+j}, \mathbf{z}_{t-1}^i \varepsilon_{i,t}), \qquad j = 0, 1, 2, \cdots,$$

and

$$\mathbf{\Sigma}_{\mathbf{z}^i,\varepsilon_i} = \mathbf{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(0) + \sum_{j=1}^{\infty}\left[\mathbf{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(j) + \mathbf{\Sigma}_{\mathbf{z}^i,\varepsilon_i}^T(j)\right].$$

Let

$$\mathbf{V}_i^* = \begin{pmatrix} \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i(\boldsymbol{\Sigma}_1^i)^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_0^i(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \end{pmatrix}, \tag{1.3.10}$$

and

$$\mathbf{U}_i^* = \begin{pmatrix} \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_1^i)^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{e}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \\ \mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i & \mathbf{e}_i^T\boldsymbol{\Sigma}_0^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i & \mathbf{w}_i^T\boldsymbol{\Sigma}_0^i\boldsymbol{\Sigma}_{\mathbf{z}^i,\varepsilon_i}(\boldsymbol{\Sigma}_0^i)^T\mathbf{w}_i \end{pmatrix}. \tag{1.3.11}$$

Theorem 3 below indicates that the estimators defined in (1.2.5) are asymptotically normal with the standard $\sqrt{n}$-rate as long as $d_i = o(\sqrt{n})$. Note that it does not impose any conditions directly on the size of $p$.

A4. (a) For $\gamma > 0$ specified in A2(b),

$$\sup_p \mathrm{E}\left|\mathbf{w}_i^T\boldsymbol{\Sigma}_0^i\mathbf{z}_t^i\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{w}_i^T\boldsymbol{\Sigma}_1^i\mathbf{z}_t^i\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T\boldsymbol{\Sigma}_0^i\mathbf{z}_t^i\right|^{4+\gamma} < \infty,$$

$$\sup_p \mathrm{E}\left|\mathbf{w}_i^T\mathbf{y}_t\right|^{4+\gamma} < \infty, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T\mathbf{y}_t\right|^{4+\gamma} < \infty.$$

and the diagonal elements of $\mathbf{V}_i^*$ defined in (1.3.10) are bounded uniformly in $p$.

(b) The rank of matrix $\mathrm{E}\{\widehat{\mathbf{Z}}_i\}$ is equal to 3, where $\widehat{\mathbf{Z}}_i$ is defined in (1.2.6).

**Theorem 3** *Let conditions A1, A2(a,b) and A4 hold. As $n \to \infty$, $p \to \infty$ and $d_i = o(\sqrt{n})$, it holds that*

$$\sqrt{n}(\mathbf{U}_i^*)^{-\frac{1}{2}}\mathbf{V}_i^*\left(\begin{pmatrix} \widetilde{\lambda}_{0i} \\ \widetilde{\lambda}_{1i} \\ \widetilde{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix}\right) \xrightarrow{d} N(0, \mathbf{I}_3), \qquad i = 1, \ldots, p,$$

*where $\mathbf{V}_i^*$ and $\mathbf{U}_i^*$ are given in* (1.3.10) *and* (1.3.11).

The fact that more such IV's does not add more information to the estimation is because condition A2(c) restrict the spatial dependence of $\mathbf{y}_t$. If we relax it to include the (special) case that elements in $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are all bounded away from 0 as $p \to \infty$, then the correlation of $y_{ti}$ and $y_{tj}$ are bounded away from 0 no matter how far they are. Under this new condition, intuitively, more IV's does add more information to the estimation, which may improve our estimation. At the same time, the noise accumulation still exists. The tradeoff is about this two effect. The new condition is condition A5, which includes the case mentioned above.

A5. For $\gamma > 0$ specified in A2(b),

$$\max\left\{ \sup_p \mathrm{E}\left|\mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{y}_t\right|^{4+\gamma}, \quad \sup_p \mathrm{E}\left|\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_t\right|^{4+\gamma}, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{y}_t\right|^{4+\gamma} \right\} = O(s_0(p)).$$

$$\max\left\{ \sup_p \mathrm{E}\left|\mathbf{w}_i^T\mathbf{y}_t\right|^{4+\gamma}, \quad \sup_p \mathrm{E}\left|\mathbf{e}_i^T\mathbf{y}_t\right|^{4+\gamma} \right\} = O(s_1(p)).$$

and the diagonal elements of $\mathbf{V}_i$ defined in (1.3.8) is in the order of $s_2(p)$, where $s_0(p)$, $s_1(p)$ and $s_2(p)$ are numbers relating to $p$.

Let us denote $C$ as a constant. When the number of nonzero elements (or elements bounded away from zero) in $\mathbf{w}_i$ increases with $p$ but is $o(p)$, we may have $s_1(p) = o(\min\{s_0(p), s_2(p)\})$. Simulation scenario 2 is under this case. When there are only finite number of nonzero elements (or elements bounded away from zero) in $\mathbf{w}_i$, we might have $s_1(p) \asymp C$, which is the case of simulation scenario 1. The reason we assume the diagonal elements of $\mathbf{V}_i$ defined in (1.3.8) are in the order of $s_2(p)$ is because we can treat $\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{\Sigma}_1^T\mathbf{w}_i, \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{\Sigma}_0\mathbf{e}_i, \mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{\Sigma}_0\mathbf{w}_i$ as the second moments of three random variables $\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{x}, \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{x}$ and $\mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{x}$ respectively, where the $p \times 1$ random vector $\mathbf{x}$ has mean 0 and covariance matrix $\mathbf{I}_p$.

**Theorem 4** *Let conditions A1, A2(a,b), A3 and A5 hold. As* $n \to \infty$, $p \to \infty$, *if* $\frac{ps_1(p)}{s_2(p)} = o(n)$ *and* $s_0^{1/2}(p) = O(ps_1^{1/2}(p)s_2(p))$, *it holds that*

$$\left\| \begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\|_2 = O_p\Big( \max\Big\{ \frac{ps_1^{3/4}(p)}{ns_2(p)}, \frac{s_0^{1/4}(p)}{\sqrt{n}s_2(p)} \Big\} \Big).$$

Let us consider some examples. (1) When $s_0(p) \asymp p$, $s_1(p) \asymp C$ and $s_2(p) \asymp p$, the convergence rate is $\max\Big\{\frac{1}{n}, \frac{1}{\sqrt{n}p^{3/4}}\Big\}$. (2) When $s_0(p) \asymp p$, $s_1(p) \asymp \sqrt{p}$ and $s_2(p) \asymp p$, if $p = o(n^2)$, the convergence rate is $\max\Big\{\frac{p^{3/8}}{n}, \frac{1}{\sqrt{n}p^{3/4}}\Big\}$. (3) When $s_0(p) \asymp C$, $s_1(p) \asymp C$ and $s_2(p) \asymp C$, if $p = o(n)$, the convergence rate is $\max\Big\{\frac{p}{n}, \frac{1}{\sqrt{n}}\Big\}$, which corresponds with Theorem 2. Theorem 4 indicates that under different situations of $s_0(p)$, $s_1(p)$ and $s_2(p)$, we may obtain different convergence rates. These observations are illustrated by simulation examples in section 4.

Example (2) is similar to the case such that the correlation of $y_{ti}$ and $y_{tj}$ are bounded away from 0 no matter how far they are. Hence Tradeoff explanations is as follows: we say more IV add more information to the estimation as the positive effect and total noise accumulation by IV as the negative effect. When $p$ is sufficient small such that $p << n^{4/9}$, the positive effect dominates the negative effects, hence more IV increase the convergence rate; When $n^{4/9} << p << n^2$, the negative effect dominates the positive effect, hence more IV reduces the convergence rate. But compared with the case when there is no positive effect, we gain some convergent rate (for instance $\frac{p^{3/8}}{n} << \frac{p}{n}$), which means the positive effect is indeed doing its job; When $p \geq Cn^2$, negative effect totally dominates positive effect and the estimator is inconsistent.

## 1.4 Simulation study

To examine the finite sample performance of the proposed estimation methods, we conduct some simulation under different scenarios.

### 1.4.1 Scenario 1

$\lambda_{0i}$, $\lambda_{1i}$ and $\lambda_{2i}$ are generated from $U(-0.6, 0.6)$. The spatial weight matrix $\mathbf{W}$ used is a block diagonal matrix formed by a $\sqrt{p} \times \sqrt{p}$ row-normalized matrix $\mathbf{W}^*$. We construct $\mathbf{W}^*$ such that the first four sub-diagonal elements are all 1 and the rest elements are all 0 before normalizing. This kind of $\mathbf{W}$ corresponds to the pooling of $\sqrt{p}$ separate districts with similar neighboring structures in each district, see Lee and Yu (2013), that is

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}^* & 0 & 0 & \dots & 0 \\ 0 & \mathbf{W}^* & 0 & \dots & 0 \\ 0 & 0 & \mathbf{W}^* & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{W}^* \end{pmatrix}.$$

The error $\varepsilon_{i,t}$ are independently generated from $N(0, \sigma_i^2)$, where we generate each $\sigma_i$ from $U(0.5, 1.5)$.

For all scenarios, we generate data from (1.2.1) with different settings for $n$ and $p$. We apply the proposed estimation method (1.2.3) and (1.2.5) (with $d_i = \min(p, n^{10/21})$) and report the mean absolute errors:

$$\text{MAE}(i) = \frac{1}{3}\sum_{j=0}^{2} |\widehat{\lambda}_{ji} - \lambda_{ji}|, \qquad \text{MAE} = \frac{1}{p}\sum_{i=1}^{p} \text{MAE}(i).$$

We replicate each setting 500 times.

Figure 1.2 depicts two boxplots of MAE with $p$ equals to, respectively, 25 and 100. As the sample size $n$ increases from 100, 250, 500, 750 to 1000, MAE decreases for both methods.

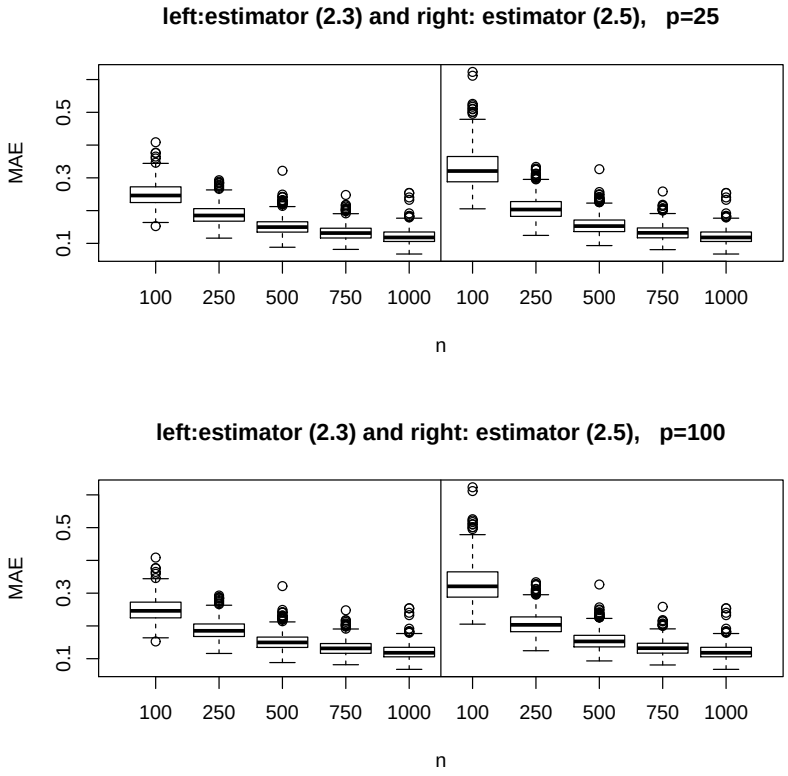


Figure 1.2: Boxplots of MAE for estimator (1.2.3) (left panels) and estimator (1.2.5) (right panels) with $p = 25$ (top panels) and 100 (bottom panels), $n = 100$, 250, 500, 750, 1000 for scenario 1.

Figure 1.3 depicts the boxplots of the MAE for the original estimator (1.2.3), the root $n$ consistent estimator (1.2.5), and the estimator (1.2.5) with the ridge penalty, where we choose the ridge tuning parameter to be $C \times \frac{p}{n}$ in order to avoid the nearly singularity problem of $\widehat{\mathbf{Z}}_i^T \widehat{\mathbf{Z}}_i$, and $C$ is chosen via cross validation. With $n = 500$, the dimension $p$ is

set at 25,49,64,81,100,169,324 and 529 respectively. The MAE for (1.2.3) remains about the same level as $p$ increases; see the panel on the left in Figure 1.3. This is in line with the asymptotic result of Theorem 4 when, for example, $s_1(p) \asymp C$, $s_0(p) \asymp p$ and $s_2(p) \asymp p$. In contrast, the MAE for estimator (1.2.5) increases sharply when $p$ increases; see the panel in the middle. This is due to the fact that $\widehat{\mathbf{Z}}_i^T \widehat{\mathbf{Z}}_i$ is nearly singular for large $p$. Adding a ridge in the estimator certainly mitigates the deterioration when $p$ increases; see the panel on the right in Figure 1.3.

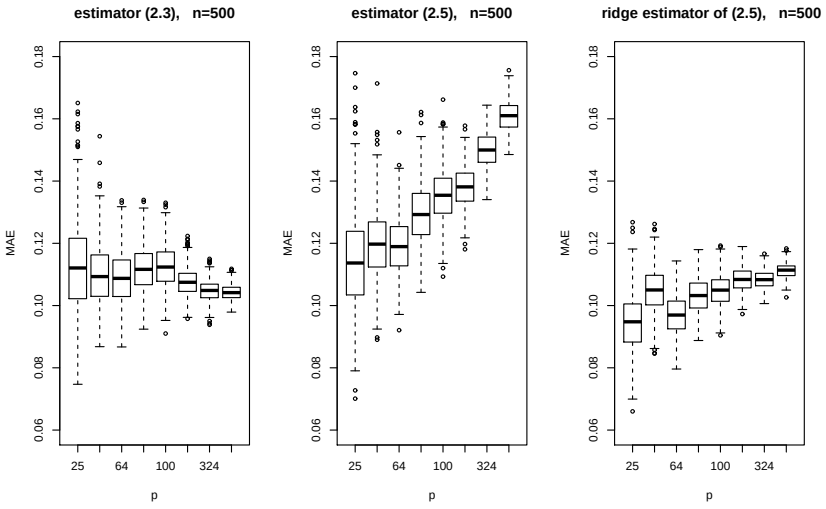


Figure 1.3: Boxplots of MAE of the original estimator (1.2.3) (the left panel), the root $n$ consistent estimator (1.2.5) (the middle panel), and the estimator (1.2.5) after adding ridge penalty (the right panel) with $n = 500$ and $p = 25, 49, 64, 81, 100, 169, 324, 529$ for scenario 1.

### 1.4.2 Scenario 2

$\lambda_{0i}$, $\lambda_{1i}$ and $\lambda_{2i}$ are generated from $U(-0.6, 0.6)$. The spatial weight matrix $\mathbf{W}$ is constructed as follows. First, we construct a $\sqrt{p} \times \sqrt{p}$ row-normalized matrix $\mathbf{W}^*$, where

$\mathbf{W}^*$ is chosen such that the first two sub-diagonal elements are all 1 and the rest elements are all 0 before normalizing. Then we treat $\mathbf{W}$ as a $\sqrt{p} \times \sqrt{p}$ block matrix and put $\mathbf{W}^*$ into the main diagonal, 2nd, 4th, 6th and etc. sub-diagonal block positions. This kind of $\mathbf{W}$ corresponds to the pooling of $\sqrt{p}$ districts (each district has $\sqrt{p}$ locations) which the evenly numbered districts are connected and the oddly numbered districts are connected but evenly numbered districts and oddly number districts are separated. Each district has similar neighboring structures. As $p$ increases, the number of the locations influencing one specific location increases in the order of $\sqrt{p}$, that is

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}^* & 0 & \mathbf{W}^* & 0 & \dots & \mathbf{W}^* \\ 0 & \mathbf{W}^* & 0 & \mathbf{W}^* & \dots & 0 \\ \mathbf{W}^* & 0 & \mathbf{W}^* & 0 & \dots & \mathbf{W}^* \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \end{pmatrix}.$$

The error $\varepsilon_{i,t}$ are independently generated from $N(0, \sigma_i^2)$, where we generate each $\sigma_i$ from $U(0.5, 1.5)$.

Figure 1.4 depicts two boxplots of MAE with $p$ equals to, respectively, 25 and 100. As the sample size $n$ increases from 100, 250, 500, 750 to 1000, MAE decreases for both methods.

Figure 1.5 depicts three boxplots as Figure 1.3. The MAE for (1.2.3) increases steadily as $p$ increases, which matches the result of Theorem 4 when, for instance, $s_1(p) \asymp \sqrt{p}$, $s_0(p) \asymp p$ and $s_2(p) \asymp p$. The MAE for (1.2.5) after adding ridge penalty is slowly increasing as well. This might be caused by the fact that, similar to A2(c), quantities in condition A4(a) is also influenced by $p$ since the number of nonzero elements in $\mathbf{w}_i$ is in the order of $\sqrt{p}$.

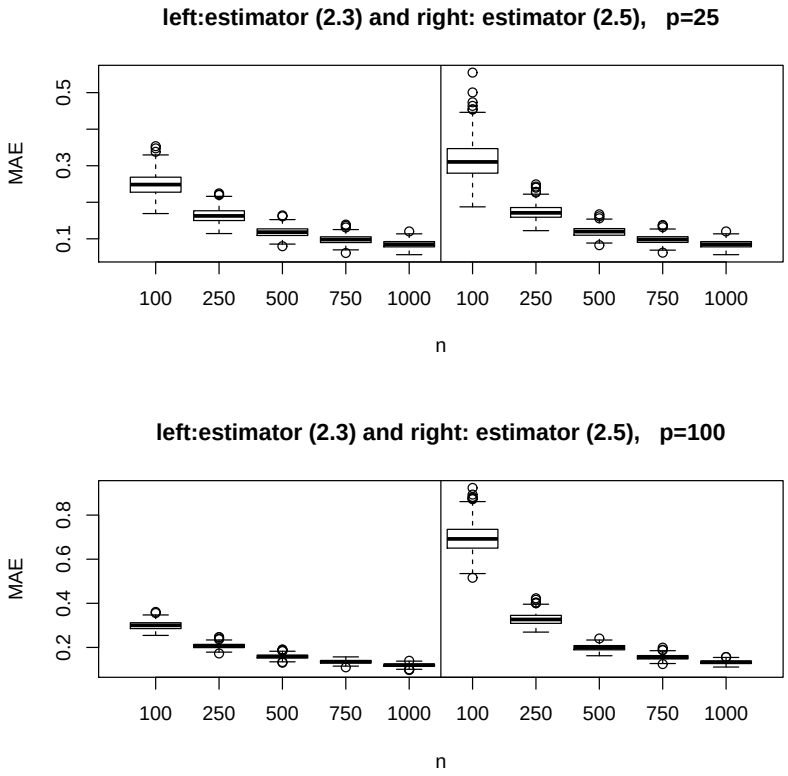


Figure 1.4: Boxplots of MAE for estimator (1.2.3) (left panels) and estimator (1.2.5) (right panels) with $p = 25$ (top panels) and 100 (bottom panels), $n = 100, 250, 500, 750, 1000$ for scenario 2.

## 1.5 Real data analysis

### 1.5.1 European Consumer Price Indices

We analyze the monthly change rates of the consumer price index (CPI) for the EU member states, over the years 2003-2010. We use the national harmonized index of consumer prices calculated by Eurostat, the statistical office of the European Union. For this data set, $n = 96$ and $p = 31$.

Figure 1.6 presents the time series plots of the monthly change rates of CPI for the 31

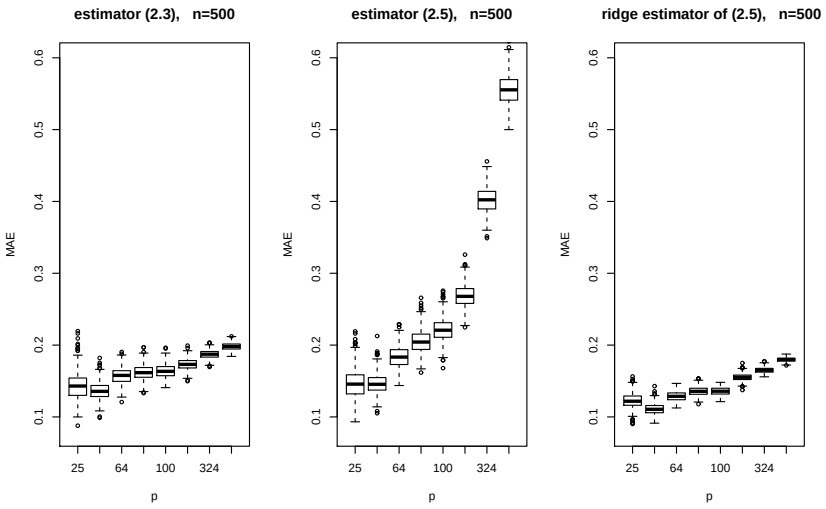


Figure 1.5: Boxplots of MAE of the original estimator (1.2.3) (the left panel), the root $n$ consistent estimator (1.2.5) (the middle panel), and the estimator (1.2.5) after adding ridge penalty (the right panel) with $n = 500$ and $p = 25, 49, 64, 81, 100, 169, 324, 529$ for scenario 2.

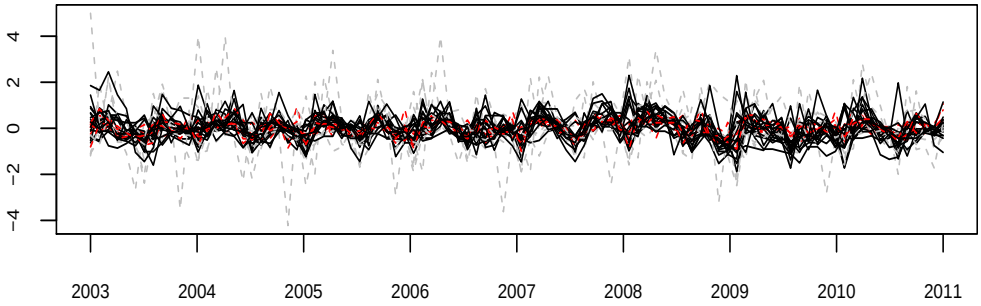


Figure 1.6: Time series plots of the monthly change rates of CPI for the 31 EU member states. Each series is subtracted by its mean value.

states. To line up the curves together, each series is centered at its mean value in Figure 1.6. There exist clearly synchronizes on the fluctuations across different states, indicating the spatial (i.e. cross-state) correlations among different states. Also noticeable is the varying degrees of the fluctuation over the different states.

Let $\mathbf{y}_t$ consist of the monthly change rates of CPI for the 31 states. We fit the proposed

spatial-temporal model (1.2.1) to this data set with the parameters estimated by (1.2.3). We take a normalized sample correlation matrix of $\mathbf{y}_t$ as the spatial weight matrix $\mathbf{W} = (w_{ij})$, i.e. we let $w_{ij}$ be the absolute value of the sample correlation between the $i$-th and $j$-th states for $i \neq j$, and $w_{ii} = 0$, and then replace $w_{ij}$ by $w_{ij}/\sum_k w_{kj}$.

Figure 1.7 presents the scatter plots of $y_{i,t}$ against, respectively, the 3 regressors in model (1.2.1), i.e. $\mathbf{w}_i^T\mathbf{y}_t$, $y_{i,t-1}$, $\mathbf{w}_i^T\mathbf{y}_{t-1}$, for four selected states Belgium, Greece, France and Iceland. We superimpose the straight line $y = \widehat{\lambda}_{ji}\, x$ in each of those 3 scatter plots with, respectively, $j = 0, 1, 2$. It is clear that the estimated slopes are very different for those 4 states. Figure 1.8 plots the true monthly change rates of the CPI for those 4 states together with the fitted values

$$\widehat{y}_{i,t} = \widehat{\lambda}_{0i}\mathbf{w}_i^T\mathbf{y}_t + \widehat{\lambda}_{1i}y_{i,t-1} + \widehat{\lambda}_{2i}\mathbf{w}_i^T\mathbf{y}_{t-1}. \tag{1.5.12}$$

Overall $\widehat{y}_{i,t}$ tracks its truth value reasonably well. Figure 1.9 shows the out-of-sample forecasting performance of our model. For the sake of comparison, predictions are made using our model and the proposed generalized Yule-Walker estimator, and using the (constant) SDPD model of Yu et al. (2008) and their Quasi-Maximum Likelihood estimator. In particular, for each location, we leave out from the sample the last six observations and we compute the (out-of-sample) forecasts with 1,2,....6 step ahead forecasting horizon; then, we compute the average prediction error over time (i.e. the mean of the 6 prediction errors). On the left panel of Figure 1.9, the two box-plots summarize the average prediction error for the 31 locations obtained with our YW estimator and the QML estimator of Yu et al. (2008), respectively. It is evident that our estimator produces unbiased predictions while the QML estimator appears to be biased. This advantage also reflects on the forecasting

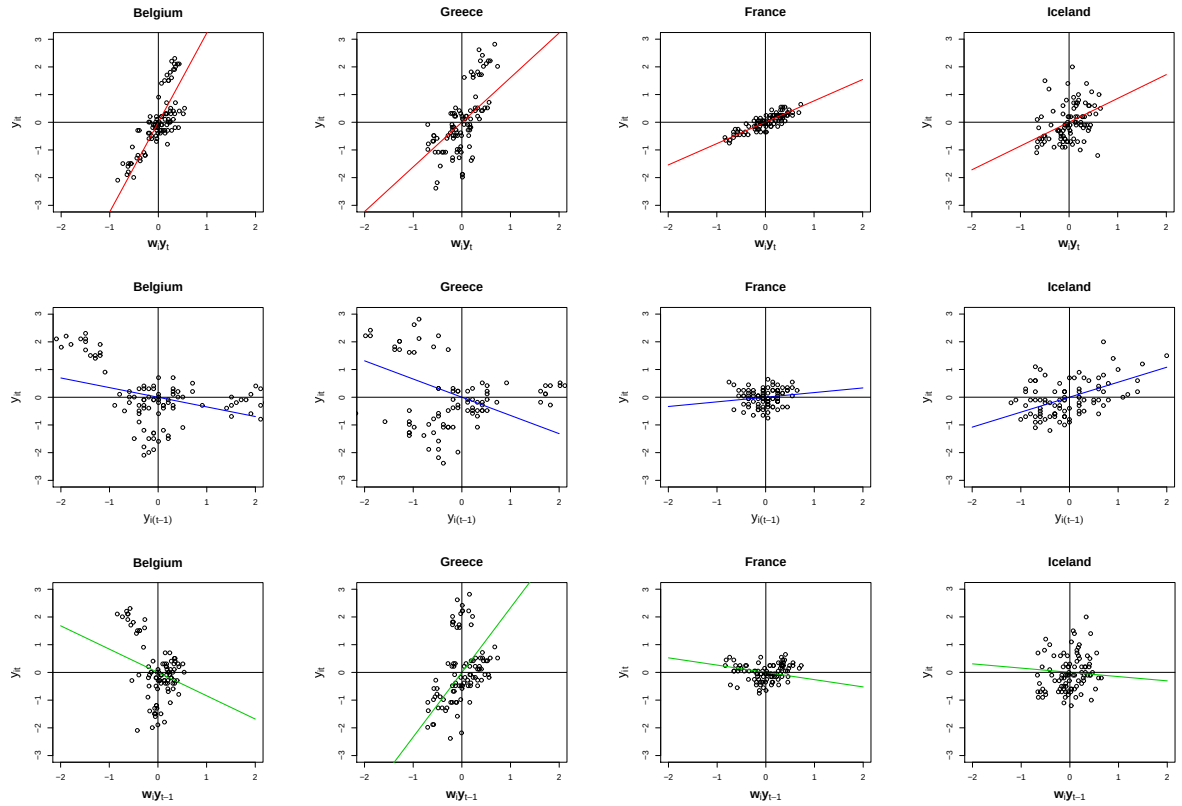


Figure 1.7: The scatter plots of $y_{i,t}$ against $\mathbf{w}_i^T\mathbf{y}_t$ (panels on the top), $y_{i,t-1}$ (panels in the middle), and $\mathbf{w}_i^T\mathbf{y}_{t-1}$ (panels on the bottom) for four selected countries Belgium, Greece, France and Iceland. The straight lines $y = \widehat{\lambda}_{ji}x$ are superimposed in the panels on the top with $j = 0$, those in the middle with $j = 1$, and those on the bottom with $j = 2$.

average square errors, reported on the right panel of Figure 1.9. In conclusion, the SDPD model of Yu et al. (2008) has a satisfying forecasting performance because several locations have similar spatial structure and for those locations a model with constant parameters is sufficient. Anyway, a marginal improvement is observed for our estimator because several locations have quite different structures and our model is able to capture this difference. Finally, it is worthwhile to notice that the variability of the two predictors appears to be the same.

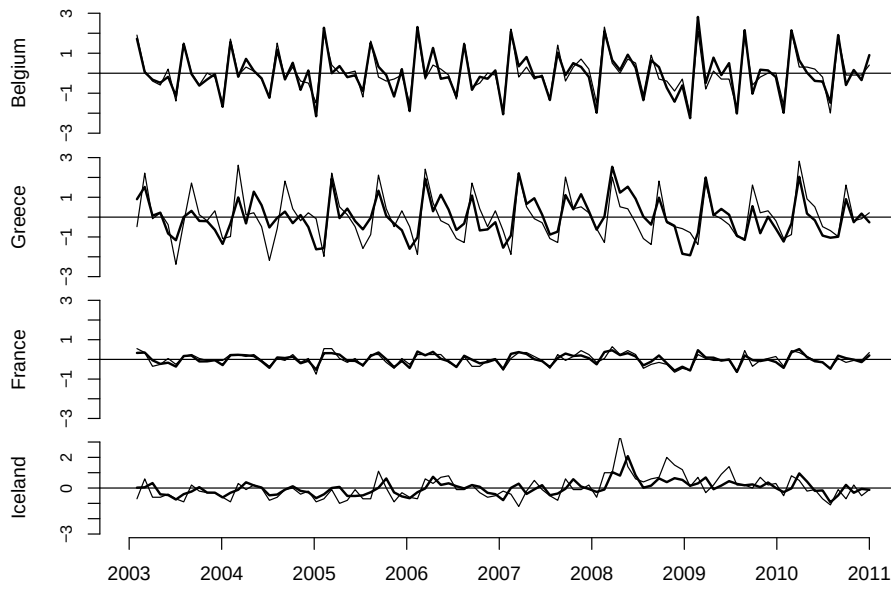


Figure 1.8: The monthly change rates of CPI (thin lines) of Belgium, Greece, France and Iceland, and their estimated values (thick lines) by model (1.2.1).

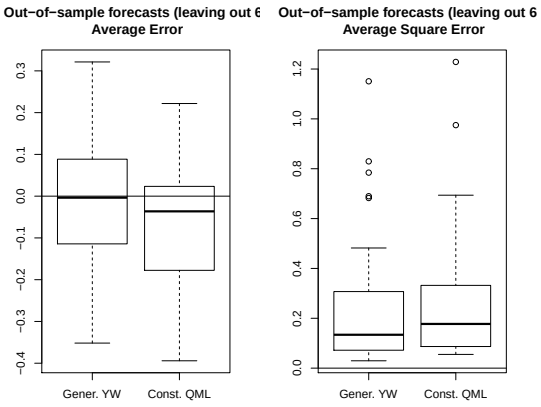


Figure 1.9: Prediction errors generated in the out-of-sample forecasting, leaving out 6 observations from the sample, using our model with the Generalized Yule-Walker estimator and using the constant SDPD model of Yu et al. (2008) with the Quasi-Maximum Likelihood estimator.

To further vindicate the necessity to use different coefficients for different states, we consider a statistical test for hypothesis

$$H_0 : \lambda_{j1} = \cdots = \lambda_{jp}, \quad j = 0,\, 1,\, 2$$

for model (1.2.1). Then the residuals resulting from the fitted model under $H_0$ will be greater than the residuals without $H_0$. However if $H_0$ is true, the difference between the two sets of residuals should not be significant. We apply a bootstrap method to test this significance. Let $\widetilde{\lambda}_0, \widetilde{\lambda}_1, \widetilde{\lambda}_2$ be the estimates under hypothesis $H_0$. Define the test statistic

$$U = \frac{1}{n}\sum_{t=1}^{n} \|\mathbf{y}_t - \widetilde{\mathbf{y}}_t\|_1, \qquad \widetilde{\mathbf{y}}_t = \widetilde{\lambda}_0 \mathbf{W}\mathbf{y}_t + \widetilde{\lambda}_1 \mathbf{y}_{t-1} + \widetilde{\lambda}_2 \mathbf{W}\mathbf{y}_{t-1}.$$

We reject $H_0$ for large values of $U$. To assess how large is large, we generate a bootstrap data from

$$\mathbf{y}_t^* = \widetilde{\lambda}_0 \mathbf{W}\mathbf{y}_t + \widetilde{\lambda}_1 \mathbf{y}_{t-1} + \widetilde{\lambda}_2 \mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t^*,$$

where $\{\boldsymbol{\varepsilon}_t^*\}$ are drawn independently from the residuals

$$\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - \widehat{\mathbf{y}}_t, \qquad t = 1, \cdots, n,$$

and $\widehat{\mathbf{y}}_t$ consists of the components defined in (1.5.12). Now the bootstrap statistic is defined as

$$U^* = \frac{1}{n}\sum_{t=1}^{n} \|\mathbf{y}_t^* - (\lambda_0^* \mathbf{W}\mathbf{y}_t + \lambda_1^* \mathbf{y}_{t-1} + \lambda_2^* \mathbf{W}\mathbf{y}_{t-1})\|_1,$$

where $(\lambda_0^*, \lambda_1^*, \lambda_2^*)$ is the estimated coefficients for the regression model

$$\mathbf{y}_t^* = \lambda_0 \mathbf{W}\mathbf{y}_t + \lambda_1 \mathbf{y}_{t-1} + \lambda_2 \mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad t = 1, \cdots, n.$$

The $P$-value for testing hypothesis $H_0$ is defined as

$$P(U^* > U | \mathbf{y}_1, \cdots, \mathbf{y}_n),$$

which is approximated by the relative frequency of the event $(U^* > U)$ in a repeated bootstrap sampling with a large number of replications. By repeating bootstrap sampling 1000 times, the estimated $P$-value is 0, exhibiting strong evidence against the null hypothesis $H_0$. Therefore the model with the equal slope parameters across different locations is inadequate for this particular data set.

### 1.5.2 Modeling mortality rates

Now we analyze the annual Italian male and female mortality rates for different ages (between 0 and 104) in the period of 1950 – 2009 based on the proposed model (1.2.1). The data were downloaded from the Human Mortality Database (see the website http://www.mortality.org/). Let $m_{i,t}$ be the log mortality rate of female or male at age $i$ and in Year $t$. Those data are plotted in Figure 1.10. Two panels on the left plot are the female and male mortality against different age in each year. More precisely the curves $\{m_{i,t},\, i = 1, \cdots, 21\}$ for $t < 1970$ are plotted in red, those for $t > 1990$ are in blue, those with $1970 \leq t \leq 1989$ are in grey. Those curves show clearly that the mortality rate decreases over the years for almost all age groups (except a few outliers at the top end). Two panels in the middle of Figure 1.10 plot the log mortality for each age group against time with the following color code: black for ages not great than 10, grey for ages between 11 and 100, and green for ages greater than 100. They indicate that the mortality for all age groups decreases over time, the most significant decreases occur at the young age groups. Furthermore, the fluctuation of the mortality rates for the top age groups reduces significantly over the years, while the mean mortality rates for those groups remain about the same. This can be seen more clearly in the two panels on the right which plot differenced log mortality rates $\{y_{i,t},\, t = 1951, \cdots, 2009\}$, using the same colour code, where $y_{i,t} = m_{i,t} - m_{i,t-1}$.

We fit the differenced log mortality data with model (1.2.1) with the parameters estimated by (1.2.5) and $d_i = 20$. Note that now $p = 104$ and $n = 59$. Let the off-diagonal elements of the spatial weight matrix $\mathbf{W}$ be

$$w_{ij} = \frac{1}{1 + |i - j|}, \qquad 1 \leq i < j \leq 104.$$

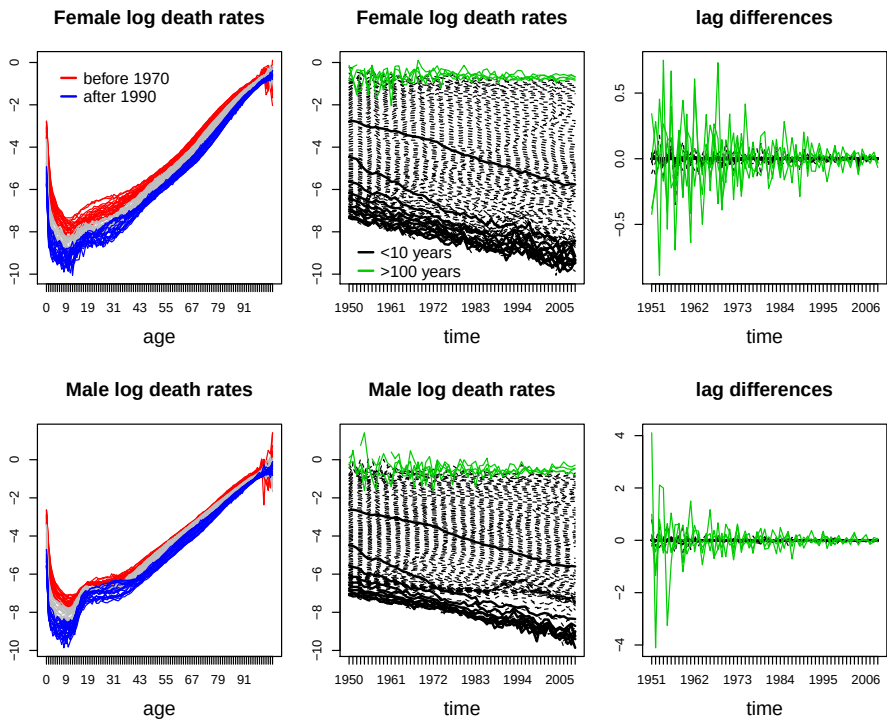


Figure 1.10: Log mortality rates of Italian female (3 top panels) and male (3 bottom panels) are plotted against age from each year in 1950-2009 (2 left panels), against year for each age group between 0 and 104 (2 middle panels). Differenced log mortality rates are plotted against year for each age in 2 right panels.

We then replace $w_{ij}$ by $w_{ij}/\sum_i w_{ij}$. Moreover, we can also fix a threshold $\tau$ and set to zero all the elements of matrix $\mathbf{W}$ such that $|x - w| > \tau$ (for simplicity, we fix $\tau = 5$ in this application, but the results are substantially invariant for different values of $\tau$).

The results of the estimation are shown in table 1.1, for a selection of cohorts of different ages. Figure 1.11 shows the fitted series for ages $i = 60, 80, 100$.

| age | $\widehat{\lambda}_{0i}$ | $\widehat{\lambda}_{1i}$ | $\widehat{\lambda}_{2i}$ | age | $\widehat{\lambda}_{0i}$ | $\widehat{\lambda}_{1i}$ | $\widehat{\lambda}_{2i}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.41 | -0.52 | 0.06 | 55 | 0.19 | -0.88 | 0.28 |
| 10 | 0.20 | -0.42 | 0.05 | 60 | -0.09 | -0.72 | 0.01 |
| 15 | 0.44 | -0.65 | 0.18 | 65 | 0.22 | -0.63 | 0.21 |
| 20 | 0.64 | -0.78 | 0.40 | 70 | 0.21 | -0.69 | 0.08 |
| 25 | -0.04 | -0.43 | 0.03 | 75 | 0.33 | -0.59 | 0.22 |
| 30 | 0.78 | -0.80 | 0.55 | 80 | 0.33 | -0.89 | 0.27 |
| 35 | 0.11 | -0.55 | 0.29 | 85 | 0.37 | -0.76 | 0.18 |
| 40 | -0.04 | -0.66 | -0.01 | 90 | 0.29 | -0.62 | 0.16 |
| 45 | 0.29 | -0.46 | 0.12 | 95 | 0.27 | -0.77 | 0.26 |
| 50 | -0.10 | -0.45 | -0.05 | 100 | 0.44 | -0.69 | -0.03 |

Table 1.1: Estimated coefficients for a selection of cohorts of different ages. The left column is the estimated pure spatial coefficients $\widehat{\lambda}_{0i}$; The middle column is the estimated pure dynamic coefficient $\widehat{\lambda}_{1i}$; The right column is the estimated spatial-dynamic coefficients $\widehat{\lambda}_{2i}$.

## 1.6 Final remark

We propose in this paper a generalized Yule-Walker estimation method for spatio-temporal models with diagonal coefficients. The setting enlarges the capacity of the popular spatial dynamic panel data models. Both the asymptotic results and numerical illustration show that the proposed estimation method works well, although the number of the estimation equations utilized should be of the order $o(\sqrt{n})$.

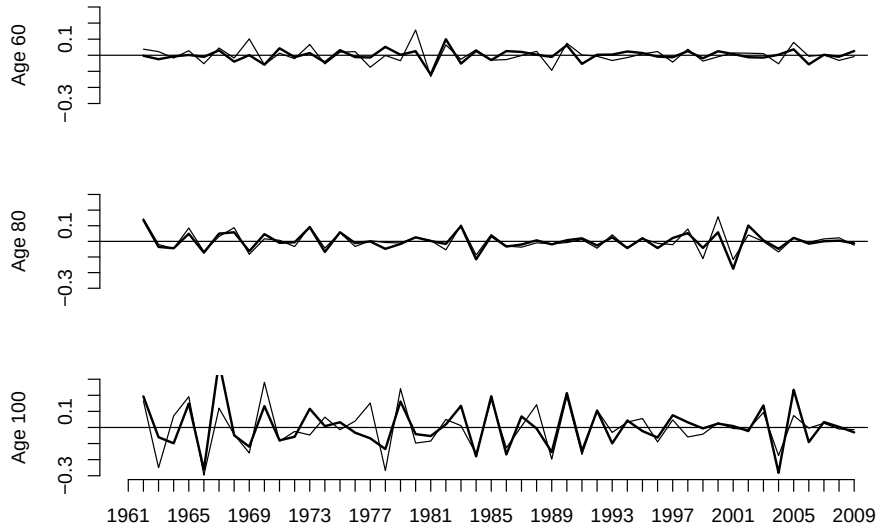


Figure 1.11: Observed time series (thin line) and fitted time series (bold line), for female mortality rate for ages $i = 60, 80, 100$.

## 1.7 Appendix: Proofs

We present the proofs for Theorems 2, Corollary 1 and Theorem 4 in this appendix. The proofs for Theorem 1 and 3 are similar and simpler than that of Theorem 2, and they are therefore omitted. We also present a lemma (i.e. Lemma 1) at the end of this appendix, which shows that condition A2 is implied by conditions A1 and B1 – B3; see Remark 1. We use $C$ to denote a generic positive constant, which may be different at different places.

**Proof of Theorem 2.** We first prove (i) of Theorem 2. We only need to prove the assertions (1) and (2) below, as then the required conclusion follows from (1) and (2) immediately.

(1)

$$\sqrt{n}\mathbf{U}_i^{-\frac{1}{2}} \begin{pmatrix} \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \end{pmatrix} \xrightarrow{d} N(0,\mathbf{I}_3).$$

(2) $\mathbf{V}_i(\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i)^{-1} \xrightarrow{P} \mathbf{I}_3$.

To prove (1), it suffices to show that for any nonzero vector $\mathbf{a} = (a_1, a_2, a_3)^T$, the linear combination

$$\mathbf{a}^T \begin{pmatrix} \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ \frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \end{pmatrix}$$

is asymptotic normal.

Let us take out the dominant term in $\frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1}$ first.

$$\begin{aligned} &\frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ =&\left[\frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t) - \mathrm{E}[\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)]\right]\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} + \mathrm{E}[\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)]\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} \\ =&\left[\frac{1}{n}\sum_{t=1}^n \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t) - \mathbf{w}_i^T\mathbf{\Sigma}_1\right]\frac{1}{n}\sum_{t=1}^n \varepsilon_{i,t}\mathbf{y}_{t-1} + \frac{1}{n}\sum_{t=1}^n \mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t} \\ =&E_1 + E_2. \end{aligned} \tag{1.7.13}$$

For term $E_1$ and $k = 1, 2, \cdots, p$, by Proposition 2.5 of Fan and Yao (2003), we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n}\sum_{t=1}^{n}(\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t - \mathbf{e}_k^T\mathbf{\Sigma}_1^T\mathbf{w}_i)\right]^2 \\
=&\frac{1}{n^2}\sum_{t=1}^{n}\mathrm{Var}(\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t) + \frac{1}{n^2}\sum_{t\neq s}\mathrm{Cov}(\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t, \mathbf{e}_k^T\mathbf{y}_{s-1}\mathbf{w}_i^T\mathbf{y}_s) \\
\leq&\frac{C}{n} + \frac{1}{n^2}\sum_{t\neq s}8\alpha(|t-s|)^{\frac{\gamma}{4+\gamma}}\left[\mathrm{E}|\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t|^{2+\frac{\gamma}{2}}\right]^{\frac{2}{4+\gamma}}\left[\mathrm{E}|\mathbf{e}_k^T\mathbf{y}_{s-1}\mathbf{w}_i^T\mathbf{y}_s|^{2+\frac{\gamma}{2}}\right]^{\frac{2}{4+\gamma}} \\
\leq&\frac{C}{n} + \frac{C}{n^2}\sum_{t\neq s}\alpha(|t-s|)^{\frac{\gamma}{4+\gamma}} \leq \frac{C}{n} + \frac{C}{n}\sum_{j=1}^{n}\alpha(j)^{\frac{\gamma}{4+\gamma}} = O(\frac{1}{n}),
\end{aligned} \tag{1.7.14}$$

where $C$ is independent of $p$. Then it holds that

$$\frac{1}{n}\sum_{t=1}^{n}(\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t - \mathbf{e}_k^T\mathbf{\Sigma}_1^T\mathbf{w}_i) = O_p(\frac{1}{\sqrt{n}}).$$

Therefore

$$\left\|\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t - \mathbf{\Sigma}_1^T\mathbf{w}_i\right\|_2 = \sqrt{\sum_{k=1}^{p}\left[\frac{1}{n}\sum_{t=1}^{n}(\mathbf{e}_k^T\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t - \mathbf{e}_k^T\mathbf{\Sigma}_1^T\mathbf{w}_i)\right]^2} = O_p(\sqrt{\frac{p}{n}}).$$

Similarly,

$$\left\|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1}\right\|_2 = O_p(\sqrt{\frac{p}{n}}).$$

Since $E_1 \leq \left\|\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}\mathbf{w}_i^T\mathbf{y}_t - \mathbf{\Sigma}_1^T\mathbf{w}_i\right\|_2 \left\|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1}\right\|_2$, it holds that $E_1 = O_p(\frac{p}{n})$. Similar to (1.7.14), we have $\mathrm{Var}(\sqrt{n}E_2) = O(1)$. Given $\frac{p}{\sqrt{n}} = o(1)$, it holds that $\sqrt{n}E_1 = o_p(1)$. Hence if $p = o(\sqrt{n})$,

$$\sqrt{n}\times\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t} + o_p(1).$$

Similarly, given $p = o(\sqrt{n})$, we have

$$\sqrt{n}\times\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t} + o_p(1),$$

$$\sqrt{n}\times\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t} + o_p(1).$$

Now it suffices to prove

$$S_{n,p} \equiv a_1 \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} + a_2 \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{e}_i^T \mathbf{\Sigma}_0 \mathbf{y}_{t-1} \varepsilon_{i,t} + a_3 \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{w}_i^T \mathbf{\Sigma}_0 \mathbf{y}_{t-1} \varepsilon_{i,t}$$

is asymptotic normal.

Note that it holds that

$$\mathrm{E}|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}|^{2+\frac{\gamma}{2}} \leq [\mathrm{E}|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1})^{4+\gamma}|^{\frac{1}{2}} [\mathrm{E}|\varepsilon_{i,t}|^{4+\gamma}]^{\frac{1}{2}} < \infty.$$

Now we calculate the variance of $S_{n,p}$. It holds that

$$\begin{aligned} &\mathrm{Var}\left( \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} \right) \\ =&\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(0) \mathbf{\Sigma}_1^T \mathbf{w}_i + \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \mathbf{w}_i^T \mathbf{\Sigma}_1 \left[ \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(j) + \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}^T(j) \right] \mathbf{\Sigma}_1^T \mathbf{w}_i, \end{aligned} \tag{1.7.15}$$

and it follows from $\sum_{j=1}^{n} \alpha(j)^{\frac{\gamma}{4+\gamma}} < \infty$ that

$$\begin{aligned} &\sup_p \sum_{j=1}^{\infty} |\mathbf{w}_i^T \mathbf{\Sigma}_1 \left[ \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(j) + \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}^T(j) \right] \mathbf{\Sigma}_1^T \mathbf{w}_i| \\ \leq& C \sup_p \sum_{j=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} \left\{ \mathrm{E}|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1}|^{4+\gamma} \right\}^{\frac{2}{4+\gamma}} \left\{ \mathrm{E}|\varepsilon_{i,t}|^{4+\gamma} \right\}^{\frac{2}{4+\gamma}} < \infty. \end{aligned}$$

Similarly,

$$\begin{aligned} &\mathrm{Cov}\left( \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}, \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{e}_i^T \mathbf{\Sigma}_0 \mathbf{y}_{t-1} \varepsilon_{i,t} \right) \\ =&\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(0) \mathbf{\Sigma}_0^T \mathbf{e}_i + \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \mathbf{w}_i^T \mathbf{\Sigma}_1 \left[ \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(j) + \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}^T(j) \right] \mathbf{\Sigma}_0 \mathbf{e}_i, \end{aligned}$$

and $\sup_p \sum_{j=1}^{\infty} |\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(j) \mathbf{\Sigma}_0 \mathbf{e}_i| < \infty$. Calculating all the variance and covariance and summing up them, it follows from dominate convergence theorem that

$$\mathrm{Var}\left( \frac{S_{n,p}}{\sqrt{\mathbf{a}^T \mathbf{U}_i \mathbf{a}}} \right) \to 1.$$

To prove the asymptotic normality of $S_{n,p}$, we employ the small-block and large-block arguments. We partition the set $\{1, 2, \cdots, n\}$ into $2k_n + 1$ subsets with large blocks of size

$l_n$, small blocks of size $s_n$ and the last remaining set of size $n - k_n(l_n + s_n)$. Put

$$l_n = [\sqrt{n}/\log n], \quad s_n = [\sqrt{n}\log n]^x, \quad k_n = [n/(l_n + s_n)],$$

where $\frac{\gamma}{4+\gamma} \leq x < 1$. Hence

$$l_n/\sqrt{n} \to 0, \quad s_n/l_n \to 0, \quad k_n = O(\sqrt{n}\log n).$$

Note that $l_n/\sqrt{n} \to 0$ is important when we derive the Lindeberg condition of the truncated partial sum $T^L_{n,p}$ defined in (1.7.16).

Since $\sum_{j=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} < \infty$, we have $\alpha(s_n) = o(s_n^{-\frac{4+\gamma}{\gamma}})$. It then holds that

$$k_n\alpha(s_n) = o(k_n/s_n^{\frac{4+\gamma}{\gamma}}) = o(\sqrt{n}\log n/[\sqrt{n}\log n]^{x\frac{4+\gamma}{\gamma}}) = o(1).$$

Then we can partition $S_{n,p}$ in the following way

$$\begin{aligned} S_{n,p} =& a_1\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\xi_j^{(1)} + a_2\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\xi_j^{(2)} + a_3\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\xi_j^{(3)} \\ &+ a_1\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\eta_j^{(1)} + a_2\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\eta_j^{(2)} + a_3\frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\eta_j^{(3)} \\ &+ a_1\frac{1}{\sqrt{n}}\zeta^{(1)} + a_2\frac{1}{\sqrt{n}}\zeta^{(2)} + a_3\frac{1}{\sqrt{n}}\zeta^{(3)}, \end{aligned}$$

where

$$\xi_j^{(1)} = \sum_{t=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}, \quad \eta_j^{(1)} = \sum_{t=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} \mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t},$$

$$\xi_j^{(2)} = \sum_{t=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t}, \quad \eta_j^{(2)} = \sum_{t=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t},$$

$$\xi_j^{(3)} = \sum_{t=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t}, \quad \eta_j^{(3)} = \sum_{t=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} \mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t},$$

$$\zeta^{(1)} = \sum_{k_n(l_n+s_n)+1}^{n} \mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}, \quad \zeta^{(2)} = \sum_{k_n(l_n+s_n)+1}^{n} \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t}, \quad \zeta^{(3)} = \sum_{k_n(l_n+s_n)+1}^{n} \mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{y}_{t-1}\varepsilon_{i,t}.$$

Note that $\alpha(n) = o(n^{-\frac{(2+\gamma/2)2}{2(2+\gamma/2-2)}})$ and $k_n s_n/n \to 0$, $(l_n + s_n)/n \to 0$, by applying proposition 2.7 of Fan and Yao (2003), it holds that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \eta_j^{(l)} = o_p(1), \quad \text{and} \quad \frac{1}{\sqrt{n}} \zeta^{(l)} = o_p(1), \quad l = 1, 2, 3.$$

Therefore

$$S_{n,p} = a_1 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)} + a_2 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(2)} + a_3 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(3)} + o_p(1) \equiv T_{n,p} + o_p(1).$$

We calculate the variance of $T_{n,p}$. Similar to (1.7.15), it holds that

$$\text{Var}\left( a_1 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)} \right) = a_1^2 \frac{k_n}{n} \text{Var}\left( \xi_1^{(1)} \right) \{1 + o(1)\} = a_1^2 \frac{k_n}{n} \text{Var}\left( \sum_{t=1}^{l_n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} \right) \{1 + o(1)\}$$

$$= a_1^2 \frac{k_n l_n}{n} \left[ \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(0) \mathbf{\Sigma}_1^T \mathbf{w}_i + \sum_{j=1}^{l_n - 1} \left( 1 - \frac{j}{l_n} \right) \mathbf{w}_i^T \mathbf{\Sigma}_1 \left[ \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}(j) + \mathbf{\Sigma}_{\mathbf{y},\varepsilon_i}^T(j) \right] \mathbf{\Sigma}_1^T \mathbf{w}_i \right] \{1 + o(1)\}.$$

Calculating all the variance and covariance and summing up them, by dominated convergence theorem and $\frac{k_n l_n}{n} \to 1$, it holds that

$$\text{Var}\left( \frac{T_{n,p}}{\sqrt{\mathbf{a}^T \mathbf{U}_i \mathbf{a}}} \right) \to 1.$$

Now it suffices to prove the asymptotic normality of $T_{n,p}$. We partition $T_{n,p}$ into two parts via truncation. Specifically, we define

$$\xi_j^{(1)L} = \sum_{t=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}| \le L\}},$$

and

$$\xi_j^{(1)R} = \sum_{t=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}| > L\}}.$$

Similarly, we have $\xi_j^{(2)L}, \xi_j^{(2)R}$ and $\xi_j^{(3)L}, \xi_j^{(3)R}$. Then

$$\begin{aligned} T_{n,p} =& \left( a_1 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)L} + a_2 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(2)L} + a_3 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(3)L} \right) \\ &+ \left( a_1 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)R} + a_2 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(2)R} + a_3 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(3)R} \right) \\ \equiv& T_{n,p}^L + T_{n,p}^R. \end{aligned} \tag{1.7.16}$$

Similar to computing the $\mathrm{Var}(T_{n,p})$, it holds that

$$\begin{aligned} &\mathrm{Var}(T_{n,p}^L) = a_1^2 \mathrm{Var}\left( \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)L} \right) + \Omega^L = a_1^2 \frac{k_n}{n} \mathrm{Var}\left( \xi_1^{(1)L} \right) \{1 + o(1)\} + \Omega^L \\ =& a_1^2 \frac{k_n}{n} \mathrm{Var}\left( \sum_{t=1}^{l_n} \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}| \le L\}} \right) \{1 + o(1)\} + \Omega^L \\ =& a_1^2 \frac{k_n l_n}{n} \Bigg[ \mathrm{Var}\left( \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}| \le L\}} \right) \\ &+ 2 \sum_{j=1}^{l_n - 1} \left( 1 - \frac{j}{l_n} \right) \mathrm{Cov}\left( \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1+j} \varepsilon_{i,t+j} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1+j} \varepsilon_{i,t+j}| \le L\}}, \mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t} I_{\{|\mathbf{w}_i^T \mathbf{\Sigma}_1 \mathbf{y}_{t-1} \varepsilon_{i,t}| \le L\}} \right) \Bigg] \\ &\{1 + o(1)\} + \Omega^L, \end{aligned}$$

where $\Omega^L$ is the sum of all the rest variance and covariance except $\mathrm{Var}\left( a_1 \frac{1}{\sqrt{n}} \sum_{j=1}^{k_n} \xi_j^{(1)L} \right)$.

Therefore

$$\mathrm{Var}\left( \frac{\mathrm{Var}(T_{n,p}^L)}{\sigma_L^2} \right) \to 1,$$

where we denote $\sigma_L^2$ as the asymptotic variance of $T_{n,p}^L$. Similarly, we have

$$\begin{aligned}
&\mathrm{Var}(T_{n,p}^R) \\
=&a_1^2\frac{k_n l_n}{n}\Bigg[\mathrm{Var}\left(\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}I_{\{|\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}|>L\}}\right) \\
&+2\sum_{j=1}^{l_n-1}\left(1-\frac{j}{l_n}\right)\mathrm{Cov}\left(\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1+j}\varepsilon_{i,t+j}I_{\{|\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1+j}\varepsilon_{i,t+j}|>L\}},\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}I_{\{|\mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{y}_{t-1}\varepsilon_{i,t}|>L\}}\right)\Bigg] \\
&\{1+o(1)\}+\Omega^R.
\end{aligned}$$

Define

$$M_{n,p}=\left|\mathrm{E}\exp\left(\frac{itT_{n,p}}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right)-\exp\left(-\frac{t^2}{2}\right)\right|,$$

where $i=\sqrt{-1}$ now. We bound $M_{n,p}$ as follows

$$\begin{aligned}
M_{n,p}\leq&\mathrm{E}\left|\exp\left(\frac{itT_{n,p}^L}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right)\left[\exp\left(\frac{itT_{n,p}^R}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right)-1\right]\right| \\
&+\left|\mathrm{E}\exp\left(\frac{itT_{n,p}^L}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right)-\prod_{j=1}^{k_n}\mathrm{E}\exp\left[\frac{it\left(a_1\frac{1}{\sqrt{n}}\xi_j^{(1)L}+a_2\frac{1}{\sqrt{n}}\xi_j^{(2)L}+a_3\frac{1}{\sqrt{n}}\xi_j^{(3)L}\right)}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right]\right| \\
&+\left|\prod_{j=1}^{k_n}\mathrm{E}\exp\left[\frac{it\left(a_1\frac{1}{\sqrt{n}}\xi_j^{(1)L}+a_2\frac{1}{\sqrt{n}}\xi_j^{(2)L}+a_3\frac{1}{\sqrt{n}}\xi_j^{(3)L}\right)}{\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}}\right]-\exp\left(-\frac{t^2}{2}\frac{\sigma_L^2}{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}\right)\right| \\
&+\left|\exp\left(-\frac{t^2}{2}\frac{\sigma_L^2}{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}\right)-\exp\left(-\frac{t^2}{2}\right)\right|.
\end{aligned}$$

Following the same arguments as part 2.7.7 of Fan and Yao (2003), for any $\epsilon>0$, it holds that $M_{n,p}<\epsilon$ as $n,p\to\infty$. Hence

$$\sqrt{n}\times\mathbf{a}^T\begin{pmatrix}\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\end{pmatrix}/\sqrt{\mathbf{a}^T\mathbf{U}_i\mathbf{a}}\xrightarrow{d}N(0,1).$$

Substituting $\mathbf{a}$ by $(\mathbf{U}_i^{-\frac{1}{2}})^T\mathbf{a}$, it holds that

$$\mathbf{a}^T\left\{\sqrt{n}\mathbf{U}_i^{-\frac{1}{2}}\begin{pmatrix}\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\end{pmatrix}\right\}\xrightarrow{d}\mathbf{a}^T N(0,\mathbf{I}_3),$$

which leads to the fact that

$$\sqrt{n}\mathbf{U}_i^{-\frac{1}{2}}\begin{pmatrix}\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\\ \frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^n\varepsilon_{i,t}\mathbf{y}_{t-1}\end{pmatrix}\xrightarrow{d}N(0,\mathbf{I}_3).$$

To prove (2), let us look at the $(1,1)$-th element of $\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i$. We have

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t)\\ &=\left(\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)-\mathbf{w}_i^T\boldsymbol{\Sigma}_1\right)\left(\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t)-\boldsymbol{\Sigma}_1^T\mathbf{w}_i\right)\\ &\quad+2\mathbf{w}_i^T\boldsymbol{\Sigma}_1\left(\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t)-\boldsymbol{\Sigma}_1^T\mathbf{w}_i\right)+\mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^T\mathbf{w}_i.\end{aligned}\tag{1.7.17}$$

Using the same arguments as (1.7.14), the first term is $O_p(\frac{p}{n})$ and the second term is $O_p(\frac{1}{\sqrt{n}})$. Hence given $p=o(n)$, it holds that

$$\frac{\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^n\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t)}{\mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^T\mathbf{w}_i}\to 1.$$

Applying the same arguments to the other elements of $\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i$, it holds that

$$\mathbf{V}_i(\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i)^{-1}\xrightarrow{P}\mathbf{I}_3.$$

To prove (ii) in Theorem 2, the required asymptotic result follows from (1.7.13) and (1.7.17) immediately when $p=o(n)$ and $\sqrt{n}=O(p)$. The proof is completed. □

**Proof of Corollary 1.** By Theorem 2, it holds that

$$\left\| \begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\|_1 = \begin{cases} O_p(\frac{1}{\sqrt{n}}) & \text{if } \frac{p}{\sqrt{n}} = O(1), \\ O_p(\frac{p}{n}) & \text{if } \frac{p}{\sqrt{n}} \to \infty \text{ and } \frac{p}{n} = o(1). \end{cases}$$

for all $i$. The required asymptotic result follows from the above result directly. □

**Proof of Theorem 4.** Let us look at term $E_1$ and $E_2$ in (1.7.13) first under the new condition (A5). Similar to the proof of (1.7.14), it holds that

$$E_1 = O_p(\frac{ps_1^{3/4}(p)}{n}), \quad E_2 = O_p(\frac{s_0^{1/4}(p)}{\sqrt{n}}).$$

Hence

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = O_p(\frac{ps_1^{3/4}(p)}{n} + \frac{s_0^{1/4}(p)}{\sqrt{n}}).$$

Similarly, we have

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = O_p(\frac{ps_1^{3/4}(p)}{n} + \frac{s_0^{1/4}(p)}{\sqrt{n}}),$$

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^{n}\varepsilon_{i,t}\mathbf{y}_{t-1} = O_p(\frac{ps_1^{3/4}(p)}{n} + \frac{s_0^{1/4}(p)}{\sqrt{n}}).$$

For the first diagonal element of $\widehat{\mathbf{X}}_i^T\widehat{\mathbf{X}}_i$, it follows from considering the three terms in (1.7.17) separately that

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_t)\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_t) = O_p(\frac{ps_1(p)}{n} + \frac{s_0^{1/4}(p)s_1^{1/4}(p)}{\sqrt{n}}) + \mathbf{w}_i^T\mathbf{\Sigma}_1\mathbf{\Sigma}_1^T\mathbf{w}_i.$$

Similarly,

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T y_{i,t-1}\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}y_{i,t-1} = O_p(\frac{ps_1(p)}{n} + \frac{s_0^{1/4}(p)s_1^{1/4}(p)}{\sqrt{n}}) + \mathbf{e}_i^T\mathbf{\Sigma}_0\mathbf{\Sigma}_0^T\mathbf{e}_i,$$

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_{t-1}^T(\mathbf{w}_i^T\mathbf{y}_{t-1})\frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_{t-1}(\mathbf{w}_i^T\mathbf{y}_{t-1}) = O_p(\frac{ps_1(p)}{n} + \frac{s_0^{1/4}(p)s_1^{1/4}(p)}{\sqrt{n}}) + \mathbf{w}_i^T\mathbf{\Sigma}_0\mathbf{\Sigma}_0^T\mathbf{w}_i.$$

Given $\frac{ps_1(p)}{s_2(p)} = o(n)$ and $\frac{s_0^{1/2}(p)}{ps_1^{1/2}(p)s_2(p)} = O(1)$, we have

$$\frac{ps_1(p)}{n} = o(s_2(p)), \quad \frac{s_0^{1/4}(p)s_1^{1/4}(p)}{\sqrt{n}} = o(s_2(p)).$$

Divide both the numerator and denominator of estimator (2.2.23) by $s_2(p)$, it holds that

$$\left\| \begin{pmatrix} \widehat{\lambda}_{0i} \\ \widehat{\lambda}_{1i} \\ \widehat{\lambda}_{2i} \end{pmatrix} - \begin{pmatrix} \lambda_{0i} \\ \lambda_{1i} \\ \lambda_{2i} \end{pmatrix} \right\|_2 = O_p\Big(\frac{ps_1^{3/4}(p)}{ns_2(p)} + \frac{s_0^{1/4}(p)}{\sqrt{n}s_2(p)}\Big).$$

The required result then follows directly. □

**Lemma 1** *Under conditions A1 and B1 – B3, condition A2 holds with* $\gamma = 4$.

**Proof.** It is apparent that part (a) of A2 is satisfied under A1 and B1 – B3. $\mathbf{y}_t$ is strictly stationary because $\varepsilon_{i,t}$ are $i.i.d$ across $i$ and $t$ and condition B3. Since the density function of $\varepsilon_{i,t}$ exists, $\alpha(n)$ decays exponentially fast, see Pham and Tran (1985). Therefore $\sum_{j=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} < \infty$. Now we prove A2(c) when $\gamma = 4$.

We present a more general result first: for any $p \times 1$ vector $\mathbf{a}$ satisfying $\sup_p \|\mathbf{a}\|_1 < \infty$, it holds that

$$\sup_p \mathrm{E}\left|\mathbf{a}^T\mathbf{y}_t\right|^8 < \infty.$$

Note that

$$\mathbf{y}_t = \sum_{h=0}^{\infty} \mathbf{A}^h\mathbf{S}^{-1}(\boldsymbol{\lambda}_0)\boldsymbol{\varepsilon}_{t-h} \equiv \sum_{h=0}^{\infty} \mathbf{B}_h\boldsymbol{\varepsilon}_{t-h}.$$

Then

$$
\begin{aligned}
&\mathrm{E}\left|\mathbf{a}^T\mathbf{y}_t\right|^8 = \mathrm{E}\left|\sum_{h=0}^{\infty}\mathbf{a}^T\mathbf{B}_h\boldsymbol{\varepsilon}_{t-h}\right|^8 \equiv \mathrm{E}\left|\sum_{h=0}^{\infty}\mathbf{b}_h^T\boldsymbol{\varepsilon}_{t-h}\right|^8 \\
=&\mathrm{E}\left|\sum_{h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8=0}^{\infty}(\boldsymbol{\varepsilon}_{t-h_1}^T\mathbf{b}_{h_1}\mathbf{b}_{h_2}^T\boldsymbol{\varepsilon}_{t-h_2})(\boldsymbol{\varepsilon}_{t-h_3}^T\mathbf{b}_{h_3}\mathbf{b}_{h_4}^T\boldsymbol{\varepsilon}_{t-h_4})(\boldsymbol{\varepsilon}_{t-h_5}^T\mathbf{b}_{h_5}\mathbf{b}_{h_6}^T\boldsymbol{\varepsilon}_{t-h_6})(\boldsymbol{\varepsilon}_{t-h_7}^T\mathbf{b}_{h_7}\mathbf{b}_{h_8}^T\boldsymbol{\varepsilon}_{t-h_8})\right| \\
=&\mathrm{E}\Bigg|\sum_{h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8=0}^{\infty}\Big(\sum_{i_1,j_1=1}^{p}[\mathbf{b}_{h_1}\mathbf{b}_{h_2}^T]_{i_1j_1}\varepsilon_{i_1,t-h_1}\varepsilon_{j_1,t-h_2}\Big)\Big(\sum_{i_2,j_2=1}^{p}[\mathbf{b}_{h_3}\mathbf{b}_{h_4}^T]_{i_2j_2}\varepsilon_{i_2,t-h_3}\varepsilon_{j_2,t-h_4}\Big) \\
&\times\Big(\sum_{i_3,j_3=1}^{p}[\mathbf{b}_{h_5}\mathbf{b}_{h_6}^T]_{i_3j_3}\varepsilon_{i_3,t-h_5}\varepsilon_{j_3,t-h_6}\Big)\Big(\sum_{i_4,j_4=1}^{p}[\mathbf{b}_{h_7}\mathbf{b}_{h_8}^T]_{i_4j_4}\varepsilon_{i_4,t-h_7}\varepsilon_{j_4,t-h_8}\Big)\Bigg| \\
=&\mathrm{E}\Bigg|\sum_{h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8=0}^{\infty}\sum_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4=1}^{p}[\mathbf{b}_{h_1}\mathbf{b}_{h_2}^T]_{i_1j_1}[\mathbf{b}_{h_3}\mathbf{b}_{h_4}^T]_{i_2j_2}[\mathbf{b}_{h_5}\mathbf{b}_{h_6}^T]_{i_3j_3}[\mathbf{b}_{h_7}\mathbf{b}_{h_8}^T]_{i_4j_4} \\
&\times\varepsilon_{i_1,t-h_1}\varepsilon_{j_1,t-h_2}\varepsilon_{i_2,t-h_3}\varepsilon_{j_2,t-h_4}\varepsilon_{i_3,t-h_5}\varepsilon_{j_3,t-h_6}\varepsilon_{i_4,t-h_7}\varepsilon_{j_4,t-h_8}\Bigg| \\
\leq&\sum_{h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8=0}^{\infty}\sum_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4=1}^{p}\Big|[\mathbf{b}_{h_1}\mathbf{b}_{h_2}^T]_{i_1j_1}[\mathbf{b}_{h_3}\mathbf{b}_{h_4}^T]_{i_2j_2}[\mathbf{b}_{h_5}\mathbf{b}_{h_6}^T]_{i_3j_3}[\mathbf{b}_{h_7}\mathbf{b}_{h_8}^T]_{i_4j_4}\Big| \\
&\times\mathrm{E}|\varepsilon_{i_1,t-h_1}\varepsilon_{j_1,t-h_2}\varepsilon_{i_2,t-h_3}\varepsilon_{j_2,t-h_4}\varepsilon_{i_3,t-h_5}\varepsilon_{j_3,t-h_6}\varepsilon_{i_4,t-h_7}\varepsilon_{j_4,t-h_8}| \\
\leq&C\sum_{h_1,h_2,h_3,h_4,h_5,h_6,h_7,h_8=0}^{\infty}\sum_{i_1,j_1,i_2,j_2,i_3,j_3,i_4,j_4=1}^{p}|\mathbf{b}_{h_1}\mathbf{b}_{h_2}^T|_{i_1j_1}|\mathbf{b}_{h_3}\mathbf{b}_{h_4}^T|_{i_2j_2}|\mathbf{b}_{h_5}\mathbf{b}_{h_6}^T|_{i_3j_3}|\mathbf{b}_{h_7}\mathbf{b}_{h_8}^T|_{i_4j_4} \\
=&C\Big[\sum_{h=0}^{\infty}\sum_{g=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p}|\mathbf{b}_h\mathbf{b}_g^T|_{ij}\Big]^4.
\end{aligned}
\tag{1.7.18}
$$

And

$$
\begin{aligned}
&\sum_{h=0}^{\infty}\sum_{g=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p}|\mathbf{b}_h\mathbf{b}_g^T|_{ij} \leq \sum_{h=0}^{\infty}\sum_{g=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p}(|\mathbf{b}_h||\mathbf{b}_g^T|)_{ij} = \sum_{i=1}^{p}\sum_{j=1}^{p}\Big(\sum_{h=0}^{\infty}\sum_{g=0}^{\infty}|\mathbf{b}_h||\mathbf{b}_g^T|\Big)_{ij} \\
=&\sum_{i=1}^{p}\sum_{j=1}^{p}\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\sum_{g=0}^{\infty}|\mathbf{b}_g^T|\Big)_{ij} = \sum_{i=1}^{p}\sum_{j=1}^{p}\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\Big)_i\Big(\sum_{g=0}^{\infty}|\mathbf{b}_g|\Big)_j \\
=&\sum_{i=1}^{p}\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\Big)_i\sum_{j=1}^{p}\Big(\sum_{g=0}^{\infty}|\mathbf{b}_g|\Big)_j,
\end{aligned}
\tag{1.7.19}
$$

where $\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\Big)_i$ is the $i$-th element of the column vector $\sum_{h=0}^{\infty}|\mathbf{b}_h|$.

Since $(\sum_{h=0}^{\infty}|\mathbf{B}_h|)_{ij} = \sum_{h=0}^{\infty}\left(\left|\mathbf{A}^h\mathbf{S}^{-1}(\boldsymbol{\lambda}_0)\right|\right)_{ij} \leq \left(\sum_{h=0}^{\infty}\left|\mathbf{A}^h\right|\left|\mathbf{S}^{-1}(\boldsymbol{\lambda}_0)\right|\right)_{ij}$ where the row and column sums of $\sum_{h=0}^{\infty}\left|\mathbf{A}^h\right|\left|\mathbf{S}^{-1}(\boldsymbol{\lambda}_0)\right|$ are bounded uniformly in $p$, it holds that the row and column sums of $\sum_{h=0}^{\infty}|\mathbf{B}_h|$ are bounded uniformly in $p$. Note that

$$\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\Big)_i = \Big(\sum_{h=0}^{\infty}|\mathbf{B}_h^T\mathbf{a}|\Big)_i \leq \Big(\sum_{h=0}^{\infty}|\mathbf{B}_h^T||\mathbf{a}|\Big)_i,$$

where the row and column sums of $\sum_{h=0}^{\infty}\left|\mathbf{B}_h^T\right|$ and $|\mathbf{a}|$ are bounded uniformly in $p$. Hence the row and column sums of $\sum_{h=0}^{\infty}|\mathbf{B}_h^T||\mathbf{a}|$ are bounded uniformly in $p$. It follows from (1.7.18) and (1.7.19) that

$$\sup_p \mathrm{E}\left|\mathbf{a}^T\mathbf{y}_t\right|^8 \leq C\Big[\sum_{i=1}^{p}\Big(\sum_{h=0}^{\infty}|\mathbf{b}_h|\Big)_i\sum_{j=1}^{p}\Big(\sum_{g=0}^{\infty}|\mathbf{b}_g|\Big)_j\Big]^4 = O(1).$$

It is easy to prove that

$$\sup_p\|\boldsymbol{\Sigma}_0\mathbf{w}_i\|_1 < \infty, \quad \sup_p\|\boldsymbol{\Sigma}_1^T\mathbf{w}_i\|_1 < \infty, \quad \sup_p\|\boldsymbol{\Sigma}_0\mathbf{e}_i\|_1 < \infty.$$

Thus $\sup_p\|\mathbf{w}_i\boldsymbol{\Sigma}_0\mathbf{y}_t\|_1 < \infty$ and etc.

The row and column sums of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are bounded uniformly in $p$. Then

$$\sup_p \mathbf{w}_i^T\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_1^T\mathbf{w}_i = O(1).$$

Similarly, we can prove the other diagonal elements of $\mathbf{V}_i$ and $\mathbf{U}_i$ are bounded uniformly in $p$.

The proof is completed. □

# Chapter 2: Sparse Factor Modelling for Vast Time Series

## 2.1 Introduction

Modelling multivariate time series has many important applications in the fields such as finance, economics and environmental studies. Based on the success of univariate time series modelling, one natural way of modelling such data is the vector autoregressive and moving average model (ARMA) models. However, without regularization, vector ARMA models suffers from the over parametrization and the lack of identification problems, see Lutkepohl (2006). By assuming the transition matrix of vector autoregressive models to be sparse, Hsu et al. (2008) proposed a lasso type estimator. Han and Liu (2013) exploited the linear programming technique and the proposed method is very fast to solve via parallel computing. Another frequently used approach is modelling using factors. Attempts include Pena and Box (1987), Stock and Watson (2002), Bai and Ng (2002), Hallin and Liska (2007), Pan and Yao (2008), Lam, Yao and Bathia (2011), Fan et al. (2013), Onatski (2014).

In this paper, we decompose the original process into a dynamic part, i.e. a common factor process and a static part, i.e. a white noise process. Motivated by practical needs

and the characteristic of high dimensional data, the sparsity assumption on factor loading matrix is imposed. Different from Lam, Yao and Bathia (2011)'s method, which is equivalent to an eigenanalysis of a non negative definite matrix, we add a constraint to control the number of nonzero elements in each column of the factor loading matrix. Our proposed sparse estimator is then the solution of a constrained optimization problem. Numerically, we solve it via the generalized deflation method (Mackey 2009) and GSLDA method (Moghaddam et al. 2006). The tuning parameter is chosen by cross validation. We establish the asymptotic results when both the sample size and dimensionality go to infinity or even when the latter is larger. Compared to Lam, Yao and Bathia (2011)'s method, when the factor is weak in the sense that $\delta > 1/2$ in their paper, our newly proposed estimator may have a faster convergence rate. Our simulation results convinced that when the common factor is weak, the newly proposed estimator has smaller error compared to Lam, Yao and Bathia (2011)'s estimator even when we allow the number of nonzero elements in each column of the factor loading matrix increases with the dimensionality.

The rest of the paper is organized as follows. Section 2.2 introduces the model, the motivation for sparsity and the new sparse estimator. The asymptotic theory for the proposed estimation method is presented in section 2.3. Simulation results and real data analysis are reported, respectively, in section 2.4 and 2.5. The technical proofs are relegated to Appendix.

## 2.2 Model and Estimation Method

### 2.2.1 Models

Let $\mathbf{y}_t = (y_{1,t}, \cdots, y_{p,t})^T$ be an observable $p \times 1$ vector time series process. The factor model decomposes $\mathbf{y}_t$ in the following form:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{2.2.20}$$

where $\mathbf{x}_t = (x_{1,t}, \cdots, x_{r,t})^T$ is a $r \times 1$ latent factor time series with unknown $r \leq p$ and $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_r)$ is a $p \times r$ unknown constant matrix. $\boldsymbol{\varepsilon}_t$ is a white noise process with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. The first part of (2.2.20) is a dynamic part and the serial dependence of $\mathbf{y}_t$ is driven by $\mathbf{x}_t$. We will achieve dimension reduction once $r \ll p$ in the sense that the dynamics of $\mathbf{y}_t$ is driven by a much lower dimensional process $\mathbf{x}_t$. Let the rank of $\mathbf{A}$ be $r$. If the rank of $\mathbf{A}$ is smaller than $r$, (2.2.20) can be expressed using a lower dimensional factor process. We also assume no linear combination of the components of $\mathbf{x}_t$ is white noise. The pair $(\mathbf{A}, \mathbf{x}_t)$ itself is not identifiable since model (2.2.20) is unchanged if we use the pair $(\mathbf{AH}, \mathbf{H}^{-1}\mathbf{x}_t)$ to replace it for any $r \times r$ nonsingular $\mathbf{H}$. But the $r$ dimensional linear space spanned by the columns of $\mathbf{A}$, denoted by $\mathcal{M}(\mathbf{A})$, is uniquely defined due to $\mathcal{M}(\mathbf{A}) = \mathcal{M}(\mathbf{AH})$. Without loss of generality, we assume $\mathbf{A}$ to be a column orthogonal matrix, that is $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$, where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. This is because $\mathbf{A}$ admits the QR decomposition $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular, then we can replace $(\mathbf{A}, \mathbf{x}_t)$ by $(\mathbf{Q}, \mathbf{R}\mathbf{x}_t)$.

We see that $\mathbf{A}$ is not identifiable. However, this lack of uniqueness of $\mathbf{A}$ can be treated as an advantage since we can choose any particular $\mathbf{A}$ of which the estimation can be

simple. Note that the $(k,i)$-th element of $\mathbf{A}$, $a_{k,i}$, measures the effect of the $i$-th common factor, $x_{i,t}$, on the $k$-th random variable of $\mathbf{y}_t$, $y_{k,t}$: large $a_{k,i}$ means $x_{i,t}$ is important to $y_{k,t}$, small $a_{k,i}$ means $x_{i,t}$ is less important to $y_{k,t}$ and $a_{k,i} = 0$ means the $x_{i,t}$ has no effect on the $y_{k,t}$.

In this paper we assume the latent process $\mathbf{x}_t$ is weakly stationary. Furthermore, we assume $\text{Cov}(\mathbf{x}_t, \boldsymbol{\varepsilon}_{t+k}) = 0$ for any $k \geq 0$. This allows the correlation between the previous white noise and the factors up to present, which enlarges the model capacity compared with most factor modelling literature. Pan and Yao (2008) handled with the non-stationary case.

Note that in model (2.2.20), only $\mathbf{y}_t$ is observable. Once we obtain the estimator $\widehat{\mathbf{A}}$ of $\mathbf{A}$, we can estimate $\mathbf{x}_t$ by $\widehat{\mathbf{A}}^T\mathbf{y}_t$. The number of the common factors $r$ has to be estimated as well but in this paper, we focus on the estimation of $\mathbf{A}$ and we directly use an estimator proposed by Lam and Yao (2012). Literature of estimating the number of common factor $r$ includes Bai and Ng (2002), Hallin and Liska (2007) and Pan and Yao (2008).

From the point of interpretation, sparsity is preferred, especially when the dimensionality $p$ is very large. If we want to recover what the common factors represent in practice, we need the following approximation:

$$\mathbf{x}_t \approx \mathbf{A}^T\mathbf{y}_t.$$

For the $i$-th common factor $x_{i,t}$ at time $t$, it holds that $x_{i,t} \approx \mathbf{a}_i^T\mathbf{y}_t$. We need to figure out the practical meaning of $x_{i,t}$ via the practical meaning of $y_{k,t}$'s and their corresponding weights $a_{k,i}$'s. When $p$ is large, it is essential to reduce the size of explicitly used $y_{k,t}$'s in order to interpret, where sparse assumption is required. From the point of practical concerns, when we have a large amount of variables and we are seeking their common factors, it is more likely that each common factor will only affect some of the variables but not all. In

practice, there might exist such common factor that influences all the variables in $\mathbf{y}_t$ and it is more likely to happen especially when $p$ is small, but this fact does not contradict with our sparsity assumption.

Let us look at one real example. Lam and Yao (2012) analyzed a multivariate environmental time series data which is a collection of monthly average sea surface air pressure records (in Pascal) for 528 month from January 1958 to December 2001. For each fixed month, the data are collected over the same $10 \times 44$ grid in a range of 22.5° longitude and 110° latitude in the North Atlantic Ocean. They denoted the air pressure in the $t$-th month at location $(u, v)$ by $P_t(u, v)$, where $t = 1, 2, \cdots, 528$ and $u = 1, 2, \cdots, 10$, $v = 1, 2, \cdots, 44$. If we vectorize these 440 locations in $\mathbf{y}_t$ for each month $t$, then we get a 440 dimensional time series data with 528 observations. They analyzed this data using common factor model above and estimated $\mathbf{A}$, $\mathbf{x}_t$ and $r$ by their proposed method. Figure 2.12 is the plot of the factor loadings of the 3 ($\widehat{r} = 3$) common factors.

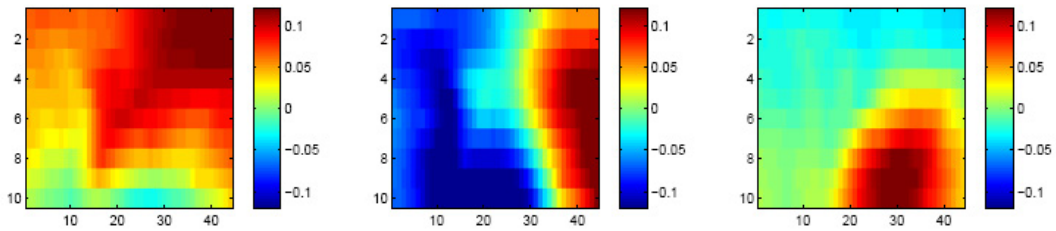


Figure 2.12: Factor loading surface of the 1st, 2nd and 3rd factors (from left to right)

The x-axis is $v$ and y-axis is $u$. The $i$-th plot represents the loadings of $i$-th common factor, which is the $i$-th column of $\widehat{\mathbf{A}}$. Some patterns are as follows: the 1st factor mainly influences the north and northeast in particular; the 2nd factor is the main factor for most part except for the narrow middle part; the southeast is mainly influenced by the 3rd factor. Also note that there are some small (sky-blue and yellow parts) or even zero (green part)

loadings of each common factor. For example, the very south part of 1st plot, the narrow middle part of 2nd plot and the north and west part of 3rd plot are with small factor loadings. These imply the sparsity condition of $\mathbf{A}$.

### 2.2.2 Estimation

We introduce some notations first. For a $p \times 1$ vector $\mathbf{v} = (v_1, \cdots, v_p)^T$, $\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$ is the Euclidean norm and $\|\mathbf{v}\|_0 =$card$\{$support$(\mathbf{v})\}$ is the number of non zero elements in $\mathbf{v}$. Let $\mathbb{V}$ be the set of $p \times r$ orthogonal matrices. Let $\mathbb{V}^\perp$ be the set of $p \times (p-r)$ orthogonal matrices such that $(\mathbf{V}, \mathbf{V}^\perp)$ is orthogonal, where $\mathbf{V} \in \mathbb{V}, \mathbf{V}^\perp \in \mathbb{V}^\perp$. For a set $\mathbb{K}$, $|\mathbb{K}|$ is its cardinality. For a $p \times r$ matrix $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_r)$, define $\|\mathbf{U}\|_0 = \sum_{j=1}^p \mathbb{I}\{\|\mathbf{u}_{j*}\|_2 \neq 0\}$ where $\mathbf{u}_{j*}$ is the $j$-th row of $\mathbf{U}$. Note that $\|\mathbf{U}\|_0$ counts the number of nonzero rows in $\mathbf{U}$.

Note that $\mathbf{A}$ equals to the matrix consisting of the first $r$ orthonormal eigenvectors of the $p \times p$ positive semidefinite matrix

$$\mathbf{M} = \sum_{k=1}^{k_0} \mathbf{\Sigma}_\mathbf{y}(k) \mathbf{\Sigma}_\mathbf{y}(k)^T, \tag{2.2.21}$$

corresponding to its $r$ non-zero eigenvalues, where $\mathbf{\Sigma}_\mathbf{y}(k) = \text{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t)$ and $k_0$ is a predetermined positive constant. Denote $\lambda_i$ as the $i$-th largest eigenvalue of $\mathbf{M}$. See Lam, Yao and Bathia (2011) for more details.

Put

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_\mathbf{y}(k) \widehat{\mathbf{\Sigma}}_\mathbf{y}(k)^T, \tag{2.2.22}$$

where

$$\widehat{\mathbf{\Sigma}}_\mathbf{y}(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} (\mathbf{y}_{t+k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^T,$$

and $\bar{\mathbf{y}} = n^{-1} \sum_{t=1}^n \mathbf{y}_t$.

We assume $\|\mathbf{A}\|_0 \leq s$. To obtain the estimator $\mathcal{M}(\widehat{\mathbf{A}})$ of $\mathcal{M}(\mathbf{A})$, it suffices to solve the following optimization problem:

$$\widehat{\mathbf{A}} = \arg\max_{\mathbf{V}\in\mathbb{V}} tr(\mathbf{V}^T\widehat{\mathbf{M}}\mathbf{V}) \quad \text{subject to} \quad \|\mathbf{V}\|_0 \leq s. \tag{2.2.23}$$

Note that for $\mathbf{V} \in \mathbb{V}$ and $\mathbf{V}^\perp \in \mathbb{V}^\perp$, we have $tr(\mathbf{V}^T\widehat{\mathbf{M}}\mathbf{V}) = tr(\widehat{\mathbf{M}}) - tr((\mathbf{V}^\perp)^T\widehat{\mathbf{M}}\mathbf{V}^\perp)$. Since $tr((\mathbf{V}^\perp)^T\widehat{\mathbf{M}}\mathbf{V}^\perp) \geq 0$, it follows that $\max tr(\mathbf{V}^T\widehat{\mathbf{M}}\mathbf{V}) = tr(\widehat{\mathbf{M}})$ when $\mathbf{V}$ consists of the $r$ eigenvectors of $\widehat{\mathbf{M}}$ corresponding its $r$ non-zero eigenvalues. This means without the sparsity constraint $\|\mathbf{V}\|_0 \leq s$. the solution of (2.2.23) is the same as the estimator of Lam, Yao and Bathia (2011). Numerically, we employ the generalized deflation method (Mackey 2009) to approximate $\widehat{\mathbf{A}}$ in (2.2.23). Specifically, the algorithm is as follows

(1) Input $\widehat{\mathbf{M}}$ and the cardinalities of $r$ columns $\{s_1, \cdots, s_r\}$.

(2) Initialize $i = 1$, $s = s_i$ and $\mathbf{B} = \mathbf{I}_p$.

(3) Solve $\widehat{\mathbf{v}} = \arg\max_{\mathbf{v}^T\mathbf{B}\mathbf{v}=1, \|\mathbf{v}\|_0\leq s} \mathbf{v}^T\widehat{\mathbf{M}}\mathbf{v}$, Compute $\mathbf{q} = \mathbf{B}\widehat{\mathbf{v}}$.

(4) Update $\widehat{\mathbf{M}}$ by $\widehat{\mathbf{M}} \leftarrow (\mathbf{I}_p - \mathbf{q}\mathbf{q}^T)\widehat{\mathbf{M}}(\mathbf{I}_p - \mathbf{q}\mathbf{q}^T)$.

Update $\mathbf{B}$ by $\mathbf{B} \leftarrow \mathbf{B}(\mathbf{I}_p - \mathbf{q}\mathbf{q}^T)$.

Update $i \leftarrow i + 1$, $s \leftarrow s_i$.

(5) Return $\widehat{\mathbf{v}}/\|\widehat{\mathbf{v}}\|_2$.

(6) Repeat step (3) to (5) until $i = r + 1$.

Apparently, we totally repeat $r$ times. The $i$-th, $i = 1, \cdots, r$ output in step (5) is the sparse estimator of the $i$-th column of $\mathbf{A}$. We also need to numerically approximate the solution of the optimization in step (3), where we adopt the GSLDA method (Moghaddam et al. 2006), specifically, the algorithm is as follows

(3.1) Decompose $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ and set $\mathbf{B}^{1/2} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^T$ and $\mathbf{C} = \mathbf{B}^{-1/2}\widehat{\mathbf{M}}\mathbf{B}^{-1/2}$.

(3.2) Initialize $t = 1$ and $\mathbf{x}_0 \in \mathbb{R}^p$.

(3.3) Compute $\mathbf{x}_t^* = T_s\left(\frac{\mathbf{B}^{-1/2}\mathbf{C}\mathbf{x}_{t-1}}{\|\mathbf{C}\mathbf{x}_{t-1}\|_2}\right)$, where $T_s(\mathbf{x})$ only keeps elements of $\mathbf{x}$ with the largest $s$ absolute values and sets all other elements to be 0.

(3.4) Compute $\mathbf{x}_t = \mathbf{x}_t^*$. Update $t \leftarrow t + 1$.

(3.5) Repeat step (3.3) and (3.4) until $\mathbf{x}_t$ is convergent.

The obtained $\mathbf{x}_t$ is the solution of the optimization problem in step (3).

In practice, we use the ratio-based estimator to get the estimator for $r$, which is defined by:

$$\widehat{r} = \arg\min_{1 \le j \le R} \widehat{\lambda}_{j+1}/\widehat{\lambda}_j \tag{2.2.24}$$

where $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_p$ are the eigenvalues of $\widehat{\mathbf{M}}$ and the integer $R(r \le R < p)$ can be chosen as, for instance, $p/2$. More details are in Lam and Yao (2012).

## 2.3 Theoretical Properties

Summarizing the assumptions, we have

A1. $\mathbf{A}$ is column orthogonal, that is, $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$; $\mathbf{a}'\mathbf{x}_t$ is not white noise for any $\mathbf{a} \in \mathbb{R}^p$; $\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$.

A2. The factor loading matrix is sparse in the sense that $\|\mathbf{A}\|_0 \le s$.

A3. The covariance matrix $\mathrm{Cov}(\mathbf{x}_t, \boldsymbol{\varepsilon}_{t+k}) = 0$ for any $k \ge 0$.

A4. The eigenvalues of $\mathbf{M}$ satisfies $\lambda_1 > \cdots > \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_p$.

A5. There exist positive constants $K_1$ and $r_1 \in (0, 1]$ such that the process $\{\mathbf{y}_t\}$ in model (2.2.20) is strictly stationary and $\alpha$-mixing with mixing coefficients satisfying

$$\alpha(u) \leq \exp(-K_1 u^{r_1}),$$

for any $u \geq 1$, where

$$\alpha(u) \equiv \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_u^\infty} \big|P(A)P(B) - P(AB)\big|, \tag{2.3.25}$$

and $\mathcal{F}_i^j$ denotes the $\sigma$-algebra generated by $\{\mathbf{y}_t, i \leq t \leq j\}$.

A6. There exist positive constants $K_2, K_3$ and $r_2 \in (0, 2]$ such that

$$P(|\mathbf{v}^T(\mathbf{y}_t - E\mathbf{y}_t)| > \tau) \leq K_2 \exp(-K_3 \tau^{r_2}),$$

for any $\tau > 0$ and unit vector $\mathbf{v}$.

Conditions A1, A3 and A4 are regularity conditions the same as Lam, Yao and Bathia (2011). Condition A2 is the sparsity assumption. Condition A6 requires the linear combination of $\mathbf{y}_t$ has exponential type tails. Together with A5, they allow us to apply the large deviation theory in Merlevéde et al. (2011). the requirements of $r_1 \leq 2$ and $r_2 \leq 1$ are not necessary. The theoretical results proposed can still be established for $r_1 > 2$ and $r_2 > 1$. They are assumed here to simplify the presentation of the theoretical results.

We first present a result for the leading eigenvector estimator $\widehat{\mathbf{a}}_1$. This is a special case of (2.2.23) when we restrict $\mathbf{V}$ to be a $p \times 1$ vector, in which case the optimization problem becomes

$$\widehat{\mathbf{a}}_1 = \arg\max_{\|\mathbf{v}\|_2 = 1} \mathbf{v}^T \widehat{\mathbf{M}} \mathbf{v} \quad \text{subject to} \quad \|\mathbf{v}\|_0 \leq s. \tag{2.3.26}$$

Theorem 5 and 6 are the asymptotic properties when $p$ is fixed.

**Theorem 5** *Let conditions A1, A3 – A5 hold and* $\widehat{\mathbf{a}}_1$ *be the solution of (2.3.26). When* $p$ *is fixed, as* $n \to \infty$*, it holds that*

$$\sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} = O_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Theorem 5 is the consistency result of the leading eigenvector estimator $\widehat{\mathbf{a}}_1$. Theorem 6 extends the above result to the $\widehat{\mathbf{A}}$.

**Theorem 6** *Let conditions A1, A3 – A5 hold and* $\widehat{\mathbf{A}}$ *be the solution of (2.2.23). When* $p$ *is fixed, as* $n \to \infty$*, it holds that*

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F = O_p\Big(\frac{1}{\sqrt{n}}\Big),$$

*where* $\|\cdot\|$ *is the Frobenius norm.*

$\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F$ is the canonical angle between two subspaces $\mathcal{M}(\widehat{\mathbf{A}})$ and $\mathcal{M}(\mathbf{A})$, see Vu and Lei (2013) for more details. When $r=1$, $\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F = \sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2}$. Theorem 5 and 6 and two trivial results since $p$ is fixed. The asymptotic properties are presented in Theorem 7 and 8 when $p$ increases with $n$.

**Theorem 7** *Let conditions A1 – A6 hold and* $\widehat{\mathbf{a}}_1$ *be the solution of (2.3.26). As* $n \to \infty, p \to \infty$*, it holds that*

$$\sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} = O_p\Big(\frac{{\lambda_1}^{1/2}}{\lambda_1-\lambda_2}\sqrt{\frac{s^3 p\log p}{n}}\Big). \tag{2.3.27}$$

**Theorem 8** *Let conditions A1 – A6 hold and* $\widehat{\mathbf{A}}$ *be the solution of (2.2.23). As* $n \to \infty, p \to \infty$*, it holds that*

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F = O_p\Big(\frac{{\lambda_1}^{1/2}}{\lambda_r}\sqrt{\frac{s^3 p\log p}{n}}\Big), \tag{2.3.28}$$

*where* $\|\cdot\|$ *is the Frobenius norm.*

Compared with Lam, Yao and Bathia (2011)'s estimator, which owns the convergent rate $\frac{p^\delta}{\sqrt{n}}$ for $\delta \in [0,1]$, if the factor is weak enough in the sense that $\delta$ is large than 0.5, our sparse estimator will obtain a faster convergent rate $\sqrt{\frac{p\log p}{n}}$, for example, when $s$ and $\lambda_i$'s are all constants.

## 2.4 Choice of Tuning Parameter

In practice, the cardinality of $\mathbf{A}$ is unknown. We can choose it via cross validation. Let the training sample size and validation sample size be $n_1$ and $n_2$ respectively, where $n_1+n_2 = n$. Assume the set of the possible cardinality $s$ is $\mathbb{S}$. For each fixed $s \in \mathbb{S}$, we fit model (2.2.20) using $\mathbf{y}_1, \cdots, \mathbf{y}_{n_1}$ with our proposed sparse estimation procedure and obtained $\widehat{\mathbf{A}}_s$. Consequently, we estimate the factors by $\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}_s^T \mathbf{y}_t, t = 1, \cdots, n_1$. We can then make a one step ahead prediction for $\mathbf{y}_t$ by $\widehat{\mathbf{y}}_{n_1+1} = \widehat{\mathbf{A}}_s \widehat{\mathbf{x}}_{n_1+1}^{(1)}$, where $\widehat{\mathbf{x}}_{n_1+1}^{(1)}$ is a one step forecast for $\mathbf{x}_t$ based on the estimated past $\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_{n_1}$, for example by fitting a autoregressive model to $\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_{n_1}$. Then we obtained the test error for $\mathbf{y}_{n_1+1}$, which is defined as $\|\mathbf{y}_{n_1+1} - \widehat{\mathbf{y}}_{n_1+1}\|_2/p$. We then perform the above procedure of $n_2$ rolling windows each of length $n_1$ and compute the test error of the one step forecast of $\mathbf{y}_t$. Hence we obtained the error for the $i$-th rolling window

$$\frac{\|\mathbf{y}_{n_1+i} - \widehat{\mathbf{y}}_{n_1+i}\|_2}{p}, i = 1, \cdots, n_2.$$

The measure of the prediction with tuning parameter $s$ is defined as

$$Err_s = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{\|\mathbf{y}_{n_1+i} - \widehat{\mathbf{y}}_{n_1+i}\|_2}{p}. \tag{2.4.29}$$

We then choose the tuning parameter minimizing $Err_s$ among $s \in \mathbb{S}$.

## 2.5 Simulation Studies

To examine the finite sample performance of the proposed estimation methods, we conduct some simulations under different scenarios.

### 2.5.1 scenario 1

We consider a simple sparse one factor model. We generate a $s \times 1$ unit vector $\mathbf{z}$ firstly, where we set $s = [\sqrt{p}] + 1$. We then construct $\mathbf{A}$ such that the 1st to $s$-th elements of $\mathbf{A}$ equal to $\mathbf{z}$, and the rest of $\mathbf{A}$ are all zeros. Note that in this simple one factor model, the true factor loading matrix we are estimating is simply the vector $\mathbf{A}$ we construct. The factor process is generate from $x_t = 0.8x_{t-1} + \eta_t$ and $\eta_t$ are independently generated from $N(0, 1)$. The noise terms $\varepsilon_{i,t}$ are independently generated from $N(0, 1)$ for all $i, t$.

We generate data from (2.2.20) with different setting for $n$ and $p$. We apply the proposed method and compare the error (2.3.27) with the estimation method of Lam, Yao and Bathia (2011). For simplicity, we set the tuning parameter cardinality to be the true number of nonzero elements $s$. In practice, we need to use cross validation to choose the cardinality as shown in section 2.4. And in later section 2.5.4, we will see that even if the chosen cardinality is not the same as the true cardinality, the performance of our proposed estimator is better than Lam, Yao and Bathia (2011)'s method. The replication time is 200 in all experiments.

Figure 2.13 depicts two boxplots of (2.3.27) with $p$ equals to, respectively, 20 and 200. The left panel is for the Lam et al,'s method and the right panel is the sparse estimator. As the sample size $n$ increases from 100, 200, 300, 500 to 1000, (2.3.27) decreases for both

methods. The performance of the right panel is better than the left.

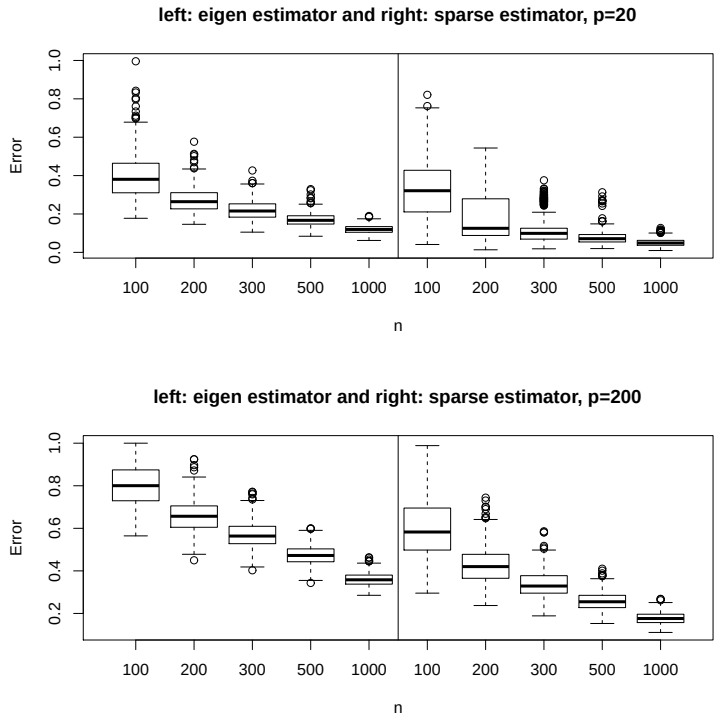


Figure 2.13: Boxplots of (2.3.27) for $p = 20, 200$ and $n = 100, 200, 300, 500, 1000$.

Figure 2.14 depicts three boxplots of (2.3.27) with $n$ equals to, respectively, 200, 300, 500. As $p$ increases from 100, 200, 300, 400 to 500, (2.3.27) increases for both methods. Again, the performance of the right panel is better than the left.

### 2.5.2 scenario 2

We consider a three common factor model, that is $r = 3$. We generate a $s \times r$ orthogonal matrix $\mathbf{Z}$ firstly, where we set $s = [\sqrt{p}]+1$. We then construct $\mathbf{A}$ such that the $(i-1)s+1$-th to $is$-th elements of $\mathbf{A}_{*,i}$ equal to $\mathbf{Z}_{*,i}$ for $i = 1, \cdots, r$ and the rest of $\mathbf{A}$ are all zeros, where $\mathbf{A}_{*,i}$ represents the $i$-th column of $\mathbf{A}$. We then independently generated three common factors from AR(1) process with coefficient 0.8, 0.6, 0.4 respectively. The noise terms $\varepsilon_{i,t}$

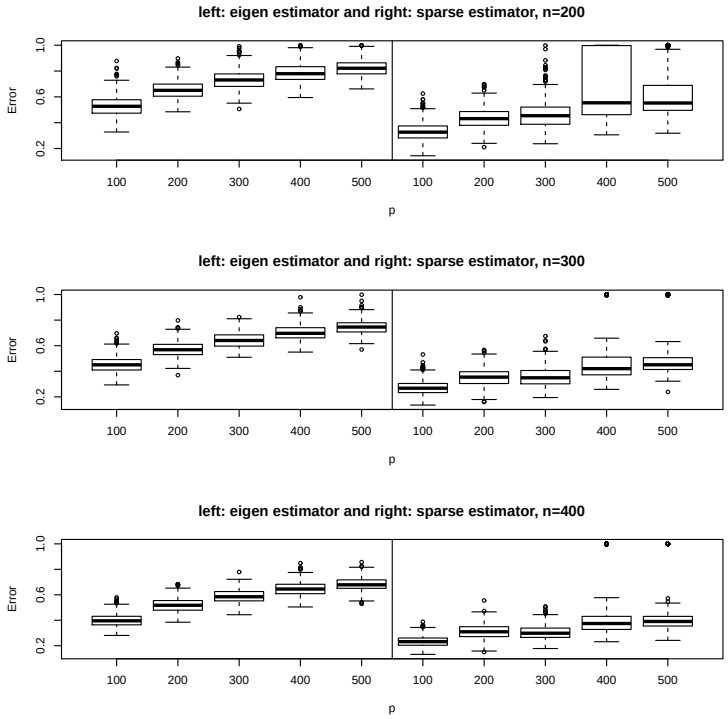


Figure 2.14: Boxplots of (2.3.27) for $n = 200, 300, 400$ and $p = 100, 200, 300, 400, 500$.

are independently generated from $N(0, 1)$ for all $i, t$.

Figure 2.15 depicts two boxplots of (2.3.28) with $p$ equals to, respectively, 20 and 200. The left panel is for Lam et al. (2011)'s method and the right panel is estimator (2.2.23). As the sample size $n$ increases from 100, 200, 300, 500 to 1000, (2.3.28) decreases for both methods. Lam et al. (2011)'s method outperforms the estimator (2.2.23) when $p = 20$. When $p$ increases to 200, the newly proposed sparse estimator performs better than Lam et al.'s estimator except for $n = 1000$.

Figure 2.16 depicts three boxplots of (2.3.28) with $n$ equals to, respectively, 200, 300, 500. As $p$ increases from 100, 200, 300, 400 to 500, (2.3.28) increases for both methods. Again, the performance of the right panel is better than the left.

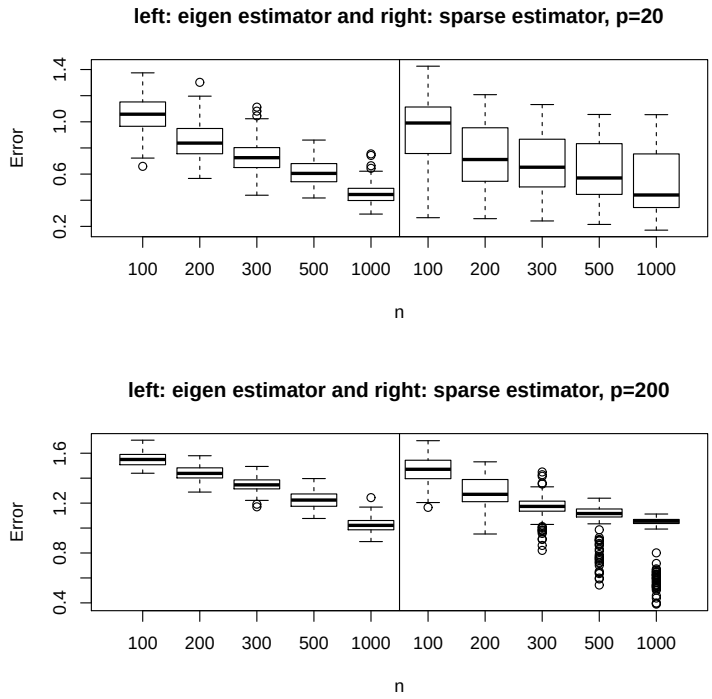


Figure 2.15: Boxplots of (2.3.28) for $p = 20, 200$ and $n = 100, 200, 300, 500, 1000$.

### 2.5.3 scenario 3

We consider a three common factor model. We generate $\mathbf{A}$ and $\varepsilon_{i,t}$ the same as scenario 2. The factor process $\mathbf{x}_t = (x_{1,t}, x_{2,t}, x_{3,t})^T$ is defined by

$$x_{1,t} = \omega_t, \quad x_{2,t} = \omega_{t-1}, \quad x_{3,t} = \omega_{t-2},$$

where $\omega_t = 0.8z_{t-1} + z_t$ and $z_t$ are independently generated from $N(0, 1)$. The noise terms $\varepsilon_{i,t}$ are independently generated from $N(0, 1)$ for all $i, t$.

Figure 2.17 depicts three boxplots of (2.3.28) with $n$ equals to, respectively, 200, 400. As $p$ increases from 100, 200, 300, 400 to 500, (2.3.28) increases for both methods. Again, the performance of the right panel is better than the left.

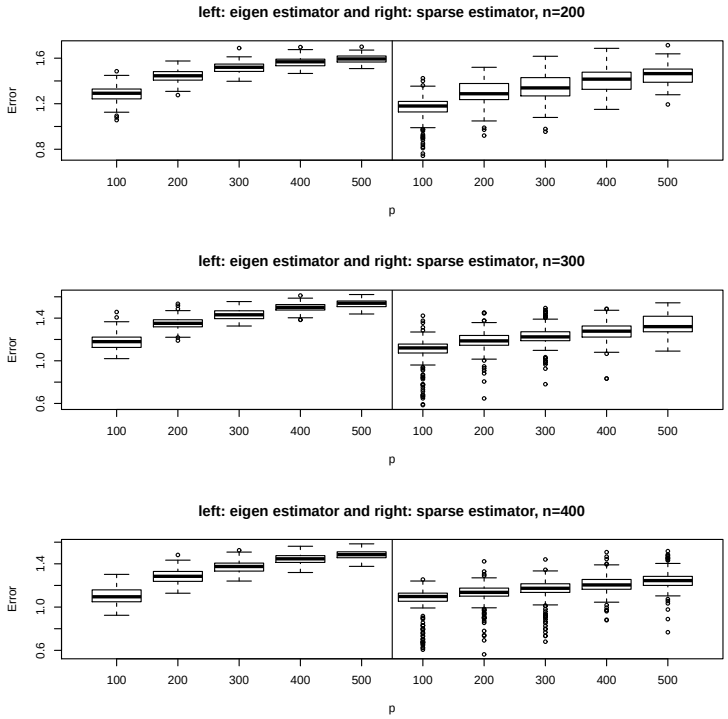


Figure 2.16: Boxplots of (2.3.28) for $n = 200, 300, 400$ and $p = 100, 200, 300, 400, 500$.

### 2.5.4 Cross Validation

We consider a single factor model and generate all parameters the same as scenario 1. We apply the proposed sparse estimation method and use the cross validation method in section 2.4 to choose the cardinality. We consider four cases: (1) $p = 50, n = 500$, (2) $p = 200, n = 500$, (3) $p = 200, n = 300$ and (4) $p = 500, n = 300$. Table 2.2 lists the mean, standard error of the chosen cardinality and the mean of test errors (2.4.29) for both methods.

Figure 2.18 depicts four boxplots of (2.3.27) of cases (1) to (4) respectively. The left panel in each plot is the performance of the eigenanalysis estimator and the right panel is the sparse estimator where we choose the number of cardinality by cross validation. As we can see from the plots, even if we the chosen cardinality might be different from the true

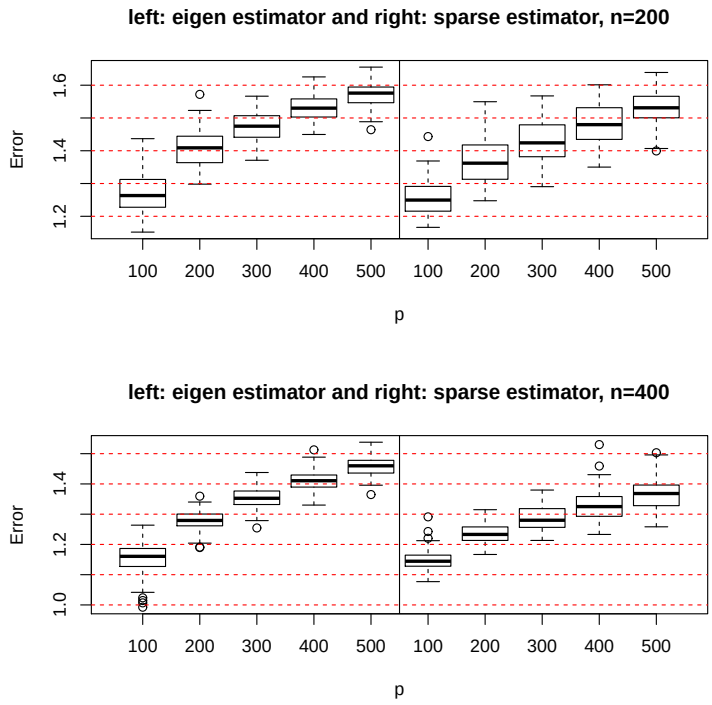


Figure 2.17: Boxplots of (2.3.28) for $n = 200, 400$ and $p = 100, 200, 300, 400, 500$.

| | p | n | true $s$ | $\widehat{s}$ | $Err^{eigen}$ | $Err^{sparse}$ |
|---|---|---|---|---|---|---|
| Case 1 | 50 | 500 | 8 | 5.7(3.6) | 0.147 | 0.142 |
| Case 2 | 200 | 500 | 15 | 12.9(3.0) | 0.070 | 0.070 |
| Case 3 | 200 | 300 | 15 | 13.2(4.5) | 0.070 | 0.070 |
| Case 4 | 500 | 300 | 23 | 11.8(5.5) | 0.045 | 0.045 |

Table 2.2: $\widehat{s}$ is the chosen cardinality by cross validation, $Err^{eigen}$ and $Err^{sparse}$ are the mean of test errors (2.4.29) for both methods.

value, the performance of the sparse estimator still dominates the original eigenanalysis estimation for all cases.

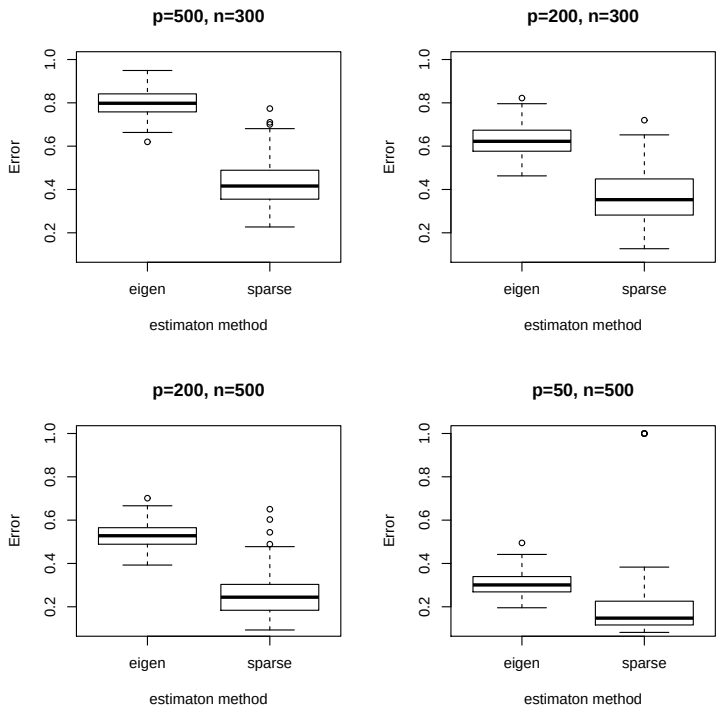


Figure 2.18: Boxplots of (2.3.27) for different $p$ and $n$.

## 2.6 Real Data Analysis

Let us revisit the monthly sea surface air pressure example in section 2.2. We observe the air pressure for 528 months and for each month (that is $n = 528$), we observe $10 \times 44$ grid, hence $p = 440$. The air pressure of $(u, v)$-th grid is denoted by $P_t(u, v)$, $u = 1, \cdots, 10$ and $v = 1, \cdots, 44$. Note that the 440 grids we use in this paper might be different from the Lam and Yao (2012). We first subtract each data point by the monthly mean over 528 months at each location. The centralized data is plotted in Figure 2.19.

We then employ the ratio-based estimator to estimate the number of common factors. The estimated eigenvalues in descending order and their ratios are plotted in Figure 2.20. Note that we choose $k_0 = 1$ since the ratio-based estimator is not sensitive to the choice

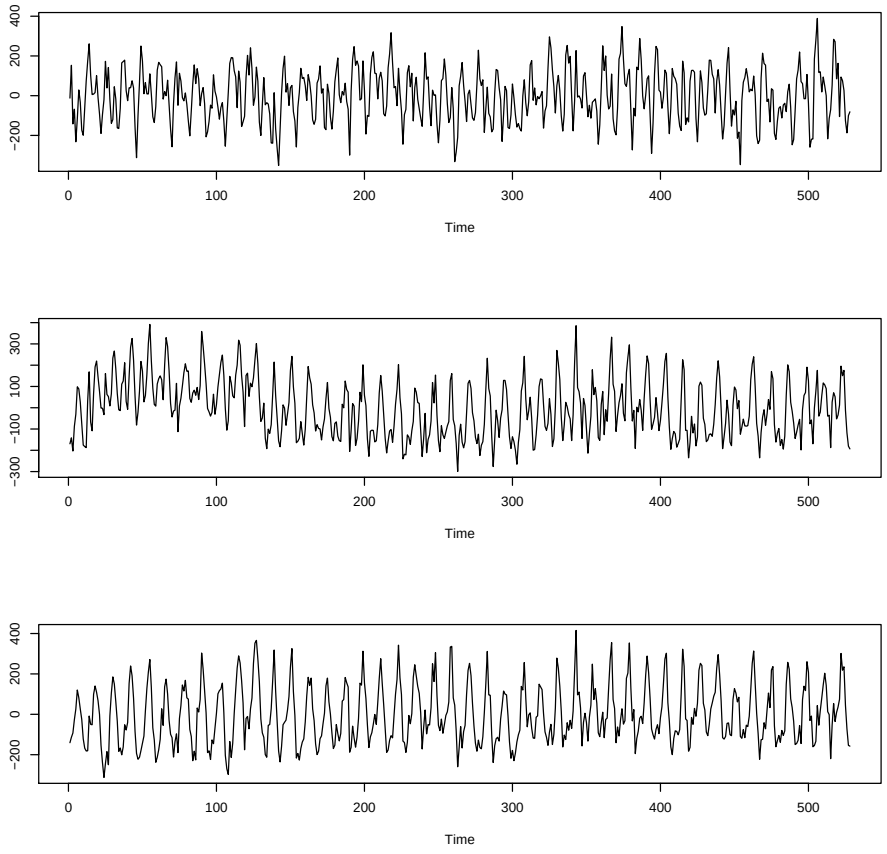


Figure 2.19: Time series plot of the sea surface air pressure data at $(u, v) = (1, 1)$ (top panel), $(u, v) = (5, 5)$ (middle panel) and $(u, v) = (10, 5)$ (bottom panel).

$k_0$. It is clear that $\widehat{r} = 3$.

We fit the model via two methods: Lam and Yao (2012) and the newly proposed sparse estimator. The number of cardinality is chosen using the cross validation method in section 2.4 and we choose the number of training observations as 475 (roughly $0.9n$) and the test dataset size as $0.1n$. The chosen cardinality is 250. Figure 2.21 is the color map of the estimated factor loading matrix. The test error (2.4.29) using the original method is 26.3, and 19.8 using the sparse estimation. From this point of view, the newly proposed method outperforms the previous eigenanalysis method.

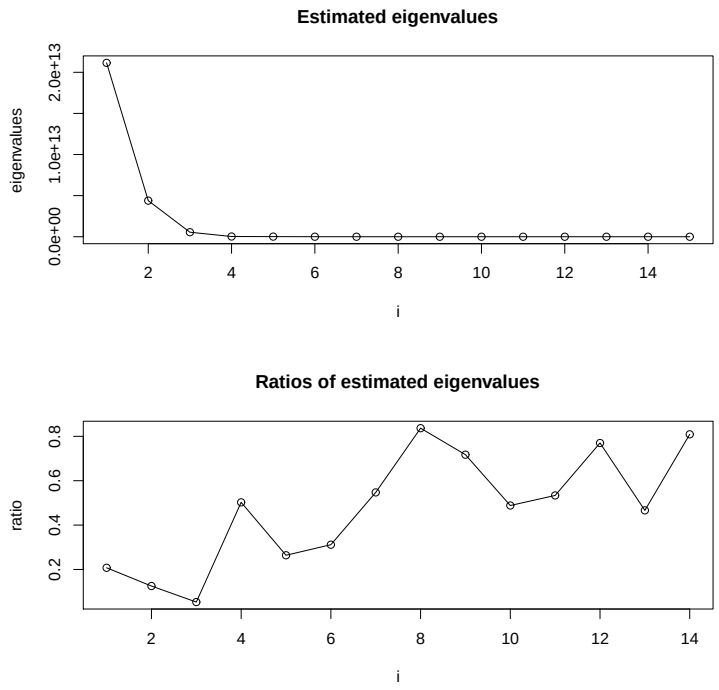


Figure 2.20: Top panel: plots of eigenvalues in descending order and bottom panel: ratios of eigenvalues of $\widehat{\mathbf{M}}$.

## 2.7 Appendix: Proofs

In this section we give the proof of theorem 7 and 8. The proof of theorem 5 and 6 are similar to theorem 7 and 8 but are simpler, hence omitted. We use $C$ to denote a generic positive constant, which may be different at different places.

**Lemma 2** *The estimator* $\widehat{\mathbf{a}}_1$ *in (2.3.26) of* $\mathbf{a}_1$ *satisfies*

$$\sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} \leq \frac{2}{\lambda_1-\lambda_2} \sup_{\|\mathbf{v}\|_2=1\cap\|\mathbf{v}\|_0\leq 2s} |\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|.$$

**Proof.** Recall that the $r$ non-zero eigenvalues of $\mathbf{M}$ are $\lambda_1 > \cdots > \lambda_r$ with $\mathbf{a}_1, \cdots, \mathbf{a}_r$ the corresponding eigenvectors. And let $\mathbf{a}_{r+1}, \mathbf{a}_p$ be the eigenvectors corresponding to the

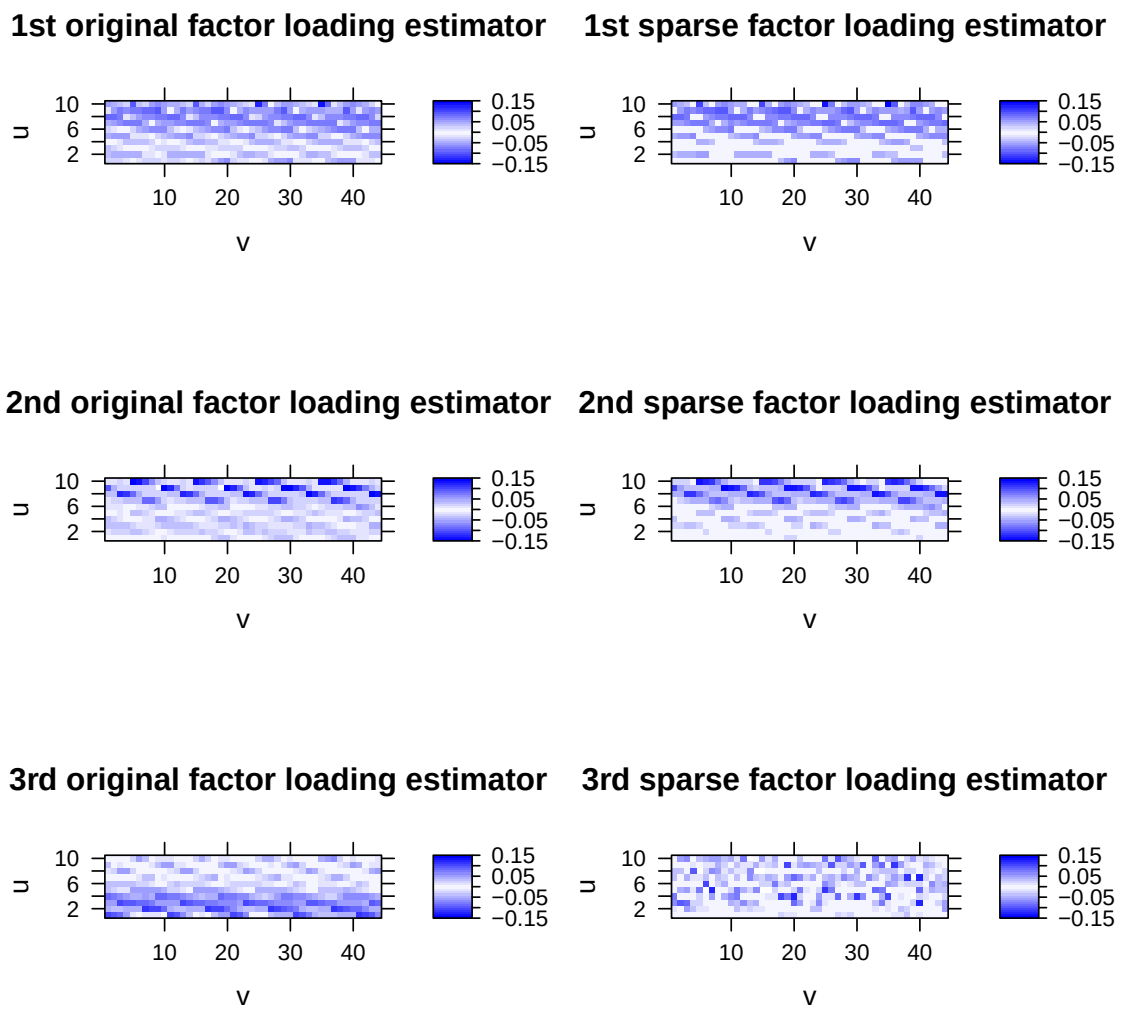


Figure 2.21: Factor loading surface of the 1st (top panel), 2nd (middle panel) and 3rd (bottom panel) factors for the eigenanalysis estimator (left panel) and sparse estimator (right panel).

0 eigenvalues of $\mathbf{M}$ such that $\mathbf{a}_i^T\mathbf{a}_j = 0$ for $i \neq j$. Since $\lambda_1\mathbf{a}_1 = \mathbf{M}\mathbf{a}_1$, it holds that

$$\begin{aligned}\mathbf{M} - \lambda_1\mathbf{a}_1\mathbf{a}_1^T =& \mathbf{M} - \lambda_1\mathbf{a}_1\mathbf{a}_1^T - \lambda_1\mathbf{a}_1\mathbf{a}_1^T + \lambda_1\mathbf{a}_1\mathbf{a}_1^T \\ =& \mathbf{M} - \mathbf{a}_1\mathbf{a}_1^T\mathbf{M} - \mathbf{M}\mathbf{a}_1\mathbf{a}_1^T + \mathbf{a}_1(\mathbf{a}_1^T\mathbf{M}\mathbf{a}_1)\mathbf{a}_1^T \\ =& (\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{M} - (\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{M}\mathbf{a}_1\mathbf{a}_1^T \\ =& (\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{M}(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T).\end{aligned}$$

Hence for any $\|\mathbf{a}\|_2 = 1$, it holds that

$$\begin{aligned}\mathrm{tr}\big\{\mathbf{M}(\mathbf{a}_1\mathbf{a}_1^T - \mathbf{a}\mathbf{a}^T)\big\} =& \mathrm{tr}\big\{\mathbf{M}\mathbf{a}_1\mathbf{a}_1^T\big\} - \mathrm{tr}\{\mathbf{M}\mathbf{a}\mathbf{a}^T\big\} \\ =& \mathrm{tr}\big\{\mathbf{M}\mathbf{a}_1\mathbf{a}_1^T\big\} - \mathrm{tr}\big\{\lambda_1\mathbf{a}_1\mathbf{a}_1^T\mathbf{a}\mathbf{a}^T\big\} - \mathrm{tr}\big\{(\mathbf{M} - \lambda_1\mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\mathbf{a}^T\big\} \\ =& \lambda_1 - \lambda_1(\mathbf{a}^T\mathbf{a}_1)^2 - \mathbf{a}^T(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{M}(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}.\end{aligned}$$

Let $\mathbf{b} = (\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}/\|(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\|_2$, since $\mathbf{b}^T\mathbf{a}_1 = 0$, we have $\mathbf{b} \in \mathcal{M}(\mathbf{a}_2, \cdots, \mathbf{a}_p)$. Since $\sum_{j=1}^p \mathbf{a}_j\mathbf{a}_j^T = \mathbf{I}$, it holds that

$$\sum_{j=2}^p (\mathbf{b}^T\mathbf{a}_j)^2 = \sum_{j=2}^p \frac{(\mathbf{a}^T\mathbf{a}_j)^2}{\|(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\|_2^2} = \frac{\mathbf{a}^T(\sum_{j=2}^p \mathbf{a}_j\mathbf{a}_j^T)\mathbf{a}}{\|(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\|_2^2} = \frac{\mathbf{a}^T(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}}{\|(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\|_2^2} = 1.$$

It then follows that

$$\mathbf{b}^T\mathbf{M}\mathbf{b} = \mathbf{b}^T\sum_{j=1}^p \lambda_j\mathbf{a}_j\mathbf{a}_j^T\mathbf{b} = \sum_{j=2}^p \lambda_j(\mathbf{b}^T\mathbf{a}_j)^2 \le \lambda_2.$$

Hence

$$\mathbf{a}^T(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{M}(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a} \le \lambda_2\|(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1^T)\mathbf{a}\|_2^2 = \lambda_2 - \lambda_2(\mathbf{a}^T\mathbf{a}_1)^2.$$

Substituting $\mathbf{a}$ by $\widehat{\mathbf{a}}_1$, we then obtain

$$\mathrm{tr}\big\{\mathbf{M}(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\} \ge (\lambda_1 - \lambda_2)(1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2).$$

Note that $\widehat{\mathbf{a}}_1 = \arg\max_{\|\mathbf{v}\|_2=1} \mathbf{v}^T\widehat{\mathbf{M}}\mathbf{v}$ subject to $\|\mathbf{v}\|_0 \le s$, since $\|\mathbf{a}_1\| \le s$, it holds that

$$\mathrm{tr}\big\{\widehat{\mathbf{M}}(\widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T - \mathbf{a}_1\mathbf{a}_1^T)\big\} = \widehat{\mathbf{a}}_1^T\widehat{\mathbf{M}}\widehat{\mathbf{a}}_1 - \mathbf{a}_1^T\widehat{\mathbf{M}}\mathbf{a}_1 > 0.$$

Then we have

$$\begin{aligned}1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2 \le& \frac{1}{\lambda_1 - \lambda_2}\Big(\mathrm{tr}\big\{\mathbf{M}(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\} + \mathrm{tr}\big\{\widehat{\mathbf{M}}(\widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T - \mathbf{a}_1\mathbf{a}_1^T)\big\}\Big) \\ =& \frac{1}{\lambda_1 - \lambda_2}\mathrm{tr}\big\{(\mathbf{M} - \widehat{\mathbf{M}})(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\}.\end{aligned}$$

The rest work is to bound $\mathrm{tr}\big\{(\mathbf{M} - \widehat{\mathbf{M}})(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\}$. Let $\mathbf{\Pi}$ be a diagonal matrix with diagonal values being 1 if and only if the corresponding entries in $\mathbf{a}_1$ or $\widehat{\mathbf{a}}_1$ are nonzero.

Then there are at most $2s$ nonzero elements in $\mathbf{\Pi}$. Then $\mathbf{\Pi}\mathbf{a}_1 = \mathbf{a}_1$ and $\mathbf{\Pi}\widehat{\mathbf{a}}_1 = \widehat{\mathbf{a}}_1$. It then holds that

$$\begin{aligned}
&\mathrm{tr}\big\{(\mathbf{M} - \widehat{\mathbf{M}})(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\} \\
=&\mathrm{tr}\big\{(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\mathbf{\Pi}\big\} \\
=&\mathrm{tr}\big\{\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\}.
\end{aligned}$$

For any two $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$, by SVD, we have

$$\begin{aligned}
&\mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{tr}(\mathbf{A}^T\mathbf{U}\mathbf{D}\mathbf{V}^T) = \mathrm{tr}(\mathbf{V}^T\mathbf{A}^T\mathbf{U}\mathbf{D}) = \sum_{i=1}^{p}\sum_{k=1}^{p}(\mathbf{V}^T\mathbf{A}^T\mathbf{U})_{ik}\mathbf{D}_{ki} \\
=&\sum_{i=1}^{p}(\mathbf{V}^T\mathbf{A}^T\mathbf{U})_{ii}\mathbf{D}_{ii} \le \|\mathbf{V}^T\mathbf{A}^T\mathbf{U}\|_2\sum_{i=1}^{p}\mathbf{D}_{ii} = \|\mathbf{A}\|_2\|\mathbf{B}\|_s,
\end{aligned}$$

where $\|\mathbf{B}\|_s$ denotes the sum of singular values of $\mathbf{B}$. Hence

$$\mathrm{tr}\big\{\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\} \le \|\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}\|_2\|\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T\|_s,$$

and

$$\mathrm{tr}\big\{(\mathbf{M} - \widehat{\mathbf{M}})(\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T)\big\} \le \|\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}\|_2 \times 2\sqrt{1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2},$$

due to lemma A.1.1 of Vu and Lei (2012) which shows $\mathbf{a}_1\mathbf{a}_1^T - \widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^T$ has the following singular values: $\sqrt{1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2}, \sqrt{1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2}, 0, 0, \cdots, 0$. Therefore,

$$\begin{aligned}
\sqrt{1 - (\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} \le&\frac{2}{\lambda_1 - \lambda_2}\|\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}\|_2 \\
=&\frac{2}{\lambda_1 - \lambda_2}\sup_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{\Pi}(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{\Pi}\mathbf{x} \\
=&\frac{2}{\lambda_1 - \lambda_2}\sup_{\|\mathbf{x}\|_2=1}\frac{\mathbf{x}^T\mathbf{\Pi}}{\|\mathbf{x}^T\mathbf{\Pi}\|_2}(\mathbf{M} - \widehat{\mathbf{M}})\frac{\mathbf{\Pi}\mathbf{x}}{\|\mathbf{\Pi}\mathbf{x}\|_2}\|\mathbf{\Pi}\mathbf{x}\|_2^2 \\
\le&\frac{2}{\lambda_1 - \lambda_2}\sup_{\|\mathbf{x}\|_2=1}\frac{\mathbf{x}^T\mathbf{\Pi}}{\|\mathbf{x}^T\mathbf{\Pi}\|_2}(\mathbf{M} - \widehat{\mathbf{M}})\frac{\mathbf{\Pi}\mathbf{x}}{\|\mathbf{\Pi}\mathbf{x}\|_2} \\
\le&\frac{2}{\lambda_1 - \lambda_2}\sup_{\|\mathbf{v}\|_2=1\cap\|\mathbf{v}\|_0\le 2s}|\mathbf{v}^T(\mathbf{M} - \widehat{\mathbf{M}})\mathbf{v}|.
\end{aligned}$$

□

**Lemma 3** *Under conditions A1 – A6, let* $\gamma_1^{-1} = r_1^{-1} + 2r_2^{-1}$ *and* $\gamma_2^{-1} = r_1^{-1} + r_2^{-1}$. *As* $n \to \infty$ *and* $p \to \infty$, *for any* $\|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_0 < 2s$ *and* $t > 0$, *it holds that*

$$
\begin{aligned}
&P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{v}| \geq t\Big) \\
\leq& Csp n \exp\Big\{ - C\frac{n^{\gamma_1}}{s^{\gamma_1}} \frac{t^{\gamma_1/2}}{p^{\gamma_1/2}} \Big\} + Csp n \exp\Big\{ - C\frac{n^{\gamma_2}}{s^{\gamma_2/2}} \frac{t^{\gamma_2/4}}{p^{\gamma_2/4}} \Big\} \\
&+ Csp \exp\Big\{ - C\frac{n}{s^2} \frac{t}{p} \Big\} + Csp \exp\Big\{ - C\frac{n}{s} \sqrt{\frac{t}{p}} \Big\} \\
&+ Csp n \exp\Big\{ - C\frac{n^{\gamma_1}}{s^{\gamma_1}} \frac{t^{\gamma_1}}{\lambda_1^{\gamma_1/2} p^{\gamma_1/2}} \Big\} + Csp n \exp\Big\{ - C\frac{n^{\gamma_2}}{s^{\gamma_2/2}} \frac{t^{\gamma_2/2}}{\lambda_1^{\gamma_2/4} p^{\gamma_2/4}} \Big\} \\
&+ Csp \exp\Big\{ - C\frac{n}{s^2} \frac{t^2}{\lambda_1 p} \Big\} + Csp \exp\Big\{ - C\frac{n}{s} \frac{t}{\sqrt{\lambda_1 p}} \Big\}.
\end{aligned}
$$

**Proof.** WLOG, for simplicity, we set $k_0 = 1$. For any $\|\mathbf{v}\|_2 = 1, \|\mathbf{v}\|_0 < 2s$, we have

$$
\begin{aligned}
&|\mathbf{v}^T(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{v}| \\
=&|\mathbf{v}^T \widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) \widehat{\mathbf{\Sigma}}_{\mathbf{y}}^T(1)\mathbf{v} - \mathbf{v}^T \mathbf{\Sigma}_{\mathbf{y}}(1) \mathbf{\Sigma}_{\mathbf{y}}^T(1)\mathbf{v}| \\
=&|\mathbf{v}^T(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) - \mathbf{\Sigma}_{\mathbf{y}}(1))(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^T(1) - \mathbf{\Sigma}_{\mathbf{y}}^T(1))\mathbf{v} + 2\mathbf{v}^T(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) - \mathbf{\Sigma}_{\mathbf{y}}(1))\mathbf{\Sigma}_{\mathbf{y}}^T(1)\mathbf{v}| \\
\leq&\mathbf{v}^T(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) - \mathbf{\Sigma}_{\mathbf{y}}(1))(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^T(1) - \mathbf{\Sigma}_{\mathbf{y}}^T(1))\mathbf{v} \\
&+ 2\sqrt{\mathbf{v}^T(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) - \mathbf{\Sigma}_{\mathbf{y}}(1))(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^T(1) - \mathbf{\Sigma}_{\mathbf{y}}^T(1))\mathbf{v}} \sqrt{\mathbf{v}^T \mathbf{\Sigma}_{\mathbf{y}}(1) \mathbf{\Sigma}_{\mathbf{y}}^T(1)\mathbf{v}} \\
\leq&E + 2\sqrt{E}\sqrt{\lambda_1},
\end{aligned}
\tag{2.7.30}
$$

where we denote $E = \mathbf{v}^T(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1) - \mathbf{\Sigma}_{\mathbf{y}}(1))(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^T(1) - \mathbf{\Sigma}_{\mathbf{y}}^T(1))\mathbf{v}$.

Denote the $(i,j)$-th element of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(1)$ and $\mathbf{\Sigma}_{\mathbf{y}}(1)$ by $\widehat{\sigma}_{i,j}^{(1)}$ and $\sigma_{i,j}^{(1)}$ respectively and $\mathbf{v} = (v_1, \cdots, v_p)^T$. WLOG, we assume the first $2s$ elements of $\mathbf{v}$ are non zeros. It then holds that

$$
E = \Big\| \Big( \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,1}^{(1)} - \sigma_{i,1}^{(1)}), \cdots, \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,p}^{(1)} - \sigma_{i,p}^{(1)}) \Big)^T \Big\|_2^2 = \sum_{j=1}^{p} \Big( \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}) \Big)^2.
$$

For any $\tau > 0$, it holds that

$$\begin{aligned}
P(E \geq \tau) =& P\Big( \sum_{j=1}^{p} \Big( \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}) \Big)^2 \geq \tau \Big) \\
\leq & \sum_{j=1}^{p} P\Big( \Big( \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}) \Big)^2 \geq \frac{\tau}{p} \Big) \\
= & \sum_{j=1}^{p} P\Big( \Big| \sum_{i=1}^{2s} v_i(\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}) \Big| \geq \sqrt{\frac{\tau}{p}} \Big) \\
\leq & \sum_{j=1}^{p} P\Big( \sum_{i=1}^{2s} |v_i| |\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}| \geq \sqrt{\frac{\tau}{p}} \Big) \\
\leq & \sum_{j=1}^{p} P\Big( \sum_{i=1}^{2s} |\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}| \geq \sqrt{\frac{\tau}{p}} \Big) \\
\leq & \sum_{j=1}^{p} \sum_{i=1}^{2s} P\Big( |\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}| \geq \frac{1}{2s} \sqrt{\frac{\tau}{p}} \Big).
\end{aligned}$$

It follows from lemma 9 of Chang, Guo and Yao (2014) that

$$\begin{aligned}
& P\Big( |\widehat{\sigma}_{i,j}^{(1)} - \sigma_{i,j}^{(1)}| \geq \frac{1}{2s} \sqrt{\frac{\tau}{p}} \Big) \\
\leq & Cn \exp\Big\{ - C\Big( \frac{1}{2s} \sqrt{\frac{\tau}{p}} \Big)^{\gamma_1} n^{\gamma_1} \Big\} + Cn \exp\Big\{ - C\Big( \frac{1}{2s} \sqrt{\frac{\tau}{p}} \Big)^{\gamma_2/2} n^{\gamma_2} \Big\} \\
& + C \exp\Big\{ - C\Big( \frac{1}{2s} \sqrt{\frac{\tau}{p}} \Big)^2 n \Big\} + C \exp\Big\{ - C \frac{1}{2s} \sqrt{\frac{\tau}{p}} n \Big\} \\
= & Cn \exp\Big\{ - C \frac{n^{\gamma_1}}{s^{\gamma_1}} \frac{\tau^{\gamma_1/2}}{p^{\gamma_1/2}} \Big\} + Cn \exp\Big\{ - C \frac{n^{\gamma_2}}{s^{\gamma_2/2}} \frac{\tau^{\gamma_2/4}}{p^{\gamma_2/4}} \Big\} \\
& + C \exp\Big\{ - C \frac{n}{s^2} \frac{\tau}{p} \Big\} + C \exp\Big\{ - C \frac{n}{s} \sqrt{\frac{\tau}{p}} \Big\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(E \geq \tau) \leq & Cspn \exp\Big\{ - C \frac{n^{\gamma_1}}{s^{\gamma_1}} \frac{\tau^{\gamma_1/2}}{p^{\gamma_1/2}} \Big\} + Cspn \exp\Big\{ - C \frac{n^{\gamma_2}}{s^{\gamma_2/2}} \frac{\tau^{\gamma_2/4}}{p^{\gamma_2/4}} \Big\} \\
& + Csp \exp\Big\{ - C \frac{n}{s^2} \frac{\tau}{p} \Big\} + Csp \exp\Big\{ - C \frac{n}{s} \sqrt{\frac{\tau}{p}} \Big\}.
\end{aligned} \tag{2.7.31}$$

By (2.7.30) and (2.7.31), for any $t > 0$, we have

$$
\begin{aligned}
&P\Big\{|\mathbf{v}^T(\widehat{\mathbf{M}} - \mathbf{M})\mathbf{v}| \geq t\Big\} \\
\leq &P\Big\{(E + 2\sqrt{E}\sqrt{\lambda_1}) \geq t\Big\} \\
\leq &P\Big\{E \geq \frac{t}{2}\Big\} + P\Big\{E \geq \frac{t^2}{16\lambda_1}\Big\} \\
\leq &Cspn \exp\Big\{-C\frac{n^{\gamma_1}}{s^{\gamma_1}}\frac{t^{\gamma_1/2}}{p^{\gamma_1/2}}\Big\} + Cspn \exp\Big\{-C\frac{n^{\gamma_2}}{s^{\gamma_2/2}}\frac{t^{\gamma_2/4}}{p^{\gamma_2/4}}\Big\} \\
&+ Csp \exp\Big\{-C\frac{n}{s^2}\frac{t}{p}\Big\} + Csp \exp\Big\{-C\frac{n}{s}\sqrt{\frac{t}{p}}\Big\} \\
&+ Cspn \exp\Big\{-C\frac{n^{\gamma_1}}{s^{\gamma_1}}\frac{t^{\gamma_1}}{\lambda_1^{\gamma_1/2}p^{\gamma_1/2}}\Big\} + Cspn \exp\Big\{-C\frac{n^{\gamma_2}}{s^{\gamma_2/2}}\frac{t^{\gamma_2/2}}{\lambda_1^{\gamma_2/4}p^{\gamma_2/4}}\Big\} \\
&+ Csp \exp\Big\{-C\frac{n}{s^2}\frac{t^2}{\lambda_1 p}\Big\} + Csp \exp\Big\{-C\frac{n}{s}\frac{t}{\sqrt{\lambda_1 p}}\Big\}.
\end{aligned}
$$

□

**Proof of Theorem 7.** Let $\mathbb{S}^{p-1}$ be the set of $p$-dimensional unit vector and $\mathbb{B}(s)$ be the set such that all elements satisfies $\|\mathbf{x}\|_0 < s$. Let $\mathbb{K}$ be a fixed subsect $\mathbb{K} \subset \{1, 2, \cdots, p\}$ with $|\mathbb{K}| = 2s$, for example, $\mathbb{K} = \{1, 2, \cdots, 2s\}$. Define

$$\mathbb{B}_{\mathbb{K}} = \{\mathbf{v}| \quad \text{for any} \quad i \in \{1, 2, \cdots, p\}/\mathbb{K}, v_i = 0\}.$$

In order to proceed, we need the following result: An $\epsilon$-net $\mathbb{N}_\epsilon$ of a sphere $\mathbb{S}^{p-1}$ is a subset of $\mathbb{S}^{p-1}$ such that for any $\mathbf{v} \in \mathbb{S}^{p-1}$, there exists $\mathbf{u} \in \mathbb{N}_\epsilon$ subject to $\|\mathbf{u} - \mathbf{v}\| \leq \epsilon$. Two existed results we will use are (1) for any $\epsilon > 0$, it holds that $|\mathbb{N}_\epsilon| \leq (1 + 2/\epsilon)^p$. (2) for any $p \times p$ matrix $\mathbf{A}$ and $\epsilon \in (0, 1/2)$, it holds that $\sup_{\mathbf{v}_1 \in \mathbb{S}^{p-1}} |\mathbf{v}_1^T \mathbf{A} \mathbf{v}_1| \leq (1 - 2\epsilon)^{-1} \sup_{\mathbf{v}_2 \in \mathbb{N}_\epsilon} |\mathbf{v}_2^T \mathbf{A} \mathbf{v}_2|$.

Now we can go on with the original proof.

Let the $\frac{1}{4}$-net of $\mathbb{S}^{p-1} \cap \mathbb{B}_{\mathbb{K}}$ be $\mathbb{N}_{\mathbb{K}}$, for any $t > 0$, according to the above results (1) and (2),

we have

$$
\begin{aligned}
&P\Big(\sup_{\mathbf{v}\in\mathbb{S}^{p-1}\cap\mathbb{B}_{\mathbb{K}}}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t\Big)\\
\leq&P\Big((1-2\times 1/4)^{-1}\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t\Big)\\
=&P\Big(\bigcup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}\Big\{|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big\}\Big)\\
=&\sum_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big)\\
\leq&|\mathbb{N}_{\mathbb{K}}|\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big)\\
\leq&(1+2/(1/4))^{2s}\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big)\\
=&9^{2s}\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big),
\end{aligned}
$$

where the $2s$ (instead of $p$) in the last inequality is because for a fixed $\mathbb{K}\subset\{1,2,\cdots,p\}$ with $|\mathbb{K}|=2s$, $\mathbb{S}^{p-1}\cap\mathbb{B}_{\mathbb{K}}$ is equivalent to a subset of $\mathbb{S}^{2s-1}$, hence we can employ the $\epsilon$-net arguments on such a subset.

Now we allow for arbitrage subset $\mathbb{K}\subset\{1,2,\cdots,p\}$ with $|\mathbb{K}|=2s$, it then follows that

$$
\begin{aligned}
&P\big(\sup_{\mathbf{v}\in\mathbb{S}^{p-1}\cap\mathbb{B}_0(2s)}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t\Big)\\
\leq&\sum_{\mathbb{K}\subset\{1,2,\cdots,p\}}P\big(\sup_{\mathbf{v}\in\mathbb{S}^{p-1}\cap\mathbb{B}_{\mathbb{K}}}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t\Big)\\
\leq&\begin{pmatrix}p\\2s\end{pmatrix}9^{2s}\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}}P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|\geq t/2\Big).
\end{aligned}
\tag{2.7.32}
$$

Therefore, by lemma 2, lemma 3 and (2.7.32), we have

$$
\begin{aligned}
&P\Big(\sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} \geq \frac{2t}{\lambda_1-\lambda_2}\Big) \\
\leq & P\Big(2\frac{2}{\lambda_1-\lambda_2}\sup_{\mathbf{v}\in\mathbb{S}^{p-1}\cap\mathbb{B}_0(2s)}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}| \geq \frac{2t}{\lambda_1-\lambda_2}\Big) \\
= & P\Big(\sup_{\mathbf{v}\in\mathbb{S}^{p-1}\cap\mathbb{B}_0(2s)}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}| \geq t\Big) \\
\leq & \begin{pmatrix} p \\ 2s \end{pmatrix} 9^{2s}\sup_{\mathbf{v}\in\mathbb{N}_{\mathbb{K}}} P\Big(|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}| \geq t/2\Big) \\
\leq & \begin{pmatrix} p \\ 2s \end{pmatrix} 9^{2s}\Big(Cspn\exp\Big\{-C\frac{n^{\gamma_1}}{s^{\gamma_1}}\frac{t^{\gamma_1/2}}{p^{\gamma_1/2}}\Big\} + Cspn\exp\Big\{-C\frac{n^{\gamma_2}}{s^{\gamma_2/2}}\frac{t^{\gamma_2/4}}{p^{\gamma_2/4}}\Big\} \\
& + Csp\exp\Big\{-C\frac{n}{s^2}\frac{t}{p}\Big\} + Csp\exp\Big\{-C\frac{n}{s}\sqrt{\frac{t}{p}}\Big\} \\
& + Cspn\exp\Big\{-C\frac{n^{\gamma_1}}{s^{\gamma_1}}\frac{t^{\gamma_1}}{\lambda_1^{\gamma_1/2}p^{\gamma_1/2}}\Big\} + Cspn\exp\Big\{-C\frac{n^{\gamma_2}}{s^{\gamma_2/2}}\frac{t^{\gamma_2/2}}{\lambda_1^{\gamma_2/4}p^{\gamma_2/4}}\Big\} \\
& + Csp\exp\Big\{-C\frac{n}{s^2}\frac{t^2}{\lambda_1 p}\Big\} + Csp\exp\Big\{-C\frac{n}{s}\frac{t}{\sqrt{\lambda_1 p}}\Big\}\Big).
\end{aligned}
$$

For a sufficient large constant $M > 0$, let

$$
t = M{\lambda_1}^{1/2}\sqrt{\frac{s^3 p\log p}{n}},
$$

(note that the $s$ in $s^3$ comes from bounding $\begin{pmatrix} p \\ 2s \end{pmatrix} 9^{2s}$ and $s^2$ in $s^3$ comes from the exponential terms). We then have

$$
\sqrt{1-(\widehat{\mathbf{a}}_1^T\mathbf{a}_1)^2} = O_p\Big(\frac{{\lambda_1}^{1/2}}{\lambda_1-\lambda_2}\sqrt{\frac{s^3 p\log p}{n}}\Big).
$$

□

**Lemma 4** *The estimator* $\widehat{\mathbf{A}}$ *in (2.2.23) of* $\mathbf{A}$ *satisfies*

$$
\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F \leq \frac{2\sqrt{r}\sup_{\|\mathbf{v}\|_2=1\cap\|\mathbf{v}\|_0\leq 2rs}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|}{\lambda_r}.
$$

**Proof.** It is straightforward by employing lemma 5, lemma 6 and lemma 7 of Wang, Han and Liu (2013). Note that the assumption $\|\mathbf{A}\|_0 \leq s$ is used in the proof of lemma 6 in Wang, Han and Liu (2013). □

**Proof of Theorem 8.** Since $\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F \leq \frac{2\sqrt{r}\sup_{\|\mathbf{v}\|_2=1\cap\|\mathbf{v}\|_0\leq 2rs}|\mathbf{v}^T(\widehat{\mathbf{M}}-\mathbf{M})\mathbf{v}|}{\lambda_r}$, replacing $s$ by $rs$ in lemma 3 and Theorem 7 and following the same arguments as above, we have

$$\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T(\mathbf{I}-\mathbf{A}\mathbf{A}^T)\|_F = O_p\Big(\frac{{\lambda_1}^{1/2}}{\lambda_r}\sqrt{\frac{s^3 p\log p}{n}}\Big).$$

□

# Chapter 3: Group Lasso for Covariance Matrix Break Detection

## 3.1 Introduction

Detecting multiple change points in univariate time series has been widely discussed, see Chen and Gupta (1997), Davis et al. (2006) and Davis et al. (2008) for example. The second order nonstationarities observed in large panel of asset returns (see Fan et al. (2011)) implies the importance of the detection of change points of the second order structure of multivariate time series. Vert and Bleakley (2010) describe other interesting examples of multivariate, nonstationary time series in many other fields, such as signal processing, biology and medicine. Current attempts on the detection of second order structure change include Cho and Fryzlewicz (2015). They considered a piecewise stationary, multivariate time series with a time varying second order structure, where the autocovariance and cross-covariance functions are asymptotically piecewise constant and hence the time series is approximately stationary between change-points in these functions. They proposed a CUSUM-based binary segmentation method for the multiple change-points case. Based on the classical CUMSUM test, Aue, Hörmann, Horváth and Reimherr (AHHR) (2009)

proposed a nonparametric method to assess the stability of volatilities and cross-volatilites of linear and nonlinear multivariate time series models, but only for a single change point.

We attempted to detect multiple change points of general multivariate time sereis, that is, unlike Aue, Hörmann, Horváth and Reimherr (AHHR) (2009), we allow more than one or even diverging number of change points and unlike Cho and Fryzlewicz (2015), we do not consider any specific models. By reformulating the problem in a variable selection context, the group least absolute shrinkage and selection operator (LASSO) is proposed to estimate the locations of the change points. Our method is model-free, it can be extensively applied to multivariate time series, such as GARCH and stochastic volatility models. It is shown that the locations of the change points can be consistently estimated by the group LASSO procedure when we have the knowledge of the number of change points, and the computation can be efficiently performed. However, the number of the change point is unknown in practice and it can be shown that the group LASSO procedure will over-estimate the number of the change points most times. Hence an improved practical version that incorporates group LASSO and the stepwise regression variable selection technique are discussed. The two-step procedure can consistently estimate both the number of change points and the locations of the change points.

The rest of the paper is organized as follows. Section 3.2 introduces the model and the two-step estimation method. The asymptotic theory for the proposed estimation method is presented in Section 3.3. Simulation results are reported in Section 3.4. A short discussion of future work is presented in Section 3.5. All the technical proofs are relegated to an Appendix.

## 3.2 Problem and Estimation Method

### 3.2.1 Problem

Let $\mathbf{y}_t = (y_{t1}, \cdots, y_{tp})^T$ be an observable $p \times 1$ vector time series process with mean zero and covariance matrix $\mathrm{Cov}(\mathbf{y}_t) = \mathbf{\Sigma}_t \equiv (\sigma_{ijt})_{p\times p}$. Our interest in this paper is to estimate the following multiple changes of the covariance structure $\mathbf{\Sigma}_t$:

$$\mathbf{\Sigma}_t I(t_{i-1} \leq t < t_i) = \mathbf{\Sigma}_{t_{i-1}},\ i = 1, \cdots, m+1, \tag{3.2.33}$$

where $1 = t_0 < t_1 < \cdots < t_{m+1} = n+1$ and $\mathbf{\Sigma}_{t_{i-1}} \neq \mathbf{\Sigma}_{t_i}$ for $i = 1, 2, \cdots, m$. When $p$ is fixed and $m = 1$, the above question is discussed by Aue, Hörmann, Horváth and Reimherr (AHHR) (2009) by based on classical CUMSUM test. Recently, Cho and Fryzlewicz (2015) proposed a CUSUM-based binary segmentation method for the multiple change-points case, but they assume that each of the components of $\mathbf{y}_t$ follows a piecewise stochastic volatility model, i.e., for each component $y_{ti}$ of $\mathbf{y}_t$,

$$y_{ti} = \sigma_i(t/n) Z_{ti}^2,\ t = 1, \cdots, n,\ i = 1, 2, \cdots, p, \tag{3.2.34}$$

where $n$ is the sample size, $\sigma_i(t/n)$ is a piecewise constant function and $Z_{ti}$ is a sequence of standard normal variables. The purpose of this paper is to propose a model-free and efficient algorithm for the estimation of the change points in (3.2.33) with big $m$ and possibly diverging with $n$.

### 3.2.2 One-step Estimation

For any matrix $\mathbf{A} = (a_{ij})_{p\times p}$, define

$$\mathrm{vec}(\mathbf{A}) = (a_{11}, \cdots, a_{1p}, a_{21}, \cdots, a_{2p}, a_{31}, \cdots, a_{3p}, \cdots, a_{p1}, \cdots, a_{pp})^T,$$

that is, the vector consists of all the elements of the matrix. Let

$$\boldsymbol{\mu}_i = \text{vec}(\boldsymbol{\Sigma}_t), \quad \text{if} \quad t_{i-1} \leq t < t_i, i = 1, \cdots, m+1.$$

Then the detection of the covariance structure in (3.2.33) is equivalent to identifying the change-points $(t_1, \cdots, t_m)$. Denote $\mathbf{x}_t = \text{vec}(\mathbf{y}_t \mathbf{y}_t')$. Since $\text{E}(\mathbf{y}_t \mathbf{y}_t') = \boldsymbol{\Sigma}_t$, we can see the change point detection problem (3.2.33) as the multiple change-points in mean of the following model:

$$\mathbf{x}_t = \sum_{i=1}^{m+1} \{\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_t\} \mathbb{I}(t_{i-1} \leq t < t_i). \tag{3.2.35}$$

Thus, we can estimate the change-points via group Lasso procedure as in Chan, Yau and Zhang (2014), see also Harchaoui and Lévy-Leduc (2010). Specifically, let $\mathbf{x}(n) = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_n^T)^T$, $\boldsymbol{\varepsilon}(n) = (\boldsymbol{\varepsilon}_1^T, \cdots, \boldsymbol{\varepsilon}_n^T)^T$, $\boldsymbol{\theta}(n) = (\boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_n^T)^T$ and $\mathbf{Q}$ be an $np^2 \times np^2$ matrix defined by

$$\mathbf{Q} = \begin{pmatrix} \mathbf{I} & 0 & 0 & \dots & 0 \\ \mathbf{I} & \mathbf{I} & 0 & \dots & 0 \\ \mathbf{I} & \mathbf{I} & \mathbf{I} & \dots & 0 \\ \vdots & & & & \\ \mathbf{I} & \mathbf{I} & \mathbf{I} & \dots & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is the $p^2 \times p^2$ identity matrix. Set $\boldsymbol{\theta}_1 = \boldsymbol{\mu}_1$ and

$$\boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\mu}_{j+1} - \boldsymbol{\mu}_j\,, & \text{when } i = t_j, \text{ where } t_j \text{ is a changepoint in (3.2.35)}\,, \\ \mathbf{0}\,, & \text{otherwise}\,, \end{cases}$$

for $i = 2, \ldots, n$. Throughout this paper, for a vector $\boldsymbol{\theta}$, the notations $\boldsymbol{\theta} = \mathbf{0}$ and $\boldsymbol{\theta} \neq 0$ mean that $\boldsymbol{\theta}$ has all entries zero and has at least one non-zero entry, respectively. It can be seen that model (3.2.35) can be expressed as a high dimensional regression model

$$\mathbf{x}(n) = \mathbf{Q}\boldsymbol{\theta}(n) + \boldsymbol{\varepsilon}(n)\,. \tag{3.2.36}$$

Since only $m+1$ of the vectors $\boldsymbol{\theta}_i$s in $\boldsymbol{\theta}(n)$ are non-zero, we look for a sparse solution to the high dimension regression model (3.2.36). A well-known solution to this problem is given by the group lasso estimation (Yuan and Lin (2006)). Thus, we propose to estimate $\boldsymbol{\theta}(n)$ by the following group LASSO equation:

$$\widehat{\boldsymbol{\theta}}(n) = \text{argmin}_{\boldsymbol{\theta}(n)} \frac{1}{n}||\mathbf{x}(n) - \mathbf{Q}\boldsymbol{\theta}(n)||^2 + \lambda_n \sum_{i=1}^{n} ||\boldsymbol{\theta}_i||\,, \tag{3.2.37}$$

where $\lambda_n > 0$ is the regularization parameter. Note that when $\widehat{\boldsymbol{\theta}}_i \neq 0$, $i \geq 2$, there is a change point at time $i$. Thus the structural breaks $t_j$, $j = 1, 2, \ldots, m$ can be estimated by identifying those $\widehat{\boldsymbol{\theta}}_i, (i \geq 2)$ which are not zero. We denote the estimates of the change points by $\mathcal{A}_n = \{t \geq 2 : \widehat{\boldsymbol{\theta}}_t \neq 0\} \equiv \{\widehat{t}_1, \cdots, \widehat{t}_{|\mathcal{A}_n|}\}$.

### 3.2.3 Two-step estimation procedure

Using the GLASSO procedure for estimating the number of change points, which is usually larger than the true number of change points, see Theorem 10 below. Two immediate issues arise: (i) how to estimate the true number of breaks, and (ii) how to estimate the change points with a nearly optimal rate? These two issues are dealt with in this subsection.

However, it is known that with probability tending to 1, all the true change points can be identified within a $n\gamma_n$ neighborhood, see Theorem 10 below. Therefore, the change-points can be consistently estimated and are identified within $\mathcal{A}_n$. One way to achieve this mission is to choose the "best possible subset" of change points in $\mathcal{A}_n$ according to some prescribed information criterion ($IC$). Given any $m$ and the change points $\mathbf{t} = (t_1, \ldots, t_m)$, an information criterion $IC(m, \mathbf{t})$ typically consists of a sum of a goodness-of-fit measure and a penalty term that accounts for the model complexity. Specifically, let $\widehat{\widehat{\boldsymbol{\mu}}}_j = (t_j - t_{j-1})^{-1} \sum_{t=t_{j-1}}^{t_j-1} \mathbf{x}_t$ be the least squares estimator and $S_n(t_{j-1}, t_j) = \sum_{t=t_{j-1}}^{t_j-1} \|\mathbf{x}_t - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2$ be

the residual sum of squares from time $t_{j-1}$ to $t_j - 1$. Consider a general information criterion of the form

$$IC(m, \mathbf{t}) = S_n(t_1, t_2, \ldots, t_m) + m\omega_n \,, \tag{3.2.38}$$

where the least squares criterion $S_n(t_1, t_2, \ldots, t_m) = \sum_{j=1}^{m+1} S_n(t_{j-1}, t_j)$ is the goodness-of-fit measure and $\omega_n$ is the penalty term. We estimate the number and locations of the change points by solving

$$(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{t}}}) = \arg \min_{\substack{m \in (0,1,\ldots,|\mathcal{A}_n|), \\ \mathbf{t}=(t_1,\ldots,t_m) \subset \mathcal{A}_n}} IC(m, \mathbf{t}) \,, \tag{3.2.39}$$

To achieve further computational efficiency, we adopt the following backward elimination algorithm (BEA) numerically. BEA starts with the set of change points $\mathcal{A}_n$, then removes the "most redundant" change points that corresponds to the largest reduction of $IC$ until no further removal is possible. The estimator $\mathcal{A}_n^* \equiv (\widehat{t}_1^*, \cdots, \widehat{t}_{|\mathcal{A}_n^*|}^*)$ is obtained as follows:

(1) Set $K = |\mathcal{A}_n|, \mathbf{t}_K \equiv \mathcal{A}_n = (t_{K,1}, \cdots, t_{K,K})$ and $V_K^* = IC(K, \mathcal{A}_n)$.

(2) For $i = 1, \cdots, K$, compute $V_{K,i} = IC(K-1, \mathbf{t}_K / \{t_{K,i}\})$. Set $V_{K-1}^* = \min_i V_{K,i}$.

(3) If $V_{K-1}^* > V_K^*$, then the estimated locations of change points are $\mathcal{A}_n^* = \mathbf{t}_K$.

If $V_{K-1}^* \leq V_K^*$ and $K = 1$, then $\mathcal{A}_n^* = \emptyset$.

If $V_{K-1}^* \leq V_K^*$ and $K > 1$, then set $j = \arg\min_i V_{K,i}, \mathbf{t}_{K-1} \equiv \mathbf{t}_K / \{t_{K-1,j}\}$ and $K = K - 1$. Then go to step 2.

## 3.3 Theoretical Properties

We introduce some notations first. Let $\mathcal{A} = \{t_i^0, i = 1, \cdots, m_0\}$ be the set of true change points and $\boldsymbol{\mu}_j^0$ be the true mean vector in the $j$-th segment, $j = 1, \cdots, m_0 + 1$. For a set $A$, we use $|A|$ to denote its cardinality. A strictly stationary process $\{\mathbf{y}_t\}$ is $\alpha$-mixing if

$$\alpha(k) \equiv \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} \big|P(A)P(B) - P(AB)\big| \to 0, \quad \text{as } k \to \infty, \tag{3.3.40}$$

where $\mathcal{F}_i^j$ denotes the $\sigma$-algebra generated by $\{\mathbf{y}_t, i \le t \le j\}$. See, e.g., Section 2.6 of Fan and Yao (2003) for a compact review of $\alpha$-mixing processes. Some regularity conditions are now in order.

A1. The process $\mathbf{y}_t$ is strictly stationary in each regime $[t_{i-1}^0, t_i^0), i = 1, \cdots, m^0 + 1$ with mixing coefficient $\alpha^{(i)}(k)$ defined in (3.3.40) and there exist a positive constant $\gamma_1$ and a positive $c$ such that

$$\alpha^{(i)}(k) \le \exp(-cn^{\gamma_1}),$$

for any positive integer $k$ and $i = 1, \cdots, m^0 + 1$.

A2. For any positive $z$, there exists positive constant $\gamma_2$ such that

$$\sup_{1 \le i \le p^2} \sup_{t > 0} P(|x_{ti} - Ex_{ti}| > z) \le \exp(1 - z^{\gamma_2}),$$

where $x_{ti}$ is the $i$-th element of $\mathbf{x}_t$ and suppose furthermore that $\gamma < 1$ where $\gamma$ is defined by $1/\gamma = 1/\gamma_1 + 1/\gamma_2$.

A3. Assume $\min_{1 \le i \le m_0+1} \|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i-1}^0\| > \nu$ for some $\nu > 0$. As $n \to \infty$, $\min_{1 \le i \le m_0+1} |t_i^0 - t_{i-1}^0|/n\gamma_n \to \infty$ for some $\gamma_n \to 0$ satisfying $\frac{(\log n)^{1/\gamma}}{n} = o(\gamma_n)$ and $\gamma_n/\lambda_n \to \infty$, where $\gamma$ is defined in A2 and $\lambda_n$ is the tuning parameter in (3.2.37).

Condition A1 and A2 allow us to obtain the large deviation result for $\alpha$ mixing process. To ensure a change occur at $t_i^0$, $\min_{1\le i\le m_0+1} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}\| > \nu$ is imposed in A3. The sample size $t_i^0 - t_{i-1}^0$ for each segment should go to infinity in order to study the asymptotic properties. A3 allows $\min_{1\le i\le m_0+1} |t_i^0 - t_{i-1}^0|$ larger than $n\gamma_n$, which enlarges the assumption that $\min_{1\le i\le m_0+1} |t_i^0 - t_{i-1}^0| > Cn$ in most literatures. This allows the fact that $m_0$ can be divergent. Now we are ready to present the theoretical results.

Theorem 9 is about the consistency result for the estimate of change points when the number of change points $m_0$ is known.

**Theorem 9** *Under assumptions A1, A2 and A3, if $m_0$ is known and $|\mathcal{A}_n| = m_0$, then as $n \to \infty$,*

$$P\{\max_{1\le i\le m_0} |\widehat{t}_i - t_i^0| \le n\gamma_n\} \to 1.$$

In practice, $m_0$ is not known. Theorem 10 shows the consistency result of the estimator $\mathcal{A}_n$ obtained from the one-step group LASSO procedure and the fact that the number of change points estimated from this step is usually larger than $m_0$. Define the Hausdorff distance between two sets $A$ and $B$ as in Boysen et. al. (2009) by

$$d_H(A, B) = \max_{b\in B} \min_{a\in A} |b - a|,$$

and $d_H(A, \emptyset) = d_H(\emptyset, B) = 1$, where $\emptyset$ is the empty set.

**Theorem 10** *Under assumptions A1, A2 and A3, as $n \to \infty$, we have*

$$P(|\mathcal{A}_n| \ge m_0) \to 1,$$

*and*

$$P(d_H(\mathcal{A}_n, \mathcal{A}) \le n\gamma_n) \to 1.$$

Though the one-step procedure overestimate $m_0$, the estimates of the locations are within a $n\gamma_n$ neighborhood of the true change points, which motivates the second step estimation. Theorem 11 gives the consistency result of the estimator $(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{t}}})$ obtained from the two-step estimation procedure.

**Theorem 11** *Suppose* $\omega_n$ *in the information criteria* (3.2.38) *satisfies* $\lim_{n\to\infty} n\gamma_n m_0/\omega_n = 0$ *and* $\lim_{n\to\infty} \omega_n / \min_{1\le i\le m_0} |t_i^0 - t_{i-1}^0| = 0$, *then under conditions A1, A2 and A3, as* $n \to \infty$, *the minimizer* $(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{t}}})$ *of* (3.2.39) *satisfies*

$$P(\widehat{\widehat{m}} = m_0) \to 1,$$

*and there exists a constant* $B > 0$ *such that*

$$P(\max_{1\le i\le m_0} |\widehat{\widehat{t}}_i - t_i^0| \le Bn\gamma_n) \to 1.$$

Theorem 12 gives the consistency result of the estimator $\mathcal{A}_n^*$ obtained from the backward elimination algorithm.

**Theorem 12** *Under conditions of Theorem 11, as* $n \to \infty$, *the estimator* $\mathcal{A}_n^*$ *obtained from BEA satisfies*

$$P(|\mathcal{A}_n^*| = m_0) \to 1,$$

*and there exists a constant* $B > 0$ *such that*

$$P(\max_{1\le i\le m_0} |\widehat{t}_i^{\,*} - t_i^0| \le Bn\gamma_n) \to 1.$$

## 3.4 Simulation Studies

To examine the finite sample performance of the proposed estimation methods, we conduct some simulations under different scenarios. We used the group LARS algorithm for the

first step and backward elimination algorithm for the second step. The maximum number of change points used in the group LARS algorithm step is set to be 20 for all scenarios. The penalty term $\omega_n$ in the second step is specified as $C\log(n)$, where $C$ can be chosen via cross validation.

### 3.4.1 Scenario 1

Let the components of $\mathbf{y}_t = (y_{t1}, \cdots, y_{tp})^T$ be $p$ realizations from AR(1) process. That is $y_{ti} = \alpha y_{t,i-1} + \varepsilon_t, i = 2, \cdots, p$. Let $n = 2048$, the first and second breaks are chosen at $t_1 = 513$ and $t_2 = 1537$. We generate the $p$ components of $\mathbf{y}_t$ for each $t$ from AR(1) with coefficient 0.7 if $1 \leq t \leq 512$, AR(1) with coefficient -0.6 if $513 \leq t \leq 1536$ and AR(1) with coefficient 0.8 if $1537 \leq t \leq 2048$, where $\varepsilon_t$ are independently generated from $N(0, 0.5)$.

We apply the proposed two step estimation method for 200 times and report the simulation times of correctly estimated number of change points, the mean and standard error of the relative location estimator with $p$ equals to 2, 5, and 10.

| | Mean (SE) of 1st break | Mean (SE) of 2nd break | times/200 of $\widehat{m} = 2$ |
|---|---|---|---|
| p=2 | 0.243 (0.018) | 0.752 (0.018) | 190/200 |
| p=5 | 0.247 (0.014) | 0.753 (0.013) | 200/200 |
| p=10 | 0.249 (0.011) | 0.752 (0.008) | 200/200 |

Table 3.3: Estimated break points from two step estimation method for scenario 1.

Table 3.3 shows that the mean of the estimated 1st and 2nd relative change point position is very close to the true relative position 0.25 and 0.75. The standard deviations are small as well. And the number of change points can correctly estimated most times.

### 3.4.2 Scenario 2

Let $n = 2048$, the first and second breaks are chosen at $t_1 = 513$ and $t_2 = 1537$. Firstly, we generate $n$ $p \times 1$ random vectors $\mathbf{z}_t$ from $p$ dimensional standard normal distribution $N(0, \mathbf{I}_p)$. Put $\mathbf{U}_1 = \mathbf{I}_p$ and $\mathbf{D}_1$ is a $p \times p$ diagonal matrix which the diagonal elements are generated from Uniform(1, 2), $\mathbf{U}_2$ equals to the Q part of the QR decomposition of a $p \times p$ matrix whose elements are generated from $N(0, 1)$ and $\mathbf{D}_2$ is a $p \times p$ diagonal matrix whose diagonal elements are generated from Uniform(0, 0.5), $\mathbf{U}_3$ is generated in the same process as $\mathbf{U}_2$ and and $\mathbf{D}_3$ is a $p \times p$ diagonal matrix whose diagonal elements are generated from Uniform(4, 5). The time series we obtain is $\mathbf{y}_t = \mathbf{U}_1\mathbf{D}_1^{1/2}\mathbf{z}_t$ if $1 \leq t \leq 512$, $\mathbf{y}_t = \mathbf{U}_2\mathbf{D}_2^{1/2}\mathbf{z}_t$ if $513 \leq t \leq 1536$ and $\mathbf{y}_t = \mathbf{U}_3\mathbf{D}_3^{1/2}\mathbf{z}_t$ if $1537 \leq t \leq 2048$.

We apply the proposed two step estimation method for 200 times and report the simulation times of correctly estimated number of change points, the mean and standard error of the relative location estimator with $p$ equals to 2, 5, and 10. Table 3.4 shows that the

| | Mean (SE) of 1st break | Mean (SE) of 2nd break | times/200 of $\widehat{m} = 2$ |
|---|---|---|---|
| p=2 | 0.248 (0.015) | 0.751 (0.002) | 192/200 |
| p=5 | 0.248 (0.012) | 0.751 (0.001) | 194/200 |
| p=10 | 0.249 (0.009) | 0.751 (0.001) | 195/200 |

Table 3.4: Estimated break points from two step estimation method for scenario 2.

mean of the estimated 1st and 2nd relative change point position is very close to the true relative position 0.25 and 0.75. The standard deviations are small as well. And the number of change points can correctly estimated most times.

### 3.4.3 Scenario 3

Let $n = 2048$, the first and second breaks are chosen at $t_1 = 513$ and $t_2 = 1537$. We generate $\mathbf{y}_t$ from a one factor model $\mathbf{y}_t = \mathbf{A}x_t + \boldsymbol{\varepsilon}_t$. We generate $x_t$ from a AR(1) process with coefficient 0.4 with $N(0,1)$ distributed noise term. Let $\mathbf{A}$ be a $p \times 1$ vector with $2\cos(2\pi i/p), i = 1, \cdots, p$ as its $i$-th element if $1 \leq t \leq 512$, $\mathbf{A}$ with all elements to be 0.2 if $513 \leq t \leq 1536$ and a $p \times 1$ vector with $3\cos(2\pi i/p), i = 1, \cdots, p$ as its $i$-th element if $1537 \leq t \leq 2048$. $\boldsymbol{\varepsilon}_t$ are generated from $N(0, \mathbf{I}_p)$.

We apply the proposed two step estimation method for 200 times and report the simulation times of correctly estimated number of change points, the mean and standard error of the relative location estimator with $p$ equals to 2, 5, and 10.

| | Mean (SE) of 1st break | Mean (SE) of 2nd break | times/200 of $\widehat{m} = 2$ |
|---|---|---|---|
| p=2 | 0.245 (0.024) | 0.753 (0.005) | 187/200 |
| p=5 | 0.245 (0.025) | 0.753 (0.004) | 188/200 |
| p=10 | 0.249 (0.020) | 0.754 (0.006) | 185/200 |

Table 3.5: Estimated break points from two step estimation method for scenario 3.

Table 3.5 shows that the mean of the estimated 1st and 2nd relative change point position is very close to the true relative position 0.25 and 0.75. The standard deviations are small as well. And the number of change points can correctly estimated most times, even though not as satisfied as scenario 1 and 2.

## 3.5 Future work

Currently, the asymptotic results are established when the dimensionality $p$ is fixed. A more challenging problem is the detection of multiple change points when $p$ goes to infinity as the sample size $n$ goes to infinity or even larger than $n$. Intuitively, when $p$ is large, we have more parameters to estimate, hence we will obtain less accurate estimators, which might influence the estimation of breaks. The future work will focus on the asymptotic results when $p$ can be divergent. If the convergent rate is bad, then future work becomes how to improve the estimation. Indeed, the asymptotic results for the first group lasso procedure has been obtained.

## 3.6 Appendix: Proofs

**Lemma 5** *Let $\widehat{\boldsymbol{\theta}}(n)$ be defined as in* (3.2.37)*, we have*

$$\sum_{l=j}^{n}\sum_{k=1}^{l}\widehat{\boldsymbol{\theta}}_k - \sum_{l=j}^{n}\mathbf{x}_l + \frac{1}{2}n\lambda_n\frac{\widehat{\boldsymbol{\theta}}_j}{\|\widehat{\boldsymbol{\theta}}_j\|} = \mathbf{0}, \quad \textit{if} \quad \widehat{\boldsymbol{\theta}}_j \neq 0, \tag{3.6.41}$$

$$\textit{and} \quad \|\sum_{l=j}^{n}\sum_{k=1}^{l}\widehat{\boldsymbol{\theta}}_k - \sum_{l=j}^{n}\mathbf{x}_l\| \leq \frac{1}{2}n\lambda_n \quad \textit{for all } j. \tag{3.6.42}$$

**Proof.** By proposition 1 in Yuan and Lin (2006), we know that a necessary and sufficient condition for $\widehat{\boldsymbol{\theta}}(n)$ to be a solution of (3.2.37) is

$$-\mathbf{Q}_j^T(\mathbf{x}(n) - \mathbf{Q}\widehat{\boldsymbol{\theta}}(n)) + \frac{1}{2}n\lambda_n\frac{\widehat{\boldsymbol{\theta}}_j}{\|\widehat{\boldsymbol{\theta}}_j\|} = \mathbf{0}, \quad \text{if} \quad \widehat{\boldsymbol{\theta}}_j \neq 0,$$

$$\text{and} \quad \|-\mathbf{Q}_j^T(\mathbf{x}(n) - \mathbf{Q}\widehat{\boldsymbol{\theta}}(n))\| \leq \frac{1}{2}n\lambda_n \quad \text{for all } j,$$

where $\mathbf{Q}_j$ is the $(j-1)p$-th to $jp$-th columns. For each $j = 1, \cdots, n$,

$$\mathbf{Q}_j^T\mathbf{x}(n) = \sum_{l=j}^{n}\mathbf{x}_l, \quad \mathbf{Q}_j^T\mathbf{Q}\widehat{\boldsymbol{\theta}}(n) = \sum_{l=j}^{n}\sum_{k=1}^{l}\widehat{\boldsymbol{\theta}}_k,$$

the required result is then obtained. □

**Lemma 6** *Under conditions A1 and A2, for any positive $a_n$ and $x$, for $1 \le i \le m_0+1$, there exist positive constants $C_1, C_2, C_3, C_4$ and $C_5$ depending only on $c, \gamma$ and $\gamma_1$ such that*

$$\begin{aligned}
&P\Big(\max_{|t-s|>a_n, t_{i-1}^0 \le s<t\le t_i^0} \max_{1\le j\le p^2} |\frac{1}{t-s}\sum_{l=s}^{t}\varepsilon_{lj}| \ge x\Big)\\
\le& p^2n^3\exp\Big(-\frac{a_n^\gamma x^\gamma}{C_1}\Big) + p^2n^2\exp\Big(-\frac{a_n^2x^2}{C_2(1+C_3(t-s))}\Big)\\
&+ p^2n^2\exp\Big(-\frac{a_nx^2}{C_4}\exp\Big(\frac{a_n^{\gamma(1-\gamma)}x^{\gamma(1-\gamma)}}{C_5(\log(t-s)x)^\gamma}\Big)\Big).
\end{aligned}$$

*Furthermore, if $a_n$ satisfies $(\log n)^{1/\gamma} = o(a_n)$, for any constant $x$, it holds that as $n \to \infty$*

$$P\Big(\max_{|t-s|>a_n, t_{i-1}^0 \le s<t\le t_i^0} \max_{1\le j\le p^2} |\frac{1}{t-s}\sum_{l=s}^{t}\varepsilon_{lj}| \ge x\Big) \to 0.$$

**Proof.** By Theorem 1 of Merlevéde, Peligrad and Rio (2009), there exist positive constant $C_1, C_2, C_3, C_4$ and $C_5$ depending only on $c, \gamma$ and $\gamma_1$ such that for any positive constant $x$,

$$\begin{aligned}
&P\Big(\max_{\substack{|t-s|>a_n\\ t_{i-1}^0\le s<t\le t_i^0}} \max_{1\le j\le p^2} \frac{1}{t-s}|\sum_{l=s}^{t-1}\varepsilon_{lj}| \ge x\Big)\\
\le&\sum_{j=1}^{p^2}\sum_{\substack{|t-s|>a_n\\ t_{i-1}^0\le s<t\le t_i^0}} P\Big(|\sum_{l=s}^{t-1}\varepsilon_{lj}| \ge (t-s)x\Big)\\
\le&\sum_{j=1}^{p^2}\sum_{\substack{|t-s|>a_n\\ t_{i-1}^0\le s<t\le t_i^0}} P\Big(\sup_{s\le k\le t-1}|\sum_{l=s}^{k}\varepsilon_{lj}| \ge (t-s)x\Big)\\
\le&p^2n^2\Big((t-s)\exp\Big(-\frac{(t-s)^\gamma x^\gamma}{C_1}\Big) + \exp\Big(-\frac{(t-s)^2x^2}{C_2(1+C_3(t-s))}\Big)\\
&+\exp\Big(-\frac{(t-s)x^2}{C_4}\exp\Big(\frac{(t-s)^{\gamma(1-\gamma)}x^{\gamma(1-\gamma)}}{C_5(\log(t-s)x)^\gamma}\Big)\Big)\Big)\\
\le&p^2n^3\exp\Big(-\frac{a_n^\gamma x^\gamma}{C_1}\Big) + p^2n^2\exp\Big(-\frac{a_n^2x^2}{C_2(1+C_3(t-s))}\Big)\\
&+p^2n^2\exp\Big(-\frac{a_nx^2}{C_4}\exp\Big(\frac{a_n^{\gamma(1-\gamma)}x^{\gamma(1-\gamma)}}{C_5(\log(t-s)x)^\gamma}\Big)\Big).
\end{aligned}$$

The required result is then straightforward for the first part.

Given $(\log n)^{1/\gamma} = o(a_n)$, for any constant $x$, we have

$$p^2 n^3 \exp\Big( - \frac{a_n^\gamma x^\gamma}{C_1} \Big) \asymp \exp\Big( 2\log p + 3\log n - a_n^\gamma \Big) \to 0.$$

Similarly, we have

$$p^2 n^2 \exp\Big( - \frac{a_n^2 x^2}{C_2(1 + C_3(t-s))} \Big) \to 0,$$

and

$$p^2 n^2 \exp\Big( - \frac{a_n x^2}{C_4} \exp\Big( \frac{a_n^{\gamma(1-\gamma)} x^{\gamma(1-\gamma)}}{C_5 (\log(t-s)x)^\gamma} \Big) \Big) \to 0,$$

hence $P\Big( \max_{|t-s|>a_n, t_{i-1}^0 \le s < t \le t_i^0} \max_{1\le j\le p^2} |\frac{1}{t-s} \sum_{l=s}^t \varepsilon_{lj}| \ge x \Big) \to 0.$ □

**Lemma 7** *Under the conditions of Theorem 11, for $m < m_0$, there exists a constant $\nu$ such that*

$$S_n(\tilde{t}_1, \cdots, \tilde{t}_m) > \sum_{i=1}^n \|\boldsymbol{\varepsilon}_i\|^2 + \nu \min_{1\le i\le m_0} |t_i^0 - t_{i-1}^0|$$

*in probability, where $S_n(\tilde{t}_1, \cdots, \tilde{t}_m) = \arg\min_{t_1, \cdots, t_m} S_n(t_1, \cdots, t_m)$.*

**Proof.** The proof is the same as Lemma 6.4 of Chan, Yau and Zhang (2014). □

**Proof of Theorem 9.** Define $A_{ni} = \{|\widehat{t}_i - t_i^0| > n\gamma_n\}, i = 1, 2, \cdots, m_0$, it holds that

$$P\{\max_{1\le i\le m_0} |\widehat{t}_i - t_i^0| > n\gamma_n\} \le \sum_{i=1}^{m_0} P\{|\widehat{t}_i - t_i^0| > n\gamma_n\} = \sum_{i=1}^{m_0} P(A_{ni}).$$

Define the set $C_n$ by $C_n = \{\max_{1\le i\le m_0} |\widehat{t}_i - t_i^0| \le \min_i |t_i^0 - t_{i-1}^0|/2\}$, it is enough to prove that $\sum_{i=1}^{m_0} P(A_{ni}C_n) \to 0$ and $\sum_{i=1}^{m_0} P(A_{ni}C_n^c) \to 0$, where $C_n^c$ is the complement of $C_n$. The proof is similar to Proposition 5 of Harchaoui and Lévy-Leduc (2010) and hence we only give the proof of $\sum_{i=1}^{m_0} P(A_{ni}C_n) \to 0$. Note that $C_n$ implies that

$$t_{i-1}^0 < \widehat{t}_i < t_{i+1}^0 \quad \text{for} \quad 1 \le i \le m_0.$$

First consider the case when $\widehat{t}_i \le t_i^0$. Applying lemma 5 with $\widehat{t}_i$ and $t_i^0$, we have

$$\|\sum_{l=\widehat{t}_i}^{n}\sum_{i=1}^{l}\widehat{\boldsymbol{\theta}}_i - \sum_{l=\widehat{t}_i}^{n}\mathbf{x}_l\| \le \frac{1}{2}n\lambda_n \quad \text{and} \quad \|\sum_{l=t_i^0}^{n}\sum_{i=1}^{l}\widehat{\boldsymbol{\theta}}_i - \sum_{l=t_i^0}^{n}\mathbf{x}_l\| \le \frac{1}{2}n\lambda_n.$$

It follows from triangle inequality that

$$\|\sum_{l=\widehat{t}_i}^{t_i^0-1}\mathbf{x}_l - \sum_{l=\widehat{t}_i}^{t_i^0-1}\sum_{k=1}^{l}\widehat{\boldsymbol{\theta}}_k\| \le n\lambda_n.$$

Note that when $l \in [\widehat{t}_i, t_i^0 - 1]$, we have $\mathbf{x}_l = \boldsymbol{\mu}_i^0 + \boldsymbol{\varepsilon}_l$ and $\sum_{k=1}^{l}\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\mu}}_{i+1}$, it holds that

$$\|\sum_{l=\widehat{t}_i}^{t_i^0-1}\boldsymbol{\varepsilon}_l + \sum_{l=\widehat{t}_i}^{t_i^0-1}(\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0) + \sum_{l=\widehat{t}_i}^{t_i^0-1}(\boldsymbol{\mu}_{i+1}^0 - \widehat{\boldsymbol{\mu}}_{i+1})\| \le n\lambda_n.$$

It follows that

$$\begin{aligned}
P\Big(A_{ni}C_n \cap \{\widehat{t}_i \le t_i^0\}\Big) \le & P\Big(\{\frac{1}{3}(t_i^0 - \widehat{t}_i)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le n\lambda_n\} \cap \{|\widehat{t}_i - t_i^0| > n\gamma_n\}\Big) \\
& + P\Big(\{\frac{1}{3}(t_i^0 - \widehat{t}_i)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le \|\sum_{l=\widehat{t}_i}^{t_i^0-1}\boldsymbol{\varepsilon}_l\|\} \cap \{|\widehat{t}_i - t_i^0| > n\gamma_n\}\Big) \\
& + P\Big(\{\frac{1}{3}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le \|\boldsymbol{\mu}_{i+1}^0 - \widehat{\boldsymbol{\mu}}_{i+1}\|\|\} \cap \{|\widehat{t}_i - t_i^0| > n\gamma_n\}\Big) \\
\equiv & P(A_{ni1}) + P(A_{ni2}) + P(A_{ni3}).
\end{aligned}$$

Since $\min_{1\le i\le m_0+1}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i-1}^0\| \ge \nu$, in the set $\{|\widehat{t}_i - t_i^0| > n\gamma_n\}$ we have

$$\frac{1}{3}(t_i^0 - \widehat{t}_i)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| > Cn\gamma_n.$$

Since $\frac{\gamma_n}{\lambda_n} \to \infty$, we have $P(A_{ni1}) \to 0$.

In the set $\{|\widehat{t}_i - t_i^0| > n\gamma_n\}$ we have

$$\|\frac{1}{t_i^0 - \widehat{t}_i}\sum_{l=\widehat{t}_i}^{t_i^0-1}\boldsymbol{\varepsilon}_l\| \le \max_{\substack{|t-s|>n\gamma_n \\ t_{i-1}^0 \le s < t \le t_i^0}}\|\frac{1}{t-s}\sum_{l=s}^{t-1}\boldsymbol{\varepsilon}_l\| \le \max_{\substack{|t-s|>n\gamma_n \\ t_{i-1}^0 \le s < t \le t_i^0}}\max_{1\le j\le p^2}|p\frac{1}{t-s}\sum_{l=s}^{t-1}\varepsilon_{lj}|.$$

Note that $\frac{1}{3}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i-1}^0\| \ge \frac{1}{3}\nu$ and by lemma 6, we have

$$\begin{aligned}
& P\Big(\{\frac{1}{3}(t_i^0 - \widehat{t}_i)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le \|\sum_{l=\widehat{t}_i}^{t_i^0-1}\boldsymbol{\varepsilon}_l\|\} \cap \{|\widehat{t}_i - t_i^0| > n\gamma_n\}\Big) \\
\le & P\Big(\{\max_{\substack{|t-s|>n\gamma_n \\ t_{i-1}^0 \le s < t \le t_i^0}}\max_{1\le j\le p^2}|p\frac{1}{t-s}\sum_{l=s}^{t-1}\varepsilon_{lj}| \ge \frac{1}{3}\nu\}\Big) \to 0.
\end{aligned}$$

Hence $P(A_{ni2}) \to 0$.

Note that $C_n \cap \{\widehat{t}_i \le t_i^0\}$ implies $\widehat{t}_{i+1} > (t_i^0 + t_{i+1}^0)/2$. Hence if $l \in [t_i^0, (t_i^0 + t_{i+1}^0)/2]$, it holds that $\mathbf{x}_l = \boldsymbol{\mu}_{i+1}^0 + \boldsymbol{\varepsilon}_l$ and $\sum_{k=1}^{l} \widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\mu}}_{i+l}$. Applying lemma 5 with $t_i^0$ and $\frac{t_i^0+t_{i+1}^0}{2}$ and using triangle inequality we have

$$\| \sum_{l=t_i^0}^{(t_i^0+t_{i+1}^0)/2-1} \boldsymbol{\varepsilon}_l + \sum_{l=t_i^0}^{(t_i^0+t_{i+1}^0)/2-1} (\boldsymbol{\mu}_{i+1}^0 - \widehat{\boldsymbol{\mu}}_{i+l}) \| \le n\lambda_n.$$

Hence

$$\frac{t_{i+1}^0 - t_i^0}{2} \|\boldsymbol{\mu}_{i+1}^0 - \widehat{\boldsymbol{\mu}}_{i+1}\| \le n\lambda_n + \| \sum_{l=t_i^0}^{(t_i^0+t_{i+1}^0)/2-1} \boldsymbol{\varepsilon}_l \|,$$

which implies

$$\begin{aligned}
P(A_{ni3}) \le & P\Big(\frac{1}{6}(t_{i+1}^0 - t_i^0)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le n\lambda_n + \| \sum_{l=t_i^0}^{(t_i^0+t_{i+1}^0)/2-1} \boldsymbol{\varepsilon}_l \|\Big) \\
\le & P\Big(\frac{1}{12}(t_{i+1}^0 - t_i^0)\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le n\lambda_n\Big) \\
& + P\Big(\frac{1}{6}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_{i+1}^0\| \le \| \frac{1}{(t_{i+1}^0 - t_i^0)/2} \sum_{l=t_i^0}^{(t_i^0+t_{i+1}^0)/2-1} \boldsymbol{\varepsilon}_l \|\Big) \\
\equiv & P_1 + P_2,
\end{aligned}$$

where $P_1 \to 0$ by $\min_{1\le i\le m_0+1} |t_i^0 - t_{i-1}^0|/(n\gamma_n) \to \infty$ and $\gamma_n/\lambda_n \to \infty$ and $P_2 \to 0$ by lemma 6. Hence $P(A_{ni3}) \to 0$. Now we finish the proof of $P\Big(A_{ni}C_n \cap \{\widehat{t}_i \le t_i^0\}\Big) \to 0$. Similarly, we can show that $P\Big(A_{ni}C_n \cap \{\widehat{t}_i > t_i^0\}\Big) \to 0$. Thus $P(A_{ni}C_n) \to 0$.

When $m_0$ is fixed, the required result is apparent. When $m_0 \to \infty$, by lemma 6, the rate of convergence of $P(A_{ni})$ can be fast enough such that $m_0 P(A_{ni}) \to 0$ for all $i = 1, \cdots, m_0$.

□

**Proof of Theorem 10.** To prove $|\mathcal{A}_n| \ge m_0$, suppose on the contrary that $|\mathcal{A}_n| < m_0$, then there exist some $t_{i_0}^0$ and $\widehat{t}_{l_0} \in \mathcal{A}_n$ such that $t_{i_0+}^0 - t_{i_0}^0 < \widehat{t}_{l_0+1} - \widehat{t}_{l_0}$, thus we have

$t^0_{i_0+1} - t^0_{i_0} \vee \widehat{t}_{l_0} \geq \frac{n\gamma_n}{3}$ and $t^0_{i_0+2} \wedge \widehat{t}_{l_0+1} - t^0_{i_0+1} \geq \frac{n\gamma_n}{3}$. Applying lemma 5 to $t^0_{i_0} \vee \widehat{t}_{l_0}$ and $t^0_{i_0+1}$ we have

$$(t^0_{i_0+1} - t^0_{i_0} \vee \widehat{t}_{l_0})\|\boldsymbol{\mu}^0_{i_0+1} - \widehat{\boldsymbol{\mu}}_{l_0}\| \leq n\lambda_n + \|\sum_{l=t^0_{i_0} \vee \widehat{t}_{l_0}}^{t^0_{i_0+1}-1} \boldsymbol{\varepsilon}_l\|,$$

and applying lemma 5 to $t^0_{i_0+1}$ and $t^0_{i_0+2} \wedge \widehat{t}_{l_0+1}$ we have

$$(t^0_{i_0+2} \wedge \widehat{t}_{l_0+1} - t^0_{i_0+1})\|\boldsymbol{\mu}^0_{i_0+2} - \widehat{\boldsymbol{\mu}}_{l_0}\| \leq n\lambda_n + \|\sum_{l=t^0_{i_0+1}}^{t^0_{i_0+2} \wedge \widehat{t}_{l_0+1}-1} \boldsymbol{\varepsilon}_l\|.$$

since $t^0_{i_0+1} - t^0_{i_0} \vee \widehat{t}_{l_0} \geq \frac{n\gamma_n}{3}$ and $t^0_{i_0+2} \wedge \widehat{t}_{l_0+1} - t^0_{i_0+1} \geq \frac{n\gamma_n}{3}$, we have

$$\|\boldsymbol{\mu}^0_{i_0+1} - \widehat{\boldsymbol{\mu}}_{l_0}\| \leq \frac{\lambda_n}{\gamma_n} + \frac{1}{t^0_{i_0+1} - t^0_{i_0} \vee \widehat{t}_{l_0}}\|\sum_{l=t^0_{i_0} \vee \widehat{t}_{l_0}}^{t^0_{i_0+1}-1} \boldsymbol{\varepsilon}_l\|,$$

lemma 6 leads to $\frac{1}{t^0_{i_0+1} - t^0_{i_0} \vee \widehat{t}_{l_0}}\|\sum_{l=t^0_{i_0} \vee \widehat{t}_{l_0}}^{t^0_{i_0+1}-1} \boldsymbol{\varepsilon}_l\| \to 0$, together with $\lambda_n/\gamma_n \to 0$ we have

$$\|\boldsymbol{\mu}^0_{i_0+1} - \widehat{\boldsymbol{\mu}}_{l_0}\| \xrightarrow{p} 0,$$

similarly,

$$\|\boldsymbol{\mu}^0_{i_0+2} - \widehat{\boldsymbol{\mu}}_{l_0}\| \xrightarrow{p} 0,$$

which means $\boldsymbol{\mu}^0_{i_0+1}$ and $\boldsymbol{\mu}^0_{i_0+2}$ are the same. This contradicts with $\boldsymbol{\mu}^0_{i_0+1} \neq \boldsymbol{\mu}^0_{i_0+2}$. Hence $P(|\mathcal{A}_n| \geq m_0) \to 1$.

The proof of $P(d_H(\mathcal{A}_n, \mathcal{A}) \leq n\gamma_n) \to 1$ is the same as the second part of Theorem 2.3 of Chan, Yau and Zhang (2014), hence omitted. □

**Proof of Theorem 11.** To prove $P(\widehat{\widehat{m}} = m_0) \to 1$, it suffices to prove $P(\widehat{\widehat{m}} < m_0) \to 0$ and $P(\widehat{\widehat{m}} > m_0) \to 0$. First let us prove $P(\widehat{\widehat{m}} < m_0) \to 0$. It follows from Theorem 10 that there exist points $\widehat{t}_{ni} \in \mathcal{A}_n, i = 1, 2, \cdots, m_0$ such that $\max_{1 \leq i \leq m_0} |\widehat{t}_{ni} - t^0_i| \leq n\gamma_n$. Now it suffices to show that if $\widehat{\widehat{m}} < m_0$, we have $IC(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{t}}}) \geq S_n(\widehat{t}_{n1}, \cdots, \widehat{t}_{nm_0}) + m_0\omega_n$ in probability.

Denote $R_n(m_0) = \{(t_1, t_2, \cdots, t_{m_0}) : |t_i - t_i^0| \leq n\gamma_n, i = 1, 2\cdots, m_0\}$. For any $\mathbf{t} \in R_n(m_0)$, we have

$$\begin{aligned} S_n(t_1, t_2, \cdots, t_{m_0}) = & \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_1\|^2 + \sum_{j=2}^{m_0} \sum_{i=t_{j-1}^0 + n\gamma_n}^{t_j^0 - n\gamma_n - 1} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 + \sum_{i=t_m^0 + n\gamma_n}^{n} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_{m_0+1}\|^2 \\ & + \sum_{j=1}^{m_0} \sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 + \sum_{j=1}^{m_0} \sum_{i=t_j^0}^{t_j^0 + n\gamma_n - 1} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_{j+1}\|^2 \\ = & L_1 + L_2 + L_3 + L_4 + L_5, \end{aligned}$$

where $\widehat{\widehat{\boldsymbol{\mu}}}_j$ are the least square estimators of $\boldsymbol{\mu}_j, 1 \leq j \leq m_0 + 1$ on $[t_{j-1}, t_j - 1]$. It can be shown that in probability

$$L_1 + L_2 + L_3 \leq \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\boldsymbol{\varepsilon}_i\|^2 + \sum_{j=2}^{m_0} \sum_{i=t_{j-1}^0 + n\gamma_n}^{t_j^0 - n\gamma_n - 1} \|\boldsymbol{\varepsilon}_i\|^2 + \sum_{i=t_m^0 + n\gamma_n}^{n} \|\boldsymbol{\varepsilon}_i\|^2 + O(m_0 n\gamma_n),$$

and the proof is as follows: take $L_1$ for an example, denote $\tilde{\boldsymbol{\mu}}_1$ as the LSE obtained by using the data on $[1, t_1^0 - n\gamma_n - 1]$, given $E\|\mathbf{x}_i\|$ exists, we have in probability

$$\begin{aligned} L_1 = & \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_1\|^2 \\ = & \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\|^2 \\ \leq & \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_1\|^2 + \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\|^2 + 2 \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_1\| \|\tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\| \\ \leq & \sum_{i=1}^{t_1^0 - n\gamma_n - 1} \|\boldsymbol{\varepsilon}_i\|^2 + (t_1^0 - n\gamma_n - 1)\|\tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\|^2 + 2O((t_1^0 - n\gamma_n - 1)\|\tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\|), \end{aligned}$$

where the $O(\cdot)$ is obtained by Markov inequality.

Since $0 \leq t_1 - t_1^0 + n\gamma_n \leq 2n\gamma_n$, given $E\|\mathbf{x}_i\|$ exists, it holds that in probability

$$\begin{aligned}
&(t_1^0 - n\gamma_n - 1)\|\tilde{\boldsymbol{\mu}}_1 - \widehat{\widehat{\boldsymbol{\mu}}}_1\| \\
\leq &\|\sum_{i=1}^{t_1^0 - n\gamma_n - 1} \mathbf{x}_i - \frac{t_1^0 - n\gamma_n - 1}{t_1 - 1}\sum_{i=1}^{t_1-1}\mathbf{x}_i\| \\
= &\|\sum_{i=1}^{t_1^0 - n\gamma_n - 1} \mathbf{x}_i - (1 - \frac{t_1 - t_1^0 + n\gamma_n}{t_1 - 1})\sum_{i=1}^{t_1-1}\mathbf{x}_i\| \\
= &\|\sum_{i=t_1^0 - n\gamma_n}^{t_1-1} \mathbf{x}_i\| + \|\frac{t_1 - t_1^0 + n\gamma_n}{t_1 - 1}\sum_{i=1}^{t_1-1}\mathbf{x}_i\| \\
= &O(t_1 - t_1^0 + n\gamma_n) = O(n\gamma_n),
\end{aligned}$$

where the $O(\cdot)$ is obtained by Markov inequality. Hence $L_1 \leq \sum_{i=1}^{t_1^0 - n\gamma_n - 1}\|\boldsymbol{\varepsilon}_i\|^2 + O(n\gamma_n)$. We then have similar results to $L_2$ and $L_3$, hence the above result has been proved.

Now let's turn to $L_4 + L_5$. It can be shown that there exists $A_0 > 0$ such that in probability

$$L_4 + L_5 \leq \sum_{j=1}^{m_0}\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\varepsilon}_i\|^2 + \sum_{j=1}^{m_0}\sum_{i=t_j^0}^{t_j^0 + n\gamma_n - 1}\|\boldsymbol{\varepsilon}_i\|^2 + A_0 m_0 n\gamma_n,$$

and the proof of this equation is as follows:

take $L_4$ for an example, $L_4 = \sum_{j=1}^{m_0}\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2$ and

$$\begin{aligned}
&\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\mathbf{x}_i - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 = \sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\mathbf{x}_i - \boldsymbol{\mu}_j + \boldsymbol{\mu}_j - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 \\
\leq &\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\varepsilon}_i\|^2 + \sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\mu}_j - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 + 2\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\varepsilon}_i\|\|\boldsymbol{\mu}_j - \widehat{\widehat{\boldsymbol{\mu}}}_j\| \\
\equiv &\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\varepsilon}_i\|^2 + A^{(j)} n\gamma_n,
\end{aligned}$$

where

$$A^{(j)} = \|\boldsymbol{\mu}_j - \widehat{\widehat{\boldsymbol{\mu}}}_j\|^2 + \frac{2}{n\gamma_n}\sum_{i=t_j^0 - n\gamma_n}^{t_j^0 - 1}\|\boldsymbol{\varepsilon}_i\|\|\boldsymbol{\mu}_j - \widehat{\widehat{\boldsymbol{\mu}}}_j\|.$$

Then

$$L_4 \leq \sum_{j=1}^{m_0} \sum_{i=t_j^0-n\gamma_n}^{t_j^0-1} \|\boldsymbol{\varepsilon}_i\|^2 + n\gamma_n \sum_{j=1}^{m_0} A^{(j)}.$$

Similarly, we have

$$L_5 \leq \sum_{j=1}^{m_0} \sum_{i=t_j^0}^{t_j^0+n\gamma_n-1} \|\boldsymbol{\varepsilon}_i\|^2 + n\gamma_n \sum_{j=1}^{m_0} B^{(j)},$$

where

$$B^{(j)} = \|\boldsymbol{\mu}_{j+1} - \widehat{\widehat{\boldsymbol{\mu}}}_{j+1}\|^2 + \frac{2}{n\gamma_n} \sum_{i=t_j^0}^{t_j^0+n\gamma_n-1} \|\boldsymbol{\varepsilon}_i\| \|\boldsymbol{\mu}_{j+1} - \widehat{\widehat{\boldsymbol{\mu}}}_{j+1}\|.$$

Then

$$A_0 = \sum_{j=1}^{m_0} A^{(j)}/m_0 + \sum_{j=1}^{m_0} B^{(j)}/m_0.$$

Note that $A_0 = O_p(1)$ given $E\|\boldsymbol{\varepsilon}_i\|$ exists.

Hence if $\mathbf{t} \in R_n(m_0)$, it holds that in probability

$$S_n(t_1, t_2, \cdots, t_{m_0}) \leq \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|^2 + (A_0 + O(1)) m_0 n\gamma_n.$$

Since $(\widehat{t}_{n1}, \cdots, \widehat{t}_{nm_0}) \in R_n(m_0)$, we have in probability

$$S_n(\widehat{t}_{n1}, \cdots, \widehat{t}_{nm_0}) \leq \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|^2 + (A_0 + O(1)) m_0 n\gamma_n.$$

At the same time, by lemma 7 we have in probability

$$S_n(\widehat{\widehat{t}}_1, \cdots, \widehat{\widehat{t}}_{\widehat{\widehat{m}}}) \geq \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|^2 + \nu \min_{1 \leq i \leq m_0} |t_i^0 - t_{i-1}^0|.$$

Hence it holds that in probability

$$\begin{aligned} IC(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{t}}}) &= S_n(\widehat{\widehat{t}}_1, \cdots, \widehat{\widehat{t}}_{\widehat{\widehat{m}}}) + \widehat{\widehat{m}}\omega_n \\ &\geq \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|^2 + \nu \min_{1 \leq i \leq m_0} |t_i^0 - t_{i-1}^0| + \widehat{\widehat{m}}\omega_n \\ &\geq S_n(\widehat{t}_{n1}, \cdots, \widehat{t}_{nm_0}) + m_0\omega_n + \nu \min_{1 \leq i \leq m_0} |t_i^0 - t_{i-1}^0| - (A_0 + O(1)) m_0 n\gamma_n - (m_0 - \widehat{\widehat{m}})\omega_n. \end{aligned}$$

Since $\omega_n/\min_{1\le i\le m_0}|t_i^0-t_{i-1}^0|\to 0$ and $n\gamma_n/\min_{1\le i\le m_0}|t_i^0-t_{i-1}^0|\to 0$, we have

$$IC(\widehat{\widehat{m}},\widehat{\widehat{\mathbf{t}}})\ge S_n(\widehat{t}_{n1},\cdots,\widehat{t}_{nm_0})+m_0\omega_n,$$

in probability and this implies

$$P(\widehat{\widehat{m}}<m_0)\to 0.$$

Now let us prove $P(\widehat{\widehat{m}}>m_0)\to 0$, which suffices to show that if $\widehat{\widehat{m}}>m_0$, we have $IC(\widehat{\widehat{m}},\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}})>IC(m_0,\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{m_0})$. Note that

$$S_n(\widehat{t}_{n1},\cdots,\widehat{t}_{nm_0})\ge S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{m_0})\ge S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}})\ge S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}},t_1^0,\cdots,t_{m_0}^0).$$

It can be shown that

$$S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}},t_1^0,\cdots,t_{m_0}^0)\ge\sum_{i=1}^n\|\boldsymbol{\varepsilon}_i\|^2-(\widehat{\widehat{m}}+m_0)n\gamma_n,$$

hence it holds that

$$\begin{aligned}&S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{m_0})-S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}})\\ \le&S_n(\widehat{t}_{n1},\cdots,\widehat{t}_{nm_0})-S_n(\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}},t_1^0,\cdots,t_{m_0}^0)\\ \le&(\widehat{\widehat{m}}+m_0+m_0A_0)n\gamma_n.\end{aligned}$$

Since $m_0n\gamma_n/\omega_n\to 0$, it then follows that

$$IC(\widehat{\widehat{m}},\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{\widehat{\widehat{m}}})-IC(m_0,\widehat{\widehat{t}}_1,\cdots,\widehat{\widehat{t}}_{m_0})\ge(\widehat{\widehat{m}}-m_0)\omega_n-(\widehat{\widehat{m}}+m_0+m_0A_0)n\gamma_n>0,$$

which implies

$$P(\widehat{\widehat{m}}>m_0)\to 0.$$

The proof of $P(\max_{1\le i\le m_0}|\widehat{\widehat{t}}_i-t_i^0|\le Bn\gamma_n)\to 1$ can be obtained following Theorem 2.4 and lemma 6.4 of Chan, Yau and Zhang (2014). □

**Proof of Theorem 12.** The proof is the same as Theorem 2.5 of Chan, Yau and Zhang (2014). □

# Chapter 4: Two Simple Results

This chapter presents two small theoretical results which extend two theorems for high dimensional, independent processes to high dimensional and dependent processes. They come from the effort of proving the results in the previous three chapters, but failing to obtain the target.

## 4.1 An extension of Bickel, P.J. and Levina, E (2008)'s result

In this section, we extend the result of Bickel and Levina (2008) to (auto)covariance matrices of high-dimensional $\alpha$-mixing dependent data. Let $\mathbf{y}_t = (y_{1,t}, \ldots, y_{p,t})^T, t = 1, \cdots, n$ be a $p \times 1$ strictly stationary $\alpha$-mixing process. Denote the (auto)covariance matrices of $\mathbf{y}_t$ at lag $k$ by $\text{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t) = \mathbf{\Sigma}_{\mathbf{y}}(k) = [\sigma_{i,j}^{(k)}]_{i,j=1,2,\cdots,p}, k \geq 0$. For each $k$, we assume it belongs to the following matrix class:

$$\left\{ \mathbf{\Sigma} : \sigma_{i,i} \leq M, \max_i \sum_{i=1}^{p} \mid \sigma_{i,j} \mid^q \leq s_1(p), \max_j \sum_{i=1}^{p} \mid \sigma_{i,j} \mid^q \leq s_2(p) \right\}, \tag{4.1.43}$$

where $0 \leq q < 1$ and $M$ is a positive constant.

Considering such a class above, we employ a thresholding estimator defined as

$$T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) = [\widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u)]_{i,j=1,2,\cdots,p},$$

where $\widetilde{\sigma}_{i,j}^{(k)}$ is the $(i,j)$-th element of the sample (auto)covariance matrix estimator $\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(\mathbf{y}_{t+k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^T$. We assume

(A1) There exist positive constants $b_1, b_2, r_1$ such that for any $s > 0$ and $i \leq p$

$$P(|\ y_{i,t} - E(y_{i,t})\ | > s) \leq b_1 \exp(-b_2 s^{r_1}).$$

(A2) $\mathbf{y}_t$ is strictly stationary and there exist positive constants $b_3, r_2$ such that the $\alpha$-mixing coefficient satisfies

$$\alpha(m) \leq \exp(-b_3 m^{r_2}) \quad \text{for any } m \geq 1.$$

We have the following result

**Theorem 13** *If $\mathbf{\Sigma}_{\mathbf{y}}(k), k \geq 0$ belongs to the sparse matrices class defined in (4.1.43), under assumptions A1, A2, assume $0 < \gamma_1 < 1$ where $1/\gamma_1 = 2/r_1 + 1/r_2$ and $r_1$, $r_2$ are defined in A1, A2, if $(\log p)^{2/\gamma_1 - 1} = o(n)$ and the thresholding parameter $u$ is*

$$u = M'\sqrt{\frac{\log p}{n}},$$

*for sufficient large $M'$. Then*

$$\|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2 = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \sqrt{s_1^{(k)}(p) s_2^{(k)}(p)}\right).$$

**Proof.** Without loss of generality, we assume $E(y_{i,t}) = E(y_{i,t}) = 0$ for $i, j = \{1, 2, \cdots, p\}$. For $k = 0, 1, 2, \cdots$ we have the following decomposition

$$\begin{aligned}\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} =& \frac{1}{n}\sum_{t=1}^{n-k}(y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k})) - \bar{y}_{j,\cdot}\frac{1}{n}\sum_{t=1}^{n-k} y_{i,t+k} - \bar{y}_{i,\cdot}\frac{1}{n}\sum_{t=1}^{n-k} y_{j,t+k} \\ &+ \frac{n-k}{n}\bar{y}_{i,\cdot}\bar{y}_{j,\cdot} - \frac{k}{n}E(y_{i,t+k}y_{j,k}),\end{aligned} \tag{4.1.44}$$

where $\bar{y}_{i,\cdot} = \frac{1}{n}\sum_{t=1}^{n} y_{i,t}$ and $\bar{y}_{j,\cdot} = \frac{1}{n}\sum_{t=1}^{n} y_{j,t}$. Note that when $k = 0$, the last four terms of (4.1.44) becomes $-\bar{y}_{i,\cdot}\bar{y}_{j,\cdot}$. This difference does not affect the asymptotic results discussed following.

Now let's consider the first term in (4.1.44). From Lemma 7 of Chang, Guo and Yao (2014), for any $s > 0$, under assumption A1 we have

$$P(|y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k})| > s) \leq 2b_1 \exp(-b_2 s^{r_1/2}).$$

By Theorem 1 of Merlevéde et al (2011), there exists constants $C_1, C_2, C_3, C_4, C_5 > 0$ only depending on $b_1, b_2, r_1, r_2$ (that is not depending on $i, j, k$) such that the upper bound for the first term in (4.1.44)

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{t=1}^{n-k}(y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k}))\right| \geq s\right) \leq & n\exp\left(-\frac{(ns)^{\gamma_1}}{C_1}\right) + \exp\left(-\frac{(ns)^2}{C_2(1+nC_3)}\right) \\ & + \exp\left(-\frac{(ns)^2}{C_4 n}\exp\left(\frac{(ns)^{\gamma_1(1-\gamma_1)}}{C_5(\log ns)^{\gamma_1}}\right)\right) \quad \text{for all } i, j. \end{aligned}$$

By Bonferroni's method we have

$$P\left(\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^{n-k}(y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k}))\right| \geq s\right) \leq p^2 \max_{i,j} P\left(\left|\frac{1}{n}\sum_{t=1}^{n-k}(y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k}))\right| \geq s\right).$$

Set $s = u = M\sqrt{\frac{\log p}{n}}$, when $(\log p)^{2/\gamma_1 - 1} = o(n)$ we have

$$p^2 n\exp\left(-\frac{(ns)^{\gamma_1}}{C_1}\right) + p^2\exp\left(-\frac{(ns)^2}{C_2(1+nC_3)}\right) + p^2\exp\left(-\frac{(ns)^2}{C_4 n}\exp\left(\frac{(ns)^{\gamma_1(1-\gamma_1)}}{C_5(\log ns)^{\gamma_1}}\right)\right) = o(1).$$

Hence

$$\max_{i,j}\left|\frac{1}{n}\sum_{t=1}^{n-k}(y_{i,t+k}y_{j,k} - E(y_{i,t+k}y_{j,k}))\right| = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

As for the second term in (4.1.44), we have

$$P\left(\left|\bar{y}_{j,\cdot}\frac{1}{n}\sum_{t=1}^{n-k} y_{i,t+k}\right| \geq s\right) \leq P(|\bar{y}_{j,\cdot}| \geq s^{1/2}) + P\left(\left|\frac{1}{n}\sum_{t=1}^{n-k} y_{i,t+k}\right| \geq s^{1/2}\right).$$

Hence there exists constants $C_1^*, C_2^*, C_3^*, C_4^*, C_5^* > 0$ only depending on $b_1, b_2, r_1, r_2$ and define $\frac{1}{\gamma_2} = \frac{1}{r_1} + \frac{1}{r_2}$ we have

$$P(\max_{i,j} |\bar{y}_{j,\cdot}| \geq s^{1/2}) \leq p^2 n \exp\left(-\frac{n^{\gamma_2} s^{\gamma_2/2}}{C_1^*}\right) + p^2 \exp\left(-\frac{n^2 s}{C_2^*(1+nC_3^*)}\right) + p^2 \exp\left(-\frac{n^2 s}{C_4^* n} \exp\left(\frac{n^{\gamma_2(1-\gamma_2)} s^{\gamma_2(1-\gamma_2)/2}}{C_5^* (\log n s^{1/2})^{\gamma_2}}\right)\right).$$

Similar to the above arguments, let $s = u = M\sqrt{\frac{\log p}{n}}$, when $(\log p)^{4/3\gamma_2 - 1/3} = o(n)$ we have $\max_{i,j} \left|\bar{y}_{j,\cdot} \frac{1}{n} \sum_{t=1}^{n-k} y_{i,t+k}\right| = O_p(\sqrt{\frac{\log p}{n}})$. Similarly when $(\log p)^{4/3\gamma_2 - 1/3} = o(n)$ we have

$$\max_{i,j} \left|\bar{y}_{i,\cdot} \frac{1}{n} \sum_{t=1}^{n-k} y_{j,t+k}\right| = O_p(\sqrt{\frac{\log p}{n}}) \quad \text{and} \quad \max_{i,j} \left|\frac{n-k}{n} \bar{y}_{i,\cdot} \bar{y}_{j,\cdot}\right| = O_p(\sqrt{\frac{\log p}{n}}).$$

And $\max_{i,j} \left|\frac{k}{n} E(y_{i,t+k} y_{j,k})\right| = O(\frac{1}{n})$ which is irrelevant with $p$. Note that when $\gamma_1 < \gamma_2 < 1$, we have $2/\gamma_1 - 1 \geq 4/3\gamma_2 - 1/3$, hence when $(\log p)^{2/\gamma_1 - 1} = o(n)$ we have

$$\max_{i,j} \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| = O_p(\sqrt{\frac{\log p}{n}}) \quad k = 0, 1, 2, \cdots.$$

The following is simply a recap of Bickel and Levina (2008).

$$\|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2 \leq \|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - T_u(\mathbf{\Sigma}_{\mathbf{y}}(k))\|_2 + \|T_u(\mathbf{\Sigma}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2.$$

Under the sparse matrices class defined in (4.1.43), the second term has the following upper bound

$$\begin{aligned}\|T_u(\mathbf{\Sigma}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2 &\leq (\max_i \sum_{j=1}^{p} 1(|\ \sigma_{i,j}^{(k)}\ | < u))^{1/2} (\max_j \sum_{i=1}^{p} 1(|\ \sigma_{i,j}^{(k)}\ | < u))^{1/2} \\ &\leq (u^{1-q} s_1^{(k)}(p))^{1/2} (u^{1-q} s_2^{(k)}(p))^{1/2} = u^{1-q} \sqrt{s_1^{(k)}(p) s_2^{(k)}(p)}.\end{aligned}$$

Hence

$$\|T_u(\mathbf{\Sigma}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2 = O\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \sqrt{s_1^{(k)}(p) s_2^{(k)}(p)}\right). \tag{4.1.45}$$

We also have

$$\begin{aligned}\|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - T_u(\mathbf{\Sigma}_{\mathbf{y}}(k))\|_2 \leq (\max_i \sum_{j=1}^{p} \left|\widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u) - \sigma_{i,j}^{(k)} 1(|\ \sigma_{i,j}^{(k)}\ | \geq u)\right|)^{1/2} \\ \times (\max_j \sum_{i=1}^{p} \left|\widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u) - \sigma_{i,j}^{(k)} 1(|\ \sigma_{i,j}^{(k)}\ | \geq u)\right|)^{1/2}.\end{aligned}$$

And

$$\begin{aligned}
&\max_i \sum_{j=1}^{p} \left| \widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u) - \sigma_{i,j}^{(k)} 1(|\ \sigma_{i,j}^{(k)}\ | \geq u) \right| \\
\leq &\max_i \sum_{j=1}^{p} \left| \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} \right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, |\ \sigma_{i,j}^{(k)}\ | \geq u) \\
&+ \max_i \sum_{j=1}^{p} \left| \widetilde{\sigma}_{i,j}^{(k)} \right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, |\ \sigma_{i,j}^{(k)}\ | < u) \\
&+ \max_i \sum_{j=1}^{p} \left| \sigma_{i,j}^{(k)} \right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | < u, |\ \sigma_{i,j}^{(k)}\ | \geq u) \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

It is easy to check

$$\begin{aligned}
I_1 &\leq \max_{i,j} \left| \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} \right| u^{-q} s_1^{(k)}(p), \\
I_3 &\leq \max_{i,j} \left| \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} \right| u^{-q} s_1^{(k)}(p) + u^{1-q} s_1^{(k)}(p).
\end{aligned}$$

Hence when $u = M\sqrt{\frac{\log p}{n}}$ and $(\log p)^{2/\gamma_1 - 1} = o(n)$, we have

$$I_1 = O_p\left( \left( \frac{\log p}{n} \right)^{\frac{1-q}{2}} s_1^{(k)}(p) \right), \quad \text{and} \quad I_3 = O_p\left( \left( \frac{\log p}{n} \right)^{\frac{1-q}{2}} s_1^{(k)}(p) \right).$$

For term $I_2$, take a constant $\tau \in (0,1)$, we have

$$\begin{aligned}
I_2 &\leq \max_i \sum_{j=1}^{p} \left| \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} \right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, |\ \sigma_{i,j}^{(k)}\ | < u) + \max_i \sum_{j=1}^{p} \left| \sigma_{i,j}^{(k)} \right| 1(|\ \sigma_{i,j}^{(k)}\ | < u) \\
&\leq \max_i \sum_{j=1}^{p} \left| \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} \right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, |\ \sigma_{i,j}^{(k)}\ | < u) + u^{1-q} s_1^{(k)}(p),
\end{aligned}$$

where

$$\begin{aligned}
&\max_i \sum_{j=1}^{p} \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, |\ \sigma_{i,j}^{(k)}\ | < u) \\
\leq &\max_i \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| \max_i \sum_{j=1}^{p} 1(|\ \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\ | \geq (1-\tau)u) \\
&+ \max_i \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| \max_i \sum_{j=1}^{p} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u, \tau u \leq |\ \sigma_{i,j}^{(k)}\ | < u) \\
&\leq \max_i \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| \max_i \sum_{j=1}^{p} 1(|\ \widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\ | \geq (1-\tau)u) \\
&+ \max_i \left|\widetilde{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}\right| s_1^{(k)}(p)(\tau u)^{-q}.
\end{aligned}$$

When $u = M\sqrt{\frac{\log p}{n}}$ and $(\log p)^{2/\gamma_1 - 1} = o(n)$, it is not hard to check

$$I_2 = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} s_1^{(k)}(p)\right).$$

Hence we have

$$\max_i \sum_{j=1}^{p} \left|\widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u) - \sigma_{i,j}^{(k)} 1(|\ \sigma_{i,j}^{(k)}\ | \geq u)\right| = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} s_1^{(k)}(p)\right).$$

Similarly, we have

$$\max_j \sum_{i=1}^{p} \left|\widetilde{\sigma}_{i,j}^{(k)} 1(|\ \widetilde{\sigma}_{i,j}^{(k)}\ | \geq u) - \sigma_{i,j}^{(k)} 1(|\ \sigma_{i,j}^{(k)}\ | \geq u)\right| = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} s_2^{(k)}(p)\right).$$

Hence

$$\|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - T_u(\mathbf{\Sigma}_{\mathbf{y}}(k))\|_2 = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \sqrt{s_1^{(k)}(p) s_2^{(k)}(p)}\right).$$

Together with (4.1.45), we get

$$\|T_u(\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)) - \mathbf{\Sigma}_{\mathbf{y}}(k)\|_2 = O_p\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}} \sqrt{s_1^{(k)}(p) s_2^{(k)}(p)}\right).$$

given $(\log p)^{2/\gamma_1 - 1} = o(n)$. □

## 4.2 A result of $U$-statistics of high dimensional $\beta$ mixing processes

Hoeffding (1948) investigated the theory of $U$-statistics for fixed dimensional independent data. Serfling (1980) presented a good summary of $U$-statistics. Zhong and Chen (2011) extended hoeffding (1948) to high dimensional independent data. The theory of $U$-statistics for fixed dimensional dependent process was discussed by Yoshihara (1976), where they considered fixed dimensional $\beta$-mixing process. Dehling and Wendler (2010) extended Yoshihara (1976) to strong mixing data where the kernel function of the $U$-statistics need to satisfy some continuity conditions. We devote this section to discuss $U$-statistics for high dimensional $\beta$-mixing process. A strictly stationary process $\{\mathbf{y}_t\}$ is $\beta$-mixing if

$$\beta(k) \equiv E\Big\{ \sup_{B\in\mathcal{F}_k^\infty} \big|P(B) - P(B|\mathbf{y}_0, \mathbf{y}_{-1}, \mathbf{y}_{-2}, \cdots)\big|\Big\} \to 0, \quad \text{as } k \to \infty,$$

where $\mathcal{F}_i^j$ denotes the $\sigma$-algebra generated by $\{\mathbf{y}_t, i \le t \le j\}$.

Suppose $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_n$ are $n$ observations of a $p \times 1$ dimensional, strictly stationary and $\beta$-mixing process $\mathbf{W}_t$ with $\beta$-mixing coefficient $\beta(n)$. We denote the distribution function of $\mathbf{W}_t$ by $F(\mathbf{W}_t)$. Consider a functional of order $s$ for a fixed $s \le n$

$$\theta(F) = \int_{R^p} \ldots \int_{R^p} h(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s) dF(\mathbf{w}_1) dF(\mathbf{w}_2) \ldots dF(\mathbf{w}_s),$$

defined over $\mathcal{F} = \{F : |\theta(F)| < \infty\}$, where the kernel function $h(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s)$ is symmetric, i.e. its value is invariant to the permutations of its $s$ arguments. We estimate $\theta(F)$ by the following $U$-statistics,

$$U_{n,p} = \begin{pmatrix} n \\ s \end{pmatrix}^{-1} \sum_{C_{n,s}} h(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}, \ldots, \mathbf{w}_{i_s}), \tag{4.2.46}$$

where $C_{n,s}$ represents all distinct combinations of $\{i_1, i_2, \ldots i_s\}$ from $\{1, 2, \ldots, n\}$.

For every $1 \leq c \leq s$, define the projection via

$$h_c(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c) = \int_{R^p} \ldots \int_{R^p} h(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s) dF(\mathbf{w}_{c+1}) dF(\mathbf{w}_{c+2}) \ldots dF(\mathbf{w}_s),$$

and denote $\tilde{h}_c = h_c - \theta(F)$. Following the notation of Zhong and Chen (2011), let

$$g_c(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c) = \tilde{h}_c(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c) - \sum_{j=1}^{c-1} \sum_{1 \leq i_1 < \ldots < i_j \leq c} g_j(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}, \ldots, \mathbf{w}_{i_j}),$$

where $g_1(\mathbf{w}_1) = \tilde{h}_1(\mathbf{w}_1)$. Denote

$$M_{nc} = \sum_{1 \leq i_1 < \ldots < i_c \leq n} g_c(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_c).$$

By Hoeffding's decomposition we have

$$U_{n,p} - \theta(F) = \sum_{c=1}^{s} \begin{pmatrix} s \\ c \end{pmatrix} \begin{pmatrix} n \\ c \end{pmatrix}^{-1} M_{nc} = \sum_{c=1}^{s} \begin{pmatrix} s \\ c \end{pmatrix} U_{n,p}^{(c)}. \tag{4.2.47}$$

For fixed dimension cases, under some regularity conditions, $E(U_{n,p}^{(c)})^2 = O(n^{-2})$ for $2 \leq c \leq s$, see lemma 2 of Liu, Chen and Yao (2010). This means the dominant term of the $U$-statistics is $U_{n,p}^{(1)}$. However, when $p \to \infty$, $E(U_{n,p}^{(c)})^2, 1 \leq c \leq s$ are also affected by $p$. Hence the dominated term may not be the first term anymore. We need to compare each $Var(U_{n,p}^{(c)})$ for $c = 1, 2, \ldots, s$. In the following, we will compute the largest order that $U_{n,p}^{(c)}$ for $c \geq 2$ can obtain under some regularity conditions. Then we can, at least, compare the relative order of $U_{n,p}^{(1)}$ to the rest terms $U_{n,p}^{(c)}, c = 2, 3, \ldots, s$.

Assume that for some $r > 2$,

(**C**1) $\mu_r = \int_{R^p} \ldots \int_{R^p} |h(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s)|^r \, dF(\mathbf{w}_1) dF(\mathbf{w}_2) \ldots dF(\mathbf{w}_s) = O(p^{\eta_1(r)})$ for $\eta_1(r) \geq 0$.

and

(**C**2) $\nu_r = E\left|h(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}, \ldots, \mathbf{w}_{i_s})\right|^r = O(p^{\eta_2(r)})$ for $\eta_2(r) \geq 0$ and all integers $i_1, i_2, \ldots, i_s$.

Now we have the following proposition,

**Theorem 14** *If there is a positive number $\delta$ such that for $r = 2 + \delta$ condition* **C**1 *and* **C**2 *hold, and $\sum_{n\geq 1} n\beta(n)^{\delta/(2+\delta)} < \infty$, then we have*

$$E(U_{n,p}^{(c)})^2 = O\left(\max\left\{\left(\frac{p^{\frac{\eta_1(2+\delta)}{2+\delta}}}{n}\right)^2, \left(\frac{p^{\frac{\eta_2(2+\delta)}{2+\delta}}}{n}\right)^2\right\}\right) \quad (2 \leq c \leq s).$$

**Proof.** The proof is straight forward following lemma 2 of Yoshihara (1976) by replacing $h$ by

$$\frac{h}{\max\left\{p^{\frac{\eta_1(2+\delta)}{2+\delta}}, p^{\frac{\eta_2(2+\delta)}{2+\delta}}\right\}}.$$

as the kernel function. □

# Bibliography

Adak, S. and Sarkar, A. (1996). A time-frequency search for stock market anomalies. In *Time-Frequency Representations: Algorithms and Applications, rm (Tolimieri, R. and An, M. Eds.)*. Boston, MA: Birkhauser.

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association* **93**, 1488–1501.

Anderson, T.W.(1984). An introduction to multivariate statistical analysis. New York: Wiley.

Andreou, E. and Ghysels, E. (2008). Structural breaks in financial time series. In *Handbook of Financial Time Series,* (Anderson, T. G., Davis, R. A., Kreiss, J. P. and Mikosch, T. Eds.) Berlin: Springer, pp. 839–866.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821–856.

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, **37**, 4046–4087.

Bai, J. (1997). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics* **4**, 551–563.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structure changes. *Econometrica* **66**, 47–78.

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

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

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structure change models. *Journal of Applied Econometrics* **18**, 1–22.

Bai, J. and Shi, S. (2011). Estimating high dimensional covariance matrices and its applications. *Annals of economics and finance*, **12-2**, 199-215.

Baltagi, B., Song, S.H., Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics*, **117**, 123-150.

Basseville, M. and Benveniste, A. (1983). Sequential segmentation of nonstationary digital signals using spectral analysis. *Informatio Sciences* **29**, 57–73.

Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Annal of Statistics*, **36(6)**, 2577-2604.

Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Boysen, L., Kempe, A., LiebsCher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Annals of Statistics* **37**, 157–183.

Brockwell, P.J. and Davis, R.A. (1990). Time series: Theory and methods (2nd ed.). *Springer, New York*

Cai, T. and Liu, W. (2011). Adaptive Thresholding for Sparse Covariance Matrix Estimation. *Journal of Amercian Statistical Association*, **106**, 672–684.

Chan, N. H., Yau, C. Y. and Zhang, R. M. (2014). Group LASSO for structural break time series. *Journal of the American Statistical Association*, **109**, 590–599.

Chang, J., Chen, S.X. and Chen X. (2014). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, to appear.

Chang, J., Guo, B., and Yao, Q. (2014). Segmenting multiple time series by contemporaneous linear transformation *Manuscript*.

Chen, J. and Gupta, A. K (1997). Testing and locating variance change-points with application to stock prices. *Journal of the American Statistical Association* **92**, 739–747.

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society*, **B**, to appear.

Cliff, A.D. and Ord, J.K. (1973). Spatial autocorrelation. *Pion Ltd., London*

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for non-stationary time series. *Journal of the American Statistical Association* **101**, 223–239.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis* **29**, 834–867.

Dehling, H. and Wendler, M. (2010). Central limit theorem and bootstrap for $U$-statistics of strong mixing data. *Journal of Multivariate Analysis*, **101**, 126–137.

El Karoui, N. (2007). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annal of Statistics*, **36(6)**, 2717–2756.

Fan, J., Lv, J. and Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of economics* **3**, 291–317.

Fan, J. and Yao, Q. (2003). Nonlinear time series analysis: nonparameteric and parametric methods. *Springer, New York.*

Fan, J., Liao, Y. and Micheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements (with discussion). *Journal of Royal Statistical Society B*, **75**, 603–680.

Golub, G. and Van Loan, C. (1996). Matrix Computations (3rd ed.). *Johns Hopkins University Press.*

Hallin, M. and Liska, R.(2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, **102**, 603–617.

Hamilton, J.D. (1994) Time series analysis. *Princeton University Press, Princeton, NJ.*

Han, F. and Liu, H. (2013). Transition Matrix Estimation in High Dimensional Vector Autoregressive Models. International Conference on Machine Learning (ICML), 30, 2013.

Harchaoui, Z. and Levy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* **105**, 1480–1493.

Hastie, T., Tibshirani, R. and Friedman, J.(2001). The elements of statistical learning: data mining, inference and prediction. New York: Springer Verlag.

Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics.*, **19**, 293–325.

Hsu,N. J., Hung, H. L., and Chang, Y. M.(2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data analysis.***52(7)**, 3645–3657.

Jolliffe, I.(1986). Principal component analysis. New York: Springer Verlag.

Kelejian, H.H. and Prucha, I.R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, **157**, 53–67.

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

Lam, C., Yao, Q and Bathia, N(2011). Estimation of latent factors for hign-dimensional time series. *Biometrika*, **98**, 901–918.

Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Lee, L.F. (2001). Generalized method of moments estimation of spatial autoregressive processes. *Manuscripts, Department of Economics, OSU, August 2001.*

Lee, R.D., Carter, L.R. (1992) Modelling and forecasting US mortality. *Journal of the American Statistical Association*, **87**, 659–671.

Lee, L.F. and Yu, J. (2010a). Some recent developments in spatial panel data models. *Regional Science and Urban Economics*, **40**, 255–271.

Lee, L.F. and Yu, J. (2010b). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, **154**, 165–185.

Lee, L.F. and Yu, J. (2013). Near Unit Root in the Spatial Autoregressive Model. *Spatial Economic analysis*, **8**, 314–351.

Liu, J. M., Chen, R. and Yao, Q (2010). Nonparametric transfer function models. *Journal of Econometrics*, **157**, 151–164.

Lütkepohl, H. (2006) *New introduction to Multiple Time Series Analysis*, Springer.

Mackey, L. (2009). Deflation methods for sparse PCA. *Advances in Neural Information Processing systems (NIPS)*, **21**, 1017–1024.

Merlevede, F., Peligrad, M. and Rio, E. (2011). A bernstein type inequaliy and moderate deviations for weakly dependent sequences *probabilty Theory and related Fields*, **151**, 435–474.

Moghaddam, B., Weiss, Y. and Avidan, S. (2006). Generalized spectral bounds for sparse LDA. *In Proc. ICML*, **2006**.

Onatski, A. (2015) Asymptotic analysis of the squared estimation error in misspecified factor models. *Journal of Econometrics*, **186**, 388–406.

Pan, J and Yao, Q(2008). Modelling multiple time series via common factors. *Biometrika*, **95**, 365–379.

Pena, D. and Box. G.(1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**, 836–843.

Pham, T.D. and Tran, L.T. (1985). Some mixing properties of time series models. *Stochastic Processes and Their Applications*, **19**, 279–303.

Rothman, A.J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of Amercian Statistical Association*, **104**, 177–186.

Serfling, R. J. (1980) Approximation Theorems of Mathematical Statistics. *New York: Wiley.*

Stock, J. H. and Watson, M (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, **97**, 1167–1179.

Su, L. (2012). Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics*, **167**, 543–560.

Vert, J. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group LARS. Advances in Neural Information Processing Systems, **23**, 2343–2351.

Vu, V. and Lei. J.(2012). Minimax Rates of Estimation for Sparse PCA in High Dimensions. http://arxiv.org/abs/1202.0786v2.

Vu, V. and Lei. J.(2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, **41**, 2905–2947.

Wang, Z., Han, F. and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. arXiv:1307.0164v1.

Xu, L. and Lee, L.F. (2010). GMM estimation of spatial autoregressive models with unknown heterokedasticity. *Journal of Econometrics*, **177**, 34–52.

Yao, Y-C. (1998) Estimating the number of change-points via schwarz criterion. *Statistics & Probability Letters*, **6**, 181–189.

Yoshihara, K. (1976) Limiting behaviour of U-statistics for stationary,absolutely regular processes. *Z. Wahrsch. view. Gebiete*, **35**, 237–252.

Yu, J., de jong, R., Lee, L.F. (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both $n$ and $T$ are large. *Journal of Econometrics*, **146**, 118–134.

Yu, J., de jong, R., Lee, L.F. (2012) Estimation for spatial dynamic panel data with fixed effects: the case of spatial cointegration. *Journal of Econometrics*, **167**, 16–37.

Yuan, X. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, **14**, 899–925.

Zhong, P. and Chen, S. (2011) Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association*, **106:493**, 260–274.