SOME TOPICS ON TIME SERIES ANALYSIS

BY CHUAN LIU

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Statistics

Written under the direction of
Han Xiao
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

Some Topics on Time Series Analysis

by Chuan Liu

Dissertation Director: Han Xiao

This thesis deals with three problems. The first problem is concerned with the classical white noise test, and the second is on the estimation of autoregressive models for matrix-valued time series. These two time series problems are treated in Chapter 2 and 3. The third problem, on the construction of exact algorithm for the linear constrained LASSO, is treated in Chapter 4.

In Chapter 2, we study the asymptotic distribution of sample canonical correlations under different distributional assumptions of the time series. The joint density of the asymptotic distribution is derived explicitly for the normal and elliptical distributions. For the general non-normal case, we propose to bootstrap the canonical correlations to obtain the p-value. We carry out an extensive simulation study to illustrate the size and power performances of the proposed tests.

In Chapter 3, we propose a novel estimator of the matrix autoregressive model, based on the weighted least squares. We derive the asymptotic distributions of the estimator, and demonstrate its performance by simulations and real examples.

Chapter 4 deals with the exact algorithm of the constrained LASSO problem.

We develop such an exact algorithm by exploring the geometric properties of the problem. We prove that the solution path of the problem is piece-wise linear. We also prove an exponential upper bound for the complexity of the constrained LASSO problem.

Acknowledgements

My deepest gratitude to my advisor Prof. Han Xiao for his patient guidance and support through out my Ph.D. journey. This dissertation might not be possible to be finished without his help and encouragement. I am especially grateful to the tremendous effort he has put in introducing me to several topics of times series and high dimensional statistics. Many fellow students are amazed by his enthusiastic teaching style and his ability to convey complex ideas in a crystal clear manner. And so do I. But I know there are more than that. To quote the famous inscription written by Yinke Chen, "Scholars strive to free themselves from the shackles of the common sense so as to develop and promote the truth through study and research. It is better to die fighting for free thought". In this respect, I may have let him down. But there is still journey ahead in my life, though I might not be able to be a good scholar, I have faith to be a qualified statistics practitioner, a person who will not distort his learnings catering to anyone's taste. I will cherish the lessons I learned from him, the notes I kept in his classes and the memory of eating the twice-cooked pork slices that he ordered for us.

My heartfelt thanks to Prof. Minge Xie for his generous advice and many useful conversations that I had with him during my Rutgers days. I would also especially like to thank Prof. David Tyler for teaching me in the early days and introducing me to the subject of multivariate analysis. I would also like to thank Prof. Xiaodong Lin for agreeing to be a part of the committee.

I have learned a lot through conversations with some fellow students during the last few years. To name a few, I have benefited from the conversations with

Dongming She, Justin Gilmer, Ritwik Mitra and Long Feng. Let me also thank them here.

Last but in no way least, I thank my parents and fiancee Han for their consistent support and unconditional love.

Dedication

To my parents Yong Liu, Xia Zou and my fiancee Han Liu.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

This dissertation includes two projects on multivariate time series analysis and one project on regularized quadratic optimization. The first project is on the multivariate white noise testing. The second project is about the weighted least square estimation of the matrix autoregressive models. The third one concerns the exact algorithm and complexity of the constrained LASSO.

One important question to ask in finance is if the efficient market hypothesis is consistent with empirical data. Testing the predictability of returns is a usual way to verify this hypothesis, since under an ideally efficient market, the returns are expected to exhibit no linear dependencies. In statistics this has been formulated and extensively studied as the white noise testing problem. In Chapter 2 we propose a new multivariate white noise test based on the canonical correlations. To be more precise, given a multivariate time series $\{\boldsymbol{X}_i\}_{i=1}^{T}$, we construct a new series $\{\boldsymbol{Y}_i \equiv (\boldsymbol{X}_i^t, \boldsymbol{X}_{i+1}^t)^t\}_{i=1}^{T-1}$ and perform canonical correlation analysis on this news series regarding the natural partition of components in its definition. Starting from the simple normal distribution assumption, we derive the asymptotic joint distribution of the sample canonical correlations calculated based on $\{\boldsymbol{Y}_i\}_{i=1}^{T-1}$. Then by combining the ideas and techniques in the proof of the normal case with the result in Eaton and Tyler (1994), we extend our results to the elliptical distribution case. For the general non-normal distribution case, such asymptotic distributions are much harder to establish so we use a bootstrap procedure as a work-around to construct the test we need.

With the advent of technology revolution, large amounts of data have become readily available and the need to extract meaningful and practical insights from the data has also become paramount with the influx of data. For example, data in finance and economics are sometimes observed in matrix format. To model such data in time series, one can forget about its matrix nature and use the traditional vector autoregressive (VAR) models. But in many scenarios, the column and row variables in the matrix data have close interactions but different interpretations. So it is more natural to keep and take advantage of the matrix structure when modeling this type of data. Chen et al. (2020) proposed a novel matrix autoregressive (MAR) model for this type of data. Based on their work, in Chapter 3, we introduce a new weighted least square estimator for the MAR model. The weights correspond to the marginal sample variances of the residual matrices. We carry out the asymptotic analysis of this new estimator, and demonstrate its superior performance over other estimators by an extensive numerical study and an example on economic indicators.

Markowitz mean-variance efficient portfolio construction (Markowitz (1952)) and the linear discriminant analysis can both be formulated as a quadratic optimization problem with linear constraints. When the number of assets or features is large, the direct implementation of this optimization cannot provide a consistent estimator of the weight vector, leading to inferior performance of the portfolio or the classifier. This is in particular due to the poor estimation of the high dimensional covariance matrix. One way to solve this problem is to incorporate better estimators of the mean vector and the covariance matrix, see for example Shao et al. (2011). Another approach is to consider the regularized optimization by penalizing the ℓ_1 norm of the weight vector (Fan et al., 2012b,c,a; Brodie et al., 2009), leading to the so called constrained LASSO problem. In Chapter 4, we propose an exact algorithm for the constrained LASSO problem, and investigate the complexity of this algorithm. This exact algorithm is motivated by and builds

upon the least angle regression proposed by the seminal work Efron et al. (2004). We prove an exponential upper bound of the complexity based on the work of Mairal and Yu (2012), and show that it is achievable by constructing a concrete worst case example.

Chapter 2

White Noise Test Based on Canonical Correlations

2.1 Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$ be observations from a p-dimensional weakly stationary time series satisfying $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$. The white noise test is testing

$$H_0 : \mathrm{Cov}(\boldsymbol{X}_{t+\tau}, \boldsymbol{X}_t) = \boldsymbol{0}, \tau \geq 1. \tag{2.1}$$

Since not all the cross covariance matrices are estimable with a given observed series, the test usually only involves lags $\tau = 1, 2, \ldots, m$, where $m \geq 1$ is a prescribed integer.

In the model diagnostic procedure of times series analysis, we often need to perform a white noise test on the residuals of a fitted model for the purpose of checking whether the model is adequate. Various testing procedures have been proposed in the literature for the univariate time series, including the Box Pierce test (Box and Pierce, 1970), the Ljung Box test (Ljung and Box, 1978), and the Lagrange Multiplier test (Breusch, 1978) and (Godfrey, 1978). Extensions of these tests have also been developed for multivariate time series by Hosking (1981), Hosking (1980) and Li and McLeod (1981), among many others. In this Chapter we propose a new approach to the white noise test for multivariate time series, based on *canonical correlations*.

Based on the series $\{\boldsymbol{X}_i\}_{i=1}^T$, one can form a new series $\{\boldsymbol{Y}_i \equiv (\boldsymbol{X}_i', \boldsymbol{X}_{i+1}')'\}_{i=1}^{T-1}$

by concatenating adjacent observations in the original series $\{\boldsymbol{X}_i\}_{i=1}^{T}$. For the new series $\{\boldsymbol{Y}_i\}_{i=1}^{T-1}$, since each observation has been partitioned into two size p parts (the first p components and the last p components), canonical correlations analysis can be naturally performed based on these two sets of variables. Under the null hypothesis (2.1), if the asymptotic distributions for the sample canonical correlations can be established, then we can use these sample canonical correlations to construct test statistics (for example, the largest sample canonical correlation) to test the validity of null.

We start with a strong *normal distribution* assumption for the *i.i.d* series $\{\boldsymbol{X}_i\}_{i=1}^{T}$. We proved the following asymptotic distribution result for the sample canonical correlations calculated based on $\{\boldsymbol{Y}_i\}_{i=1}^{T-1}$.

Theorem. *Let $\hat{r}_1 \geq \hat{r}_2 \geq \cdots \geq \hat{r}_p$ be the sample canonical correlations calculated from the series $\{\boldsymbol{Y}_i\}_{i=1}^{T-1}$ and denote $w_i \equiv T\hat{r}_i^2$ for $i = 1, 2, ..., p$. Under the i.i.d normal assumptions, the limiting joint density function of $w_1, ..., w_p$ as $T \to \infty$ is*

$$\frac{\pi^{p^2/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} w_i\right)}{2^{p^2/2}\Gamma_p(p/2)\Gamma_p(p/2)} \cdot \prod_{i=1}^{p} w_i^{-1/2} \prod_{j<i,j=1}^{p} (w_j - w_i), \tag{2.2}$$

where $w_1 \geq w_2 \geq ... \geq w_p \geq 0$ and $\Gamma_n(.)$ denotes the multivariate gamma function.

Combining the techniques we introduced in the proof of the previous theorem and a random matrix result established in Eaton and Tyler (1994), we can extend our result to the scenario when the series $\{\boldsymbol{X}_i\}_{i=1}^{T}$ is *i.i.d* elliptical. The result can be state as follows.

Theorem. *Let $\hat{r}_1 \geq \hat{r}_2 \geq \cdots \geq \hat{r}_p$ be the sample canonical correlations calculated from the series $\{\boldsymbol{Y}_i\}_{i=1}^{T-1}$ and denote $w_i \equiv T\hat{r}_i^2$ for $i = 1, 2, ..., p$. Under the i.i.d elliptical assumptions, the limiting joint density function of $w_1, ..., w_p$ as $T \to \infty$ is*

$$\frac{\pi^{p^2/2}}{(2+2\gamma)^{p^2/2}\Gamma_p(p/2)\Gamma_p(p/2)}\exp(-\sum_{i=1}^{p}\frac{w_i}{2+2\gamma})\prod_{i=1}^{p}w_i^{-1/2}\prod_{i<j}^{p}(w_i-w_j) \qquad (2.3)$$

where $w_1 \geq w_2 \geq ... \geq w_p \geq 0$ *and* $\Gamma_n(.)$ *denotes the multivariate gamma function.*

Note that even with the asymptotic joint distribution results for $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p$ under the normal or elliptical distribution assumptions, expressing the marginal distributions (like the largest sample canonical correlation $\hat{r}_1$) explicitly might not be an easy task, not to mention the fact that when the underlying distribution of the series $\{\boldsymbol{X}_i\}_{i=1}^{T}$ belongs to general non-normal family, the asymptotic distribution results for these sample canonical correlations can be difficult to establish.

Fortunately, we have a tailored bootstrap procedure for the general non-normal scenario and it is validated by the following consistency result.

Theorem. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^{T}$ *to be a series of i.i.d random vector with finite fourth moment* $\mathbb{E}\|\boldsymbol{X}_i\|^4 < \infty$ *and* $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$*. Then along almost all the sample sequences* $\omega = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T, \ldots)$*, as* T *tends to* ∞*, we have the difference between* $\mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \leq c|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T)$ *and* $\mathbb{P}(\sqrt{T}\hat{r_1} \leq c)$ *converges uniformly to 0, or equivalently we have*

$$\sup_{-\infty<c<\infty}|\mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \leq c|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T) - \mathbb{P}(\sqrt{T}\hat{r_1} \leq c)| \overset{a.s.}{\to} 0.$$

Based on this consistency result, under the general distribution assumptions, we propose to bootstrap the series $\{\boldsymbol{X}_i\}_{i=1}^{T}$ to calculate a series of values of bootstrapped largest sample canonical correlation and use its empirical quantile as threshold for testing the hypothesis (2.1).

The rest of this chapter is organized as follows. In Section 2.2 we discuss our approach under normality and derive the asymptotic distribution of the lag-1 sample canonical correlations. In Section 2.3 we consider the extension to elliptical distributions, and establish the corresponding asymptotic results. For more

general distributions (with only a moment condition), we propose to bootstrap the sample canonical correlations in Section 2.4 and 2.5, and prove the consistency of the bootstrap test. We carry out an extensive simulation study in Section 2.6 to compare the sizes and powers of the proposed tests with other state-of-art testing procedures in literature. The proof are relegated in Section 2.7.

2.2 The Multivariate Normal Case

Let $\{\boldsymbol{X}_i\}_{i=1}^T$ be an $i.i.d$ sequence of multivariate normal time series in $\mathbb{R}^p$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_x = \{\sigma_{ij}\}$ and denote $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$. Now we consider the lag-1 series $\{\boldsymbol{Y}_i = (\boldsymbol{X}_i', \boldsymbol{X}_{i+1}')'\}_{i=1}^{T-1}$ and construct the sample covariance matrix from the following matrix of lag-1 sample (note that this sample has serial correlations):

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{X}_2' \\ \boldsymbol{X}_2' & \boldsymbol{X}_3' \\ \vdots & \vdots \\ \boldsymbol{X}_{T-2}' & \boldsymbol{X}_{T-1}' \\ \boldsymbol{X}_{T-1}' & \boldsymbol{X}_T' \end{pmatrix} \tag{2.4}$$

which is of dimension $(T-1) \times 2p$.

Here our target is to investigate the asymptotic distribution of sample canonical correlations of the the lag-1 correlated $\boldsymbol{Y}$ sample in (2.4) and see whether it is same as the independent case. And now we explain what does the **independent case** here mean.

From the same series $\{\boldsymbol{X}_i\}_{i=1}^T$, this time we consider a new random vector series $\{\boldsymbol{V}_i = (\boldsymbol{X}_i', \boldsymbol{U}_i')'\}_{i=1}^{T-1}$ where the series $\{\boldsymbol{U}_i\}$ is an independent copy of the series $\{\boldsymbol{X}_i\}$. Based on this series, we introduce a new matrix $\boldsymbol{V}$ from the independent sample $\{\boldsymbol{V}_i = (\boldsymbol{X}_i', \boldsymbol{U}_i')'\}_{i=1}^{T-1}$ where

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{X}_1' & \boldsymbol{U}_1' \\ \boldsymbol{X}_2' & \boldsymbol{U}_2' \\ \vdots & \vdots \\ \boldsymbol{X}_{T-2}' & \boldsymbol{U}_{T-2}' \\ \boldsymbol{X}_{T-1}' & \boldsymbol{U}_{T-1}' \end{pmatrix} \tag{2.5}$$

which is also of dimension $T \times 2p$. This is what we mean by **independent case** and we denoted by (2.5).

For the purpose of the later discussions, let us denote the mean zero sample covariance matrix (we use T instead of $T-1$ in the denominator of the expression for simplicity and this clearly does not affect the asymptotic distribution of the sample covariance matrix) in the lag-1 case (2.4) by $\boldsymbol{N}$ and

$$\boldsymbol{N} \equiv \frac{1}{T}\boldsymbol{Y}'\boldsymbol{Y} = \frac{1}{T}\begin{pmatrix} \sum_{i=1}^{T-1} \boldsymbol{X}_i\boldsymbol{X}_i' & \sum_{i=1}^{T-1} \boldsymbol{X}_i\boldsymbol{X}_{i+1}' \\ \sum_{i=1}^{T-1} \boldsymbol{X}_{i+1}\boldsymbol{X}_i' & \sum_{i=1}^{T-1} \boldsymbol{X}_{i+1}\boldsymbol{X}_{i+1}' \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{N}_{11} & \boldsymbol{N}_{12} \\ \boldsymbol{N}_{21} & \boldsymbol{N}_{22} \end{pmatrix}.$$

For the independent case (2.5), we denote the mean zero covariance matrix by $\boldsymbol{M}$ and

$$\boldsymbol{M} \equiv \frac{1}{T}\boldsymbol{V}'\boldsymbol{V} = \frac{1}{T}\begin{pmatrix} \sum_{i=1}^{T-1} \boldsymbol{X}_i\boldsymbol{X}_i' & \sum_{i=1}^{T-1} \boldsymbol{X}_i\boldsymbol{U}_i' \\ \sum_{i=1}^{T-1} \boldsymbol{U}_i\boldsymbol{X}_i' & \sum_{i=1}^{T-1} \boldsymbol{U}_i\boldsymbol{U}_i' \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{pmatrix}.$$

Before we discuss our main results, we state a result from matrix theory here which will be used later.

Proposition 1. *Suppose we have two $p \times p$ matrices A and B and λ_{Ai} and λ_{Bi} are the i-the largest eigenvalue of A and B respectively. Then we have*

$$\sum_{i=1}^{p} |\lambda_{Ai} - \lambda_{Bi}|^2 \leq ||A - B||_2^2$$

where the norm $||.||_2$ is the Frobenius norm.

For the lag-1 sample in (2.4), the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_y = \frac{1}{T}\boldsymbol{Y}'\boldsymbol{Y} - \frac{1}{T^2}\boldsymbol{Y}'\mathbf{1}\mathbf{1}'\boldsymbol{Y}$. Note that if we multiply the second term by $\sqrt{T}$, then the term $\frac{1}{T^{3/2}}\boldsymbol{Y}'\mathbf{1}\mathbf{1}'\boldsymbol{Y} \xrightarrow{p} \mathbf{0}$ when $T \to \infty$ by law of large numbers, the central limit theorem and Slusky's theorem. So the key here is the first term $\frac{1}{\sqrt{T}}\boldsymbol{Y}'\boldsymbol{Y}$, or $\sqrt{T}\boldsymbol{N}$ in our notation. Also note that the true covariance structure of the lag-1 random vector $\boldsymbol{Y}_i$ is

$$\boldsymbol{\Sigma}_y = \begin{pmatrix} \boldsymbol{\Sigma}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_x \end{pmatrix}.$$

Now we are going to investigate how the off-diagonal block of $\boldsymbol{N}$ and $\boldsymbol{M}$ relate to the distribution of canonical correlations in the two cases. But we need two lemmas on the asymptotic normality of the mean zero sample covariance matrix for the two cases (2.4) and (2.5) respectively before proceeding. Since the independent case is more simpler than the lag-1 case, we are going to present the result for the independent case (2.5) first. Here we note that in order to establish the asymptotic normality for the case (2.4) as in Lemma 2, we are going to use Cramer-Wold device and a martingale central limit theorem result from Hall and Heyde (2014).

For the case (2.5) we have the following result:

Lemma 1. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^T$ *be an i.i.d sequence of multivariate normal time series in* $\mathbb{R}^p$ *with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}_x$*. Then for the case (2.5), the random matrix* $\boldsymbol{M}$ *(in its vectorized sense* $\mathrm{vec}(\boldsymbol{M})$*) has the following asymptotic distribution property:*

$$\sqrt{T}(\boldsymbol{M} - \boldsymbol{\Sigma}_y) \xrightarrow{d} \boldsymbol{D}_0$$

where $\boldsymbol{D}_0$ *follows* $N_{2p}^{2p}(0, (\boldsymbol{I} + \boldsymbol{K}_{2p})(\boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_y))$ *and* $\boldsymbol{K}_n$ *is the* ***commutation matrix****, a block matrix whose block in position* (i, j) *is* $\boldsymbol{e}_j\boldsymbol{e}_i' \in \mathbb{R}_n^n$*,*

$$\boldsymbol{K}_n = (\boldsymbol{e}_j\boldsymbol{e}_i') \in \mathbb{R}_{n^2}^{n^2}.$$

.

For the case (2.4) we have the following result:

Lemma 2. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^T$ *be an i.i.d sequence of multivariate normal time series in* $\mathbb{R}^p$ *with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}_x$*. Then for the case (2.4), the matrix* $\boldsymbol{N}$ *(or in its vectorized sense* $\mathrm{vec}(\boldsymbol{N})$ *) has the following asymptotic distribution property:*

$$\sqrt{T}(\boldsymbol{N} - \boldsymbol{\Sigma}_y) \xrightarrow{d} \boldsymbol{D}.$$

Here $\boldsymbol{D}$ *follows* $N_{2p}^{2p}(0, (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}))(\boldsymbol{I} + \boldsymbol{K}_{2p} + \boldsymbol{J} + \boldsymbol{J}')(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}))$ *where* $\boldsymbol{\Gamma} \equiv \boldsymbol{\Sigma}_y^{1/2}$ *and*

$$\boldsymbol{J} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{J}_0 & \mathbf{0} \end{pmatrix}$$

which is of size $4p^2 \times 4p^2$ *where* $\boldsymbol{J}_0$ *is a* $p \times p$ *partitioned block matrix of size* $2p^2 \times 2p^2$ *with the* (i, j) *block equal to*

$$= \begin{cases} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{e}_j \boldsymbol{e}_i' & \mathbf{0} \end{pmatrix}, & \text{when } i \neq j \\ \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_p + \boldsymbol{e}_i \boldsymbol{e}_i' & \mathbf{0} \end{pmatrix}, & \text{when } i = j \end{cases}$$

and $\boldsymbol{K}_n$ *is the* ***commutation matrix****, a block matrix whose block in position* (i, j) *is* $\boldsymbol{e}_j \boldsymbol{e}_i' \in \mathbb{R}_n^n$*,*

$$\boldsymbol{K}_n = (\boldsymbol{e}_j \boldsymbol{e}_i') \in \mathbb{R}_{n^2}^{n^2}.$$

.

With the above preparations, we can continue our discussions on how the distribution of the sample canonical correlations are related under the two cases.

From the canonical correlation sections of the multivariate statistics textbooks like Anderson (2003) and Bilodeau and Brenner (2008), we first note that the squared sample canonical correlations are the roots of the following equation

$$|\boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.6}$$

and we know from Lemma 2 that $T^{1/2}(\boldsymbol{N} - \boldsymbol{\Sigma}_y) \xrightarrow{d} \boldsymbol{D}$ for some distribution $\boldsymbol{D}$ of which the entries are jointly normally distributed. Now if we consider the normalized transformation $\boldsymbol{Z}_i = \boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{X}_i$ and replace the corresponding $\boldsymbol{X}_i$ by $\boldsymbol{Z}_i$ in the determinant equation (2.6), we notice that such a transformation does not change the roots of the equation. So without loss of generality, we can assume $\boldsymbol{\Sigma}_x = \boldsymbol{I}_p$.

Let us denote $T^{1/2}(\boldsymbol{N}-\boldsymbol{\Sigma}_y)$ by $\boldsymbol{W}$ with blocks $\boldsymbol{W}_{11}$, $\boldsymbol{W}_{12}$, $\boldsymbol{W}_{21}$, $\boldsymbol{W}_{22}$ similarly partitioned as $\boldsymbol{N}$. Now we can write

$$\boldsymbol{N}_{11} = \boldsymbol{I}_p + T^{-1/2}\boldsymbol{W}_{11}$$

$$\boldsymbol{N}_{22} = \boldsymbol{I}_p + T^{-1/2}\boldsymbol{W}_{22}$$

$$\boldsymbol{N}_{12} = T^{-1/2}\boldsymbol{W}_{12}.$$

Then from the matrix power series expansion

$$(\boldsymbol{I} - t\boldsymbol{A})^{-1} = \sum_{i=0}^{\infty} t^i \boldsymbol{A}^i,$$

we have the inverse identities

$$\boldsymbol{N}_{11}^{-1} = \boldsymbol{I}_p - T^{-1/2}\boldsymbol{W}_{11} + O_p(T^{-1})$$

and

$$\boldsymbol{N}_{22}^{-1} = \boldsymbol{I}_p - T^{-1/2}\boldsymbol{W}_{22} + O_p(T^{-1}).$$

So we can expand and calculate $\boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21}$ using the above equalities and obtain

$$\boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} = T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} + O_p(T^{-3/2}). \tag{2.7}$$

The above arguments work similarly for the independent case (2.5). For this case, let us denote $T^{1/2}(\boldsymbol{M}-\boldsymbol{\Sigma}_y)$ by $\boldsymbol{K}$ with blocks $\boldsymbol{K}_{11}$, $\boldsymbol{K}_{12}$, $\boldsymbol{K}_{21}$, $\boldsymbol{K}_{22}$ similarly partitioned as $\boldsymbol{M}$. So we can write

$$\boldsymbol{M}_{11} = \boldsymbol{I}_p + T^{-1/2}\boldsymbol{K}_{11}$$

$$\boldsymbol{M}_{22} = \boldsymbol{I}_p + T^{-1/2}\boldsymbol{K}_{22}$$

$$\boldsymbol{M}_{12} = T^{-1/2}\boldsymbol{K}_{12}.$$

Then from the same matrix power series expansion, we have the inverse identities

$$\boldsymbol{M}_{11}^{-1} = \boldsymbol{I}_p - T^{-1/2}\boldsymbol{K}_{11} + O_p(T^{-1})$$

and

$$\boldsymbol{M}_{22}^{-1} = \boldsymbol{I}_p - T^{-1/2}\boldsymbol{K}_{22} + O_p(T^{-1}).$$

So we can expand and calculate $\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}$ using the above equalities and obtain

$$\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21} = T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} + O_p(T^{-3/2}). \tag{2.8}$$

If we can prove the a $O_p(T^{-3/2})$ matrix perturbation on $T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21}$ does not change the asymptotic distribution of the roots of

$$|T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} - \lambda\boldsymbol{I}_p| = 0$$

then we might be able to work on the block $\boldsymbol{M}_{12}$ instead of the whole matrix.

From now on we denote the true ordered canonical correlations under the two cases to be $r_1, r_2, \ldots, r_p$ (actually they all equal to 0 by our i.i.d assumption), the corresponding ordered sample canonical correlations to be $\hat{r}_1 \geq \hat{r}_2 \geq \cdots \geq \hat{r}_p$ and adopt the expansion approach illustrated above for both cases to prove our main theorem which can be stated as follows:

Theorem 1. *The asymptotic joint distribution of the sample canonical correlations calculated from (2.4) and (2.5) are identical. Let* $w_i = T\hat{r}_i^2$ *for* $i = 1, 2, ..., p$. *Then the limiting joint density function of* $w_1, ..., w_p$ *as* $T \to \infty$ *is*

$$\frac{\pi^{p^2/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} w_i\right)}{2^{p^2/2}\Gamma_p(p/2)\Gamma_p(p/2)} \cdot \prod_{i=1}^{p} w_i^{-1/2} \prod_{j<i,j=1}^{p} (w_j - w_i), \tag{2.9}$$

where $w_1 \geq w_2 \geq ... \geq w_p \geq 0$ *and* $\Gamma_n(.)$ *denotes the multivariate gamma function.*

Among all the sample canonical correlation, we are often most interested in the largest one. Here we present a result on the distribution of $\hat{r}_1^2$ when $r_1 = r_2 = ... = r_p = 0$.

Proposition 2. *Under the assumptions of Lemma 1, with further condition that* $t = \frac{1}{2}(T - 2p - 2)$ *is a positive integer, the distribution function of* $\hat{r}_1^2$ *can be expressed as*

$$P(\hat{r}_1^2 \leq x) = x^{p^2/2} \sum_{k=0}^{pt} \sum_{\kappa}^{\star} 2^{2k}(1-x)^k (\frac{1}{2}p)_\kappa (\frac{1}{2}p)_\kappa \frac{\prod_{i<j}^{s}(2k_i - 2k_j - i + j)}{\prod_{i=1}^{s}(2k_i + s - i)!} \tag{2.10}$$

where $\kappa = (k_1, k_2, ..., k_p)$ *is an descending ordered partition of* k, $\sum_{\kappa}^{\star}$ *denotes summation over those partitions with largest part* $k_1 \leq t$, s *is the number of nonzero parts of the partition* κ *of* k *and* $(a)_\kappa$ *is the generalized hypergeometric coefficients given by*

$$(a)_\kappa = \prod_{i=1}^{p} (a - \frac{1}{2}(i-1))_{k_i} \tag{2.11}$$

where $(a)_k = a(a+1)...(a+k-1)$, $(a)_0 = 1$.

Proof. See Corollary 11.3.4 of Muirhead (2009). □

2.3 The Multivariate Elliptical Case

We assume normality for the results presented in the previous section, but a question of theoretical and practical importance is the effect of non-normality

on the problem. In this section, we are going to investigate the elliptical distribution case. The elliptical distribution can be viewed as a generalization of the multivariate normal distribution and are usually defined as follows:

Definition 1 (Elliptical distribution)**.** *A random vector $\boldsymbol{x} \in \mathbb{R}^p$ follows an elliptical distribution with location parameter $\boldsymbol{\mu}$ and scale parameter $\boldsymbol{\Lambda}$ if it has a density of the form*

$$f_{\boldsymbol{x}}(\boldsymbol{x}) = |\boldsymbol{\Lambda}|^{-1/2} g[(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})],$$

where $g : [0,\infty] \to [0,\infty]$ is a fixed function independent of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda} = (\Lambda_{ij})$ and depends on $\boldsymbol{x}$ only through $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. We denote this elliptical distribution by $\boldsymbol{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$.

Here we list some useful properties of elliptical distribution in the form of proposition. Some of these properties are used in this section for our investigation. For the detailed discussions and proof of the properties, one can refer to Chapter 13 of Bilodeau and Brenner (2008).

Proposition 3. *The elliptical distributions have the following properties:*

1. *If we have $\boldsymbol{x} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, then the linear transformation $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{B}$ is a $p \times p$ matrix and $\boldsymbol{b} \in \mathbb{R}^p$ satisfies $\boldsymbol{y} \sim E_p(\boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}')$. In particular, $\boldsymbol{z} = \boldsymbol{\Lambda}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}) \sim E_p(\boldsymbol{0}, \boldsymbol{I}_p)$ has a rotationally invariant distribution.*

2. *If $\boldsymbol{z} \sim E_p(\boldsymbol{0}, \boldsymbol{I}_p)$ is rotationally invariant, then its characteristic function $\psi_{\boldsymbol{z}}(\boldsymbol{s})$ can be expressed as a function of $|\boldsymbol{s}|$. Or equivalently, there exists a function $\phi(.)$ such that $\psi_{\boldsymbol{z}}(\boldsymbol{s}) = \phi(\boldsymbol{s}'\boldsymbol{s})$. And we have $\boldsymbol{x} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{z} + \boldsymbol{\mu}$ has characteristic function $\psi_{\boldsymbol{x}}(\boldsymbol{s}) = \exp(i\boldsymbol{s}'\boldsymbol{\mu})\phi(\boldsymbol{s}'\boldsymbol{\Lambda}\boldsymbol{s})$. Moreover, if $\boldsymbol{z}$ has finite second moment, $\mathbb{E}(z) = \boldsymbol{0}$ and $\operatorname{var}\boldsymbol{z} = \alpha\boldsymbol{I}$ for some constant α, and we have $\mathbb{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\operatorname{var}\boldsymbol{x} = \alpha\boldsymbol{\Lambda}$. Here the constant $\alpha = -2\phi'(0)$.*

3. *The marginal and conditional distributions of a elliptical distribution are elliptical. To be more precise, let* $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')' \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ *with* $\boldsymbol{x}_i \in \mathbb{R}^{p_i}$, $i = 1, 2$, $p = p_1 + p_2$, *and partition* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Lambda}$ *as* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ *and* $\begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$. *Then we have* $\boldsymbol{x}_2 \sim E_p(\boldsymbol{\mu}_2, \boldsymbol{\Lambda}_{22})$ *and* $\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim E_p(\boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21})$.

If we consider a more general *i.i.d.* elliptical distribution instead of normal distribution for our problem and look at the approaches we use in the previous section, one might notice that if we could establish similar results on the asymptotic normality of the sample covariance matrix, the arguments for Theorem 1 could essentially work to produce a similar theorem for the *i.i.d.* elliptical case. So we are going to prove analogues of Lemma 2 and 1. Here we assume $\{\boldsymbol{X}_i\}_{i=1}^T$ to be an *i.i.d* sequence of elliptical time series in $\mathbb{R}^p$ with location parameter $\mathbf{0}$ and non-singular scale matrix parameter $\boldsymbol{\Lambda}_x$ (we denoted by $\boldsymbol{X}_i \sim E_p(\mathbf{0}, \boldsymbol{\Lambda}_x)$) and finite fourth moments exist for this underlying elliptical distribution. Then by Proposition 3, we have $\text{var}(\boldsymbol{X}_i) \equiv \boldsymbol{\Sigma}_x = \alpha\boldsymbol{\Lambda}_x$ for some constant $\alpha = -2\phi'(0)$ where $\phi(.)$ is the function introduced in the second property of Proposition 3. Throughout this section we denote $\boldsymbol{\Lambda}_y \equiv \begin{pmatrix} \boldsymbol{\Lambda}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_x \end{pmatrix}$ so we have $\boldsymbol{\Sigma}_y = \begin{pmatrix} \boldsymbol{\Sigma}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_x \end{pmatrix} = \alpha\boldsymbol{\Lambda}_y$ under our setup. We also denote $\boldsymbol{U}_i \equiv \boldsymbol{\Lambda}_x^{-1/2}\boldsymbol{X}_i$ so that $\{\boldsymbol{U}_i\} \sim E_p(\mathbf{0}, \boldsymbol{I}_p)$ is rotationally invariant. Without loss of generality, we assume $\text{var}(U_{11}) = 1$ where U_{11} is the first coordinate of $\boldsymbol{U}_1$. All other notations for the independent and lag-1 case in the last section are assumed in this section.

First we have a result on the asymptotic distribution of the sample covariance matrix for the independent case.

Lemma 3. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^T$ *be an i.i.d sequence of elliptical time series in* $\mathbb{R}^p$ *where* $\boldsymbol{X}_i \sim E_p(\mathbf{0}, \boldsymbol{\Lambda}_x)$ *and its finite fourth moments exist. Then for the case (2.5), the random matrix* $\boldsymbol{M}$ *(in the vectorized sense* $\text{vec}\,\boldsymbol{M}$*) has the following*

asymptotic distribution property:

$$\sqrt{T}(\operatorname{vec} \boldsymbol{M} - \boldsymbol{\Sigma}_y) \xrightarrow{d} \boldsymbol{D}_0$$

where $\boldsymbol{D}_0$ *follows* $N_{2p}^{2p}(0, (1+\kappa)(\boldsymbol{I} + \boldsymbol{K}_{2p})(\boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_y) + \kappa \operatorname{vec} \boldsymbol{\Sigma}_y [\operatorname{vec} \boldsymbol{\Sigma}_y]')$ *and the parameter* $\kappa \equiv \frac{\phi''(0) - \phi'(0)^2}{\phi'(0)^2} = \frac{\mathbb{E}z_1^4 - 3\mathbb{E}z_1^2}{3\mathbb{E}z_1^2}$.

Remark. *We note that for the normal special case, the function* $\phi(x) = e^{-x/2}$. *Hence it is easy to compute and derive that the parameters in the lemma satisfy* $\alpha = 1$ *and* $\kappa = 0$. *Thus the covariance structure produced here coincides with the covariance structure in Lemma 1.*

Second we have a result for the lag-1 elliptical case on the asymptotic distribution of the sample covariance matrix.

Lemma 4. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^T$ *be an i.i.d sequence of elliptical time series in* $\mathbb{R}^p$ *where* $\boldsymbol{X}_i \sim E_p(\boldsymbol{0}, \boldsymbol{\Lambda}_x)$ *and its finite fourth moments exist. Then for the case (2.4), the matrix* $\boldsymbol{N}$ *(in the vectorized sense* $\operatorname{vec} \boldsymbol{N}$*) has the following asymptotic distribution property:*

$$\sqrt{T}(\operatorname{vec} \boldsymbol{N} - \boldsymbol{\Sigma}_y) \xrightarrow{d} \boldsymbol{D}.$$

Here $\boldsymbol{D}$ *follows* $N_{2p}^{2p}(0, (1+\kappa)(\boldsymbol{I} + \boldsymbol{K}_{2p})(\boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_y) + \kappa \operatorname{vec} \boldsymbol{\Sigma}_y [\operatorname{vec} \boldsymbol{\Sigma}_y]' + (\boldsymbol{\Lambda}_y^{1/2} \otimes \boldsymbol{\Lambda}_y^{1/2})(\boldsymbol{L} + \boldsymbol{L}')(\boldsymbol{\Lambda}_y^{1/2} \otimes \boldsymbol{\Lambda}_y^{1/2}))$ *where*

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{L}_0 & \boldsymbol{0} \end{pmatrix}$$

which is of size $4p^2 \times 4p^2$ *and* $\boldsymbol{L}_0$ *is a* $p \times p$ *partitioned block matrix of size* $2p^2 \times 2p^2$ *with the* (i, j) *block equal to* $\operatorname{Cov}(z_{2i} \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', z_{2j} \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')')$ *and* $\kappa \equiv \frac{\phi''(0) - \phi'(0)^2}{\phi'(0)^2} = \frac{\mathbb{E}z_1^4 - 3\mathbb{E}z_1^2}{3\mathbb{E}z_1^2}$.

Remark. *Again for the normal case, we have* $\alpha = 1$ *and* $\kappa = 0$. *And with the special moments' properties of normal (up to the fourth moment), one can verify that* $\boldsymbol{L}_0 = \boldsymbol{J}_0$ *hence the results in this lemma coincide with results in lemma 2.*

With Lemma 3, 4 and distribution results from Eaton and Tyler (1994) and Muirhead (2009), we can establish the following theorem:

Theorem 2. *Under the assumptions made in this section, the asymptotic distributions of the sample canonical correlations calculated from (2.4) and (2.5) are still identical. Let* $\boldsymbol{Z}$ *be a* $p \times p$ *random matrix distributed as* $N(\mathbf{0}, (1+\gamma)I_p \otimes I_p)$ *where* $\gamma = (E(U_{11}^4) - 3)/3$ *is the kurtosis parameter. Then the asymptotic joint distribution of* $w_1, w_2, ..., w_p$ *where* $w_i = T\hat{r}_i^2$*, is the same as the joint distribution of the decreasingly ordered eigenvalues of the random matrix* $\boldsymbol{Z}\boldsymbol{Z}'$ *and its density is given as*

$$\frac{\pi^{p^2/2}}{(2+2\gamma)^{p^2/2}\Gamma_p(p/2)\Gamma_p(p/2)} \exp(-\sum_{i=1}^{p} \frac{w_i}{2+2\gamma}) \prod_{i=1}^{p} w_i^{-1/2} \prod_{j<i}^{p} (w_j - w_i) \qquad (2.12)$$

where $w_1 \geq w_2 \geq ... \geq w_p \geq 0$ *and* $\Gamma_n(.)$ *denotes the multivariate gamma function.*

Remark. *Note that when the distribution assumption in Theorem 2 is further restricted to be normal, we have the kurtosis parameter* $\gamma = 0$ *and the density in (2.12) now coincides with the density in (2.9).*

2.4 The Multivariate non-Normal Case

For the non-normal case, it is difficult to obtain asymptotic results for the sample canonical correlations like Theorem 2. But we can still construct a white noise test based on canonical correlations with the idea of bootstrapping.

In this section we assume $\{\boldsymbol{X}_i\}_{i=1}^T$ to be a general $i.i.d$ time series with finite fourth moment condition $\mathbb{E}\|\boldsymbol{X}_i\|^4 < \infty$ and $\mathbb{E}\boldsymbol{X}_i = \mathbf{0}$. To make our theoretical argument easier, first we centralize the series and denote $\boldsymbol{X}_i^c \equiv \boldsymbol{X}_i - \bar{\boldsymbol{X}}_i$, then we perform bootstrap on the centralized time series $\{\boldsymbol{X}_i^c\}_{i=1}^T$ to obtain a bootstrap sample $\{\boldsymbol{X}_{Ti}^*\}_{i=1}^T$ of size T. And we define two modified version of sample

covariance matrices as follow:

$$\boldsymbol{S}^{(b)} \equiv \frac{1}{T}\begin{pmatrix} \sum_{i=1}^{T} \boldsymbol{X}^*_{Ti}(\boldsymbol{X}^*_{Ti})' & \sum_{i=1}^{T-1} \boldsymbol{X}^*_{Ti}(\boldsymbol{X}^*_{T(i+1)})' \\ \sum_{i=1}^{T-1} \boldsymbol{X}^*_{T(i+1)}(\boldsymbol{X}^*_{Ti})' & \sum_{i=1}^{T} \boldsymbol{X}^*_{Ti}(\boldsymbol{X}^*_{Ti})' \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{S}^{(b)}_{11} & \boldsymbol{S}^{(b)}_{12} \\ \boldsymbol{S}^{(b)}_{21} & \boldsymbol{S}^{(b)}_{22} \end{pmatrix}$$

and

$$\boldsymbol{S} \equiv \frac{1}{T}\begin{pmatrix} \sum_{i=1}^{T} \boldsymbol{X}_i\boldsymbol{X}_i' & \sum_{i=1}^{T-1} \boldsymbol{X}_i\boldsymbol{X}_{i+1}' \\ \sum_{i=1}^{T-1} \boldsymbol{X}_{i+1}\boldsymbol{X}_i' & \sum_{i=1}^{T} \boldsymbol{X}_i\boldsymbol{X}_i' \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}.$$

Note that the first one is for the bootstrap sample while the second one is for the original sample. It is easy to see that the matrix $\boldsymbol{S}$ differs from the matrix $\boldsymbol{N}$ (defined in Section 2.2) only in minor terms which appear in the diagonal blocks and the minor terms are negligible when T tends to infinity. Let us also denote the sample canonical correlations calculated from the bootstrap sample $\{\boldsymbol{X}^*_{Ti}\}_{i=1}^{T}$ by $\hat{r_i}^{(b)}$, $i = 1, 2, ..., p$, and the matrix $T^{1/2}(\boldsymbol{S}^{(b)} - \mathbb{E}(\boldsymbol{S}^{(b)}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T))$ by $\boldsymbol{W}^{(b)}$ with blocks $\boldsymbol{W}^{(b)}_{11}$, $\boldsymbol{W}^{(b)}_{12}$, $\boldsymbol{W}^{(b)}_{21}$, $\boldsymbol{W}^{(b)}_{22}$ similarly partitioned as $\boldsymbol{S}^{(b)}$. We want to establish the asymptotic relationship between $T^{1/2}(\boldsymbol{S}^{(b)} - \mathbb{E}(\boldsymbol{S}^{(b)}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T))$ and $T^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma}_y)$.

Before stating such a result, we present some necessary probability results as propositions.

Proposition 4. *Let $a_1, a_2, a_3, \ldots, a_n, \ldots$ be an infinite sequence of real numbers which satisfies the condition $\lim_{n\to\infty} A_n < \infty$ where $A_n \equiv \frac{\sum_{i=1}^{n} a_i^4}{n}$. Construct an infinite sequence of random variables $\{X_i\}_{i=1}^{\infty}$ based on $\{a_i\}_{i=1}^{\infty}$ as follows(no need to impose independence here):*

$$X_i \sim \boldsymbol{uniform}(a_1, a_2, \ldots, a_i)$$

for every $i \in \mathbb{Z}^+$. Then we have for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{E}\{X_n^4 I(X_n^2 \geq \epsilon\sqrt{n})\} \to 0. \tag{2.13}$$

The next proposition can be viewed as a centralized version of Proposition 4.

Proposition 5. *Let $a_1, a_2, a_3, \ldots, a_n, \ldots$ be an infinite sequence of real numbers which satisfies the following three conditions:*

1. $\lim_{n\to\infty} A_n < \infty$ *where* $A_n \equiv \frac{\sum_{i=1}^n a_i^4}{n}$,
2. $\bar{a}_n \equiv \frac{\sum_{i=1}^n a_i}{n}$ *is bounded for all* n,
3. $\frac{\sum_{i=1}^n (a_i - \bar{a}_n)^2}{n}$ *is bounded for all* n.

Construct an infinite sequence of random variables $\{X_i\}_{i=1}^\infty$ based on $\{a_i\}_{i=1}^\infty$ as follows(no need to impose independence here):

$$X_i \sim \boldsymbol{uniform}(a_1 - \bar{a}_i, a_2 - \bar{a}_i, \ldots, a_i - \bar{a}_i)$$

for every $i \in \mathbb{Z}^+$. Then we have for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{E}\{(X_n^2 - \mathbb{E}(X_n^2))^2 I(|X_n^2 - \mathbb{E}(X_n^2)| \geq \epsilon\sqrt{n})\} \to 0. \tag{2.14}$$

Proposition 6. *Let $\{a_i\}_{i=1}^\infty$ and $\{b_i\}_{i=1}^\infty$ be two infinite sequences of real numbers which satisfies the following conditions:*

1. $\lim_{n\to\infty} A_n < \infty$ *where* $A_n \equiv \frac{\sum_{i=1}^n a_i^2 b_i^2}{n}$,
2. $\bar{a}_n \equiv \frac{\sum_{i=1}^n a_i}{n}$ *and* $\bar{b}_n \equiv \frac{\sum_{i=1}^n b_i}{n}$ *is bounded for all* n,

Construct two sequences of random variables $\{X_i\}_{i=1}^\infty$ and $\{Y_i\}_{i=1}^\infty$ based on the two number sequences $\{a_i\}_{i=1}^\infty$ and $\{b_i\}_{i=1}^\infty$ as follows(no need to impose independence here):

$$X_i \sim \boldsymbol{uniform}(a_1 - \bar{a}_i, a_2 - \bar{a}_i, \ldots, a_i - \bar{a}_i)$$

and

$$Y_i \sim \boldsymbol{uniform}(b_1 - \bar{b}_i, b_2 - \bar{b}_i, \ldots, b_i - \bar{b}_i)$$

for every $i \in \mathbb{Z}^+$. Then we have for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{E}\{(X_nY_n - \mathbb{E}X_n\mathbb{E}Y_n)^2 I(|X_nY_n - \mathbb{E}X_n\mathbb{E}Y_n| \geq \epsilon\sqrt{n})\} \to 0. \tag{2.15}$$

Proposition 7. *Let* $\{a_i\}_{i=1}^{\infty}$ *and* $\{b_i\}_{i=1}^{\infty}$ *be two infinite sequences of real numbers which satisfies the following conditions:*

1. $\lim_{n\to\infty} A_n < \infty$ *where* $A_n \equiv \frac{\sum_{i=1}^{n} a_i^2 b_i^2}{n}$,
2. $\bar{a}_n \equiv \frac{\sum_{i=1}^{n} a_i}{n}$ *and* $\bar{b}_n \equiv \frac{\sum_{i=1}^{n} b_i}{n}$ *is bounded for all* n,
3. $\frac{\sum_{i=1}^{n}(a_i-\bar{a}_n)(b_i-\bar{b}_n)}{n}$ *is bounded for all* n.

Construct a sequence of random vectors $\{(X_i, Y_i)\}_{i=1}^{\infty}$ *based on the two number sequences* $\{a_i\}_{i=1}^{\infty}$ *and* $\{b_i\}_{i=1}^{\infty}$ *as follows:*

$$(X_i, Y_i) \sim \textbf{\textit{uniform}}((a_1 - \bar{a}_i, b_1 - \bar{b}_i), (a_2 - \bar{a}_i, b_2 - \bar{b}_i), \ldots, (a_i - \bar{a}_i, b_i - \bar{b}_i))$$

for every $i \in \mathbb{Z}^+$. *Then we have for every fixed* $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{E}\{(X_nY_n - \mathbb{E}(X_nY_n))^2 I(|X_nY_n - \mathbb{E}(X_nY_n)| \geq \epsilon\sqrt{n})\} \to 0. \tag{2.16}$$

Now we are ready to present a lemma characterizing the asymptotic normality of the conditional distribution of $T^{1/2}(\boldsymbol{S}^{(b)} - \mathbb{E}(\boldsymbol{S}^{(b)}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T))$ given the sample path $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T, \ldots$ under some conditions.

Lemma 5. *The conditional distribution of* $T^{1/2}(\boldsymbol{S}^{(b)} - \mathbb{E}(\boldsymbol{S}^{(b)}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T))$ *given the infinite sample path* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T, \ldots$ *converges in distribution to multivariate normal distribution as the bootstrap sample size* T *tends to infinity almost surely. And the covariance structure of this multivariate normal is the same as the covariance structure of* $T^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma}_y)$.

Now we can state the main result for the non-normal case.

Theorem 3. *Assuming* $\{\boldsymbol{X}_i\}_{i=1}^{T}$ *to be a series of i.i.d random vector with finite fourth moment* $\mathbb{E}\|\boldsymbol{X}_i\|^4 < \infty$ *and* $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$. *Then along almost all the sample sequences* $\omega = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T, \ldots)$, *as* T *tends to* ∞, *we have the difference between* $\mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \leq c|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T)$ *and* $\mathbb{P}(\sqrt{T}\hat{r_1} \leq c)$ *converges uniformly to 0, or equivalently we have*

$$\sup_{-\infty<c<\infty} |\mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \leq c|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T) - \mathbb{P}(\sqrt{T}\hat{r_1} \leq c)| \xrightarrow{a.s.} 0.$$

Remark. *If should be pointed out that when there are multiple roots in the covariance structure, the bootstrap consistency might not hold, see for example Section 5 of Eaton and Tyler (1991). But under our special problem setup, it can be achieved.*

2.5 Testing Procedures

Now we elaborate on how our test is performed base on the asymptotic results obtained in previous sections. Since the asymptotic distribution for the sample canonical correlations under the lag-1 case (2.4) can be obtained explicitly when the observations follow normal or elliptical distribution, when the generating scheme of data is normal or elliptical, one might adopt the largest canonical correlation calculated from the lag-1 sample as test statistics and use the limiting distribution obtained in Theorem 1 and 2 to compute the rejection threshold. But such distribution assumptions might be too strong for real data. So we propose the following testing procedure based on the bootstrap sample results in section 2.4.

Step 1. Given the original sample observations $\{\boldsymbol{X}_i\}_{i=1}^T$ where $\boldsymbol{X}_i \in \mathbb{R}^p$ for $i = 1, \ldots, T$, calculate the centralized series $\boldsymbol{X}_i^c \equiv \boldsymbol{X}_i - \bar{\boldsymbol{X}}_i$. Construct the lag-1 sample based on the centralized series $\{\boldsymbol{Y}_i = (\boldsymbol{X}_i^{(c)}, \boldsymbol{X}_{i+1}^{(c)})' \in \mathbb{R}^{2p}\}_{i=1}^{T-1}$.

Step 2. Calculate the sample covariance matrix of the lag-1 sample $\{\boldsymbol{Y}_i\}$. Based on this, calculate the squared largest sample canonical correlation $\hat{r_1}^2$.

Step 3. Let $b = 1$(use as iterator), $B \in \mathbb{Z}^+$ be a predetermined value. While $b \leq B$, repeat the following:

1. Draw a bootstrap sample $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_T^*$ from the from the centralized series.

2. Construct the lag-1 sample $\{\boldsymbol{Y}_i^*\}$ from $\{\boldsymbol{X}_i^*\}$. Based on $\{\boldsymbol{Y}_i^*\}$, record the value of the squared largest sample canonical correlation $\hat{r_{1*}}^2$.

3. increase b by 1.

Step 4. Order and denote the recorded B values of $\hat{r_{1*}}^2$ as follows:

$$\hat{r_{11}}^2 \leq \hat{r_{12}}^2 \leq \cdots \leq \hat{r_{1B}}^2$$

Step 5. Let α be the prespecified size and $q \equiv \lfloor (1-\alpha)B \rfloor$. Our approximate size α test based on canonical correlation rejects when $\hat{r_1}^2 \geq \hat{r_{1q}}^2$.

2.6 Simulation Results

To examine the performance of our test, we perform some simulation studies comparing the sizes and powers of our proposed test with several well-known white noise tests in literature. Other tests involved are the likelihood ratio test, the Ljung Box test and the Lagrangian multiplier test.

For type I error comparison, we consider two types of white noise: normal and t-distribution with 10 degree of freedom. Theses empirical results are shown in Table 2.1 to 2.9 for dimension $p = 5, 10, 25$, size threshold $\alpha = 0.1, 0.05, 0.01$ and sample size $T = 100, 300, 500, 1000, 2000$.

Throughout this section, we denote **CCT** to be the proposed white noise test based on canonical correlations, **LRT** to be the likelihood ratio test, **LBT1** and **LBT2** to be the Ljung Box test at lag-1 and lag-2 respectively, **LMT1** and **LMT2** to be the Lagrangian multiplier test at lag-1 and lag-2 respectively.

We also perform an experiment to compare the powers of these white noise tests when the null hypothesis is **NOT** true. 10 settings which violate the null hypothesis are considered: MA(1) with $\theta = 0.3, 0.5$, AR(1) with $\phi = 0.3, 0.4$, ARMA(1,1) with $(\theta, \phi) = (0.3, 0.3), (0.2, 0.2)$, MA(2) with $(\theta_1, \theta_2) = (0.2, 0.3), (0.3, 0.2)$ and AR(2) with $(\phi_1, \phi_2) = (0.2, 0.3), (0.3, 0.2)$.

Under each setting, the empirical rejection rate among 1000 replications are computed for various dimension p, quantile q, sample size T setups. From the

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.097	0.088	0.097	0.098	0.082	0.077
T=300	0.096	0.108	0.112	0.112	0.108	0.102
T=500	0.091	0.086	0.089	0.105	0.085	0.108
T=1000	0.098	0.096	0.097	0.093	0.096	0.093
T=2000	0.096	0.106	0.106	0.101	0.105	0.101

Table 2.1: type I error when the white noise is normal and $\alpha = 0.1$.

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.042	0.037	0.042	0.046	0.033	0.034
T=300	0.050	0.056	0.057	0.049	0.053	0.042
T=500	0.048	0.047	0.049	0.057	0.047	0.051
T=1000	0.060	0.052	0.051	0.047	0.049	0.045
T=2000	0.044	0.052	0.052	0.042	0.052	0.041

Table 2.2: type I error when the white noise is normal and $\alpha = 0.05$.

Tables 2.19 to 2.24 , we observe that our proposed test outperforms the other five tests in terms of power under most of the model setups. We also note that as the dimension p increase, the power of our test decrease in an insignificantly manner comparing to all other tests considered.

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.013	0.009	0.010	0.015	0.007	0.008
T=300	0.012	0.013	0.013	0.008	0.011	0.007
T=500	0.006	0.011	0.011	0.007	0.010	0.005
T=1000	0.018	0.011	0.010	0.010	0.010	0.011
T=2000	0.009	0.011	0.011	0.005	0.011	0.006

Table 2.3: type I error when the white noise is normal and $\alpha = 0.01$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.099	0.108	0.132	0.141	0.093	0.078
T=300	0.098	0.096	0.098	0.111	0.091	0.095
T=500	0.107	0.107	0.111	0.117	0.105	0.100
T=1000	0.088	0.098	0.100	0.099	0.096	0.092
T=2000	0.099	0.093	0.093	0.094	0.092	0.094

Table 2.4: type I error when the white noise is normal and $\alpha = 0.1$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.050	0.054	0.062	0.072	0.042	0.022
T=300	0.050	0.056	0.056	0.062	0.052	0.054
T=500	0.061	0.055	0.054	0.057	0.050	0.052
T=1000	0.045	0.043	0.045	0.046	0.043	0.047
T=2000	0.053	0.049	0.050	0.054	0.049	0.051

Table 2.5: type I error when the white noise is normal and $\alpha = 0.05$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.006	0.013	0.011	0.010	0.007	0.003
T=300	0.012	0.013	0.012	0.014	0.012	0.010
T=500	0.012	0.009	0.009	0.016	0.009	0.010
T=1000	0.009	0.010	0.009	0.010	0.009	0.009
T=2000	0.009	0.008	0.007	0.009	0.007	0.007

Table 2.6: type I error when the white noise is normal and $\alpha = 0.01$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.089	0.191	0.169	0.257	0.067	0.057
T=300	0.093	0.094	0.103	0.130	0.084	0.081
T=500	0.090	0.104	0.110	0.118	0.099	0.079
T=1000	0.106	0.100	0.101	0.108	0.097	0.095
T=2000	0.102	0.111	0.112	0.107	0.110	0.096

Table 2.7: type I error when the white noise is normal and $\alpha = 0.1$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.037	0.102	0.067	0.131	0.036	0.008
T=300	0.055	0.045	0.045	0.059	0.036	0.027
T=500	0.039	0.048	0.052	0.054	0.040	0.037
T=1000	0.054	0.045	0.047	0.051	0.045	0.040
T=2000	0.055	0.060	0.061	0.060	0.058	0.054

Table 2.8: type I error when the white noise is normal and $\alpha = 0.05$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.008	0.035	0.010	0.016	0.001	0.000
T=300	0.008	0.008	0.007	0.011	0.005	0.005
T=500	0.010	0.009	0.010	0.014	0.009	0.007
T=1000	0.013	0.005	0.005	0.007	0.004	0.004
T=2000	0.016	0.015	0.015	0.011	0.013	0.009

Table 2.9: type I error when the white noise is normal and $\alpha = 0.01$.

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.100	0.094	0.106	0.109	0.090	0.077
T=300	0.104	0.098	0.107	0.108	0.097	0.096
T=500	0.124	0.117	0.117	0.108	0.116	0.098
T=1000	0.107	0.107	0.108	0.105	0.106	0.098
T=2000	0.095	0.088	0.089	0.099	0.088	0.098

Table 2.10: type I error when the white noise is t(10) and $\alpha = 0.1$.

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.046	0.040	0.043	0.053	0.032	0.031
T=300	0.052	0.058	0.061	0.050	0.058	0.041
T=500	0.058	0.072	0.073	0.058	0.071	0.056
T=1000	0.058	0.056	0.057	0.047	0.056	0.052
T=2000	0.051	0.042	0.042	0.054	0.042	0.054

Table 2.11: type I error when the white noise is t(10) and $\alpha = 0.05$.

p=5	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.009	0.010	0.007	0.010	0.006	0.007
T=300	0.008	0.007	0.008	0.008	0.007	0.008
T=500	0.021	0.011	0.011	0.016	0.010	0.015
T=1000	0.011	0.007	0.006	0.015	0.005	0.012
T=2000	0.008	0.008	0.008	0.011	0.008	0.012

Table 2.12: type I error when the white noise is t(10) and $\alpha = 0.01$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.096	0.108	0.119	0.132	0.098	0.075
T=300	0.096	0.097	0.097	0.120	0.090	0.109
T=500	0.110	0.105	0.106	0.106	0.101	0.102
T=1000	0.098	0.102	0.104	0.116	0.099	0.103
T=2000	0.090	0.091	0.092	0.091	0.092	0.090

Table 2.13: type I error when the white noise is t(10) and $\alpha = 0.1$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.060	0.049	0.060	0.069	0.039	0.037
T=300	0.049	0.048	0.051	0.061	0.046	0.051
T=500	0.057	0.054	0.054	0.061	0.051	0.054
T=1000	0.053	0.056	0.055	0.052	0.053	0.045
T=2000	0.046	0.048	0.048	0.045	0.047	0.044

Table 2.14: type I error when the white noise is t(10) and $\alpha = 0.05$.

p=10	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.009	0.011	0.010	0.014	0.008	0.000
T=300	0.013	0.008	0.008	0.003	0.008	0.004
T=500	0.013	0.008	0.008	0.011	0.006	0.009
T=1000	0.011	0.011	0.011	0.006	0.010	0.007
T=2000	0.006	0.006	0.006	0.002	0.006	0.002

Table 2.15: type I error when the white noise is t(10) and $\alpha = 0.01$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.085	0.180	0.155	0.210	0.073	0.028
T=300	0.104	0.115	0.118	0.136	0.095	0.086
T=500	0.103	0.107	0.117	0.113	0.100	0.090
T=1000	0.094	0.092	0.096	0.098	0.092	0.085
T=2000	0.106	0.110	0.113	0.102	0.110	0.097

Table 2.16: type I error when the white noise is t(10) and $\alpha = 0.1$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.045	0.113	0.075	0.118	0.020	0.006
T=300	0.054	0.058	0.058	0.066	0.041	0.035
T=500	0.050	0.048	0.051	0.056	0.043	0.035
T=1000	0.053	0.045	0.047	0.046	0.046	0.041
T=2000	0.048	0.060	0.061	0.051	0.059	0.052

Table 2.17: type I error when the white noise is t(10) and $\alpha = 0.05$.

p=25	CCT	LRT	LBT1	LBT2	LMT1	LMT2
T=100	0.004	0.028	0.008	0.020	0.004	0.001
T=300	0.008	0.008	0.009	0.012	0.007	0.002
T=500	0.012	0.015	0.013	0.011	0.011	0.006
T=1000	0.010	0.011	0.011	0.008	0.010	0.009
T=2000	0.007	0.008	0.007	0.008	0.007	0.006

Table 2.18: type I error when the white noise is t(10) and $\alpha = 0.01$.

Table 2.19: empirical rejection rates when quantile q=0.90 and the white noise is normal

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.827	0.777	0.778	0.618	0.764	0.647	0.542	0.470	0.469	0.383	0.451	0.351	0.191	0.215	0.217	0.226	0.181	0.156
T=500	0.996	0.990	0.991	0.947	0.990	0.966	0.944	0.861	0.856	0.696	0.850	0.720	0.506	0.421	0.422	0.316	0.395	0.287
T=750	1.000	1.000	1.000	0.999	1.000	1.000	0.999	0.978	0.976	0.893	0.976	0.911	0.820	0.575	0.569	0.420	0.550	0.413
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.994	0.968	0.994	0.982	0.963	0.779	0.763	0.553	0.753	0.565
MA(1) theta=0.5																		
T=250	0.997	0.996	0.995	0.974	0.995	0.992	0.969	0.884	0.869	0.705	0.856	0.789	0.499	0.398	0.388	0.346	0.337	0.279
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	1.000	0.998	0.990	0.836	0.813	0.636	0.794	0.664
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.974	0.965	0.812	0.961	0.886
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.960	0.998	0.988
AR(1) phi=0.3																		
T=250	0.910	0.867	0.866	0.750	0.861	0.705	0.658	0.573	0.574	0.466	0.544	0.388	0.206	0.260	0.273	0.235	0.217	0.145
T=500	1.000	0.996	0.996	0.989	0.996	0.985	0.983	0.928	0.926	0.820	0.921	0.762	0.630	0.489	0.484	0.378	0.456	0.296
T=750	1.000	1.000	1.000	0.998	1.000	0.998	1.000	0.988	0.986	0.965	0.986	0.948	0.922	0.708	0.695	0.540	0.672	0.453
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	1.000	0.993	0.990	0.881	0.864	0.736	0.852	0.641
AR(1) phi=0.4																		
T=250	0.999	0.994	0.994	0.966	0.994	0.959	0.961	0.873	0.864	0.769	0.848	0.675	0.518	0.446	0.432	0.395	0.371	0.216
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.996	0.988	0.996	0.978	0.977	0.812	0.782	0.688	0.746	0.532
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.968	0.953	0.885	0.949	0.800
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.995	0.977	0.995	0.951
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.964	0.997	0.963	0.941	0.729	0.681	0.545	0.611	0.393
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.984	0.923	0.982	0.890
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	1.000	0.987
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.19: empirical rejection rates when quantile q=0.90 and the white noise is normal

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.993	0.981	0.982	0.929	0.980	0.936	0.904	0.805	0.794	0.656	0.777	0.610	0.366	0.371	0.379	0.314	0.313	0.198
T=500	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.998	0.959	0.998	0.961	0.934	0.732	0.710	0.552	0.683	0.487
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	0.999	0.925	0.915	0.757	0.906	0.722
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.987	0.923	0.987	0.917
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.633	0.589	0.587	0.841	0.579	0.753	0.373	0.352	0.361	0.576	0.337	0.431	0.168	0.196	0.199	0.302	0.169	0.183
T=500	0.919	0.890	0.891	0.995	0.890	0.992	0.718	0.634	0.630	0.899	0.621	0.809	0.301	0.309	0.320	0.487	0.289	0.344
T=750	0.985	0.978	0.978	1.000	0.977	1.000	0.938	0.848	0.847	0.993	0.843	0.979	0.535	0.409	0.413	0.664	0.391	0.517
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	0.985	0.955	0.955	0.998	0.954	0.995	0.758	0.551	0.540	0.831	0.529	0.705
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.921	0.888	0.888	0.865	0.880	0.794	0.736	0.639	0.637	0.609	0.621	0.472	0.275	0.289	0.293	0.317	0.235	0.187
T=500	1.000	0.997	0.997	0.994	0.997	0.991	0.986	0.951	0.947	0.921	0.943	0.850	0.747	0.569	0.565	0.516	0.543	0.377
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	0.982	0.959	0.765	0.747	0.706	0.732	0.540
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.998	0.916	0.904	0.867	0.900	0.752
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.786	0.750	0.754	0.975	0.744	0.949	0.559	0.499	0.494	0.785	0.474	0.640	0.222	0.224	0.235	0.432	0.198	0.222
T=500	0.979	0.973	0.973	1.000	0.973	1.000	0.900	0.820	0.818	0.990	0.810	0.962	0.551	0.452	0.452	0.735	0.423	0.554
T=750	0.998	0.998	0.998	1.000	0.998	1.000	0.987	0.959	0.958	1.000	0.958	1.000	0.813	0.612	0.605	0.930	0.592	0.815
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.993	1.000	0.993	1.000	0.950	0.809	0.795	0.991	0.782	0.955
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.977	0.960	0.959	0.978	0.956	0.955	0.864	0.767	0.762	0.854	0.751	0.713	0.407	0.364	0.359	0.447	0.297	0.256
T=500	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.988	0.988	0.995	0.988	0.986	0.900	0.712	0.689	0.806	0.657	0.558
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.932	0.923	0.968	0.921	0.852
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.983	0.979	0.996	0.979	0.960

Table 2.20: empirical rejection rates when quantile q=0.95 and the white noise is normal

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.742	0.666	0.662	0.481	0.657	0.512	0.430	0.346	0.345	0.246	0.320	0.231	0.104	0.126	0.127	0.130	0.098	0.079
T=500	0.994	0.979	0.978	0.908	0.977	0.924	0.904	0.764	0.752	0.555	0.740	0.583	0.373	0.289	0.284	0.202	0.257	0.165
T=750	1.000	1.000	1.000	0.997	1.000	0.998	0.995	0.951	0.949	0.812	0.949	0.852	0.739	0.435	0.423	0.277	0.406	0.276
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.991	0.942	0.991	0.959	0.934	0.650	0.636	0.423	0.625	0.423
MA(1) theta=0.5																		
T=250	0.995	0.991	0.990	0.940	0.988	0.975	0.931	0.797	0.774	0.553	0.762	0.641	0.364	0.268	0.252	0.215	0.199	0.152
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.969	0.998	0.991	0.973	0.746	0.693	0.480	0.659	0.515
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.952	0.927	0.702	0.918	0.788
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.992	0.924	0.992	0.970
AR(1) phi=0.3																		
T=250	0.856	0.784	0.779	0.619	0.765	0.568	0.544	0.437	0.420	0.336	0.389	0.240	0.119	0.157	0.148	0.129	0.104	0.072
T=500	0.997	0.992	0.992	0.972	0.992	0.963	0.962	0.873	0.860	0.699	0.856	0.621	0.509	0.345	0.337	0.246	0.305	0.178
T=750	1.000	1.000	1.000	0.997	1.000	0.996	0.998	0.977	0.974	0.922	0.974	0.896	0.876	0.572	0.540	0.406	0.522	0.314
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	1.000	0.982	0.983	0.780	0.750	0.595	0.737	0.500
AR(1) phi=0.4																		
T=250	0.997	0.988	0.984	0.937	0.981	0.909	0.922	0.786	0.764	0.648	0.747	0.509	0.380	0.312	0.280	0.249	0.230	0.115
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.993	0.977	0.993	0.952	0.958	0.681	0.643	0.528	0.611	0.373
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	0.998	1.000	0.934	0.902	0.804	0.891	0.661
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.995	0.960	0.995	0.900
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.992	0.932	0.987	0.913	0.889	0.597	0.519	0.393	0.453	0.235
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986	0.968	0.850	0.961	0.795
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.981	1.000	0.976
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.20: empirical rejection rates when quantile q=0.95 and the white noise is normal

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.984	0.967	0.963	0.868	0.962	0.873	0.860	0.705	0.681	0.511	0.648	0.474	0.258	0.247	0.229	0.186	0.186	0.102
T=500	1.000	1.000	1.000	0.999	1.000	0.999	0.999	0.994	0.994	0.928	0.994	0.933	0.886	0.600	0.563	0.399	0.529	0.333
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	0.997	0.999	0.877	0.851	0.634	0.832	0.598
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.980	0.973	0.851	0.969	0.831
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.514	0.456	0.454	0.757	0.443	0.642	0.272	0.241	0.244	0.430	0.224	0.279	0.103	0.120	0.122	0.194	0.085	0.085
T=500	0.875	0.833	0.830	0.990	0.824	0.973	0.624	0.494	0.485	0.832	0.475	0.705	0.204	0.198	0.198	0.348	0.183	0.235
T=750	0.974	0.963	0.963	1.000	0.963	1.000	0.896	0.773	0.768	0.983	0.763	0.950	0.412	0.267	0.258	0.507	0.236	0.348
T=1000	0.999	0.997	0.997	1.000	0.997	1.000	0.972	0.921	0.919	0.998	0.914	0.992	0.669	0.421	0.415	0.740	0.407	0.578
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.859	0.824	0.823	0.802	0.817	0.699	0.635	0.511	0.497	0.460	0.471	0.316	0.176	0.178	0.178	0.201	0.146	0.094
T=500	0.999	0.997	0.996	0.991	0.996	0.977	0.975	0.900	0.894	0.867	0.890	0.729	0.622	0.426	0.405	0.376	0.375	0.243
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.995	0.990	0.995	0.966	0.925	0.653	0.635	0.568	0.616	0.398
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.999	0.996	0.996	0.843	0.821	0.787	0.814	0.624
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.699	0.646	0.639	0.949	0.633	0.904	0.458	0.355	0.349	0.690	0.318	0.495	0.134	0.142	0.148	0.278	0.103	0.114
T=500	0.971	0.957	0.957	1.000	0.956	0.999	0.852	0.729	0.720	0.977	0.713	0.930	0.452	0.339	0.325	0.604	0.299	0.399
T=750	0.997	0.997	0.997	1.000	0.997	1.000	0.981	0.934	0.927	1.000	0.922	1.000	0.751	0.501	0.480	0.878	0.463	0.690
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.987	0.985	1.000	0.984	1.000	0.922	0.698	0.679	0.980	0.670	0.902
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.960	0.922	0.917	0.959	0.915	0.902	0.801	0.680	0.675	0.786	0.660	0.558	0.286	0.236	0.219	0.327	0.177	0.120
T=500	1.000	0.998	0.997	1.000	0.997	0.999	0.996	0.976	0.969	0.992	0.967	0.967	0.857	0.584	0.550	0.695	0.523	0.425
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.885	0.863	0.935	0.845	0.729
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.966	0.956	0.990	0.954	0.919

Table 2.21: empirical rejection rates when quantile q=0.99 and the white noise is normal

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.528	0.427	0.415	0.249	0.397	0.245	0.174	0.143	0.131	0.067	0.120	0.055	0.030	0.030	0.025	0.027	0.014	0.009
T=500	0.967	0.914	0.909	0.755	0.907	0.808	0.786	0.544	0.515	0.296	0.502	0.320	0.189	0.119	0.103	0.059	0.085	0.041
T=750	1.000	1.000	1.000	0.983	1.000	0.992	0.971	0.859	0.842	0.591	0.838	0.649	0.527	0.220	0.203	0.115	0.192	0.105
T=1000	1.000	0.999	0.999	0.997	0.999	0.999	1.000	0.978	0.972	0.843	0.971	0.887	0.836	0.387	0.355	0.168	0.339	0.178
MA(1) theta=0.5																		
T=250	0.984	0.946	0.938	0.792	0.933	0.907	0.823	0.582	0.522	0.311	0.492	0.355	0.160	0.110	0.077	0.065	0.055	0.033
T=500	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.985	0.893	0.984	0.958	0.922	0.485	0.406	0.216	0.378	0.233
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	0.814	0.755	0.454	0.735	0.562
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.976	0.959	0.736	0.957	0.873
AR(1) phi=0.3																		
T=250	0.664	0.573	0.547	0.377	0.534	0.309	0.318	0.206	0.183	0.131	0.160	0.081	0.042	0.030	0.024	0.033	0.016	0.015
T=500	0.984	0.976	0.975	0.894	0.975	0.857	0.896	0.674	0.645	0.426	0.632	0.348	0.284	0.125	0.108	0.078	0.102	0.046
T=750	1.000	0.998	0.998	0.986	0.998	0.981	0.993	0.931	0.920	0.797	0.918	0.716	0.744	0.303	0.275	0.190	0.261	0.118
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.995	0.948	0.994	0.925	0.958	0.545	0.506	0.319	0.498	0.235
AR(1) phi=0.4																		
T=250	0.979	0.933	0.922	0.859	0.920	0.763	0.807	0.579	0.522	0.388	0.488	0.247	0.186	0.105	0.081	0.080	0.064	0.020
T=500	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.983	0.977	0.926	0.975	0.848	0.892	0.443	0.380	0.274	0.352	0.149
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.993	0.999	0.988	0.999	0.805	0.756	0.562	0.736	0.383
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.981	0.958	0.848	0.955	0.705
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	0.999	0.999	0.994	0.999	0.998	0.998	0.956	0.929	0.766	0.922	0.705	0.717	0.323	0.213	0.147	0.162	0.062
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.938	0.859	0.629	0.843	0.516
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.928	0.996	0.907
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	0.996

Table 2.21: empirical rejection rates when quantile q=0.99 and the white noise is normal

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.950	0.886	0.874	0.675	0.864	0.663	0.710	0.464	0.404	0.257	0.382	0.192	0.107	0.069	0.053	0.047	0.034	0.025
T=500	1.000	1.000	1.000	0.997	1.000	0.998	0.998	0.968	0.956	0.783	0.952	0.775	0.739	0.342	0.291	0.157	0.256	0.103
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.977	1.000	0.980	0.996	0.691	0.619	0.379	0.601	0.312
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.921	0.894	0.645	0.883	0.598
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.300	0.235	0.230	0.561	0.219	0.374	0.117	0.097	0.084	0.193	0.072	0.083	0.026	0.025	0.017	0.051	0.013	0.016
T=500	0.739	0.660	0.653	0.960	0.649	0.907	0.427	0.288	0.271	0.643	0.256	0.417	0.078	0.062	0.062	0.151	0.047	0.070
T=750	0.933	0.900	0.896	1.000	0.895	0.998	0.773	0.564	0.554	0.933	0.544	0.838	0.230	0.091	0.082	0.264	0.074	0.136
T=1000	0.992	0.983	0.983	1.000	0.981	1.000	0.931	0.762	0.745	0.991	0.742	0.975	0.476	0.204	0.199	0.493	0.189	0.293
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.734	0.655	0.631	0.592	0.618	0.423	0.399	0.253	0.229	0.211	0.207	0.097	0.069	0.050	0.040	0.056	0.027	0.017
T=500	0.993	0.980	0.976	0.971	0.976	0.939	0.932	0.756	0.730	0.685	0.720	0.490	0.413	0.193	0.176	0.170	0.156	0.073
T=750	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.982	0.977	0.961	0.974	0.879	0.805	0.393	0.343	0.299	0.321	0.161
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.997	0.995	0.997	0.987	0.987	0.657	0.603	0.563	0.589	0.348
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.542	0.474	0.454	0.866	0.449	0.750	0.261	0.183	0.161	0.449	0.149	0.255	0.030	0.030	0.021	0.081	0.013	0.024
T=500	0.921	0.883	0.879	0.999	0.877	0.998	0.732	0.536	0.510	0.934	0.498	0.815	0.243	0.128	0.110	0.347	0.100	0.161
T=750	0.995	0.988	0.987	1.000	0.987	1.000	0.952	0.846	0.836	0.999	0.830	0.990	0.586	0.268	0.255	0.670	0.243	0.411
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.956	0.948	1.000	0.948	1.000	0.831	0.460	0.426	0.923	0.410	0.746
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.897	0.817	0.798	0.898	0.790	0.776	0.631	0.448	0.401	0.576	0.376	0.290	0.131	0.082	0.062	0.114	0.046	0.021
T=500	0.999	0.995	0.995	0.998	0.995	0.996	0.985	0.931	0.910	0.974	0.905	0.888	0.728	0.361	0.303	0.431	0.273	0.187
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.997	1.000	0.996	0.997	0.984	0.699	0.645	0.795	0.618	0.480
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.904	0.881	0.951	0.876	0.782

Table 2.22: empirical rejection rates when quantile q=0.90 and the white noise is t(10)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.827	0.785	0.783	0.609	0.772	0.646	0.536	0.456	0.453	0.349	0.428	0.324	0.192	0.211	0.230	0.231	0.172	0.153
T=500	0.997	0.993	0.993	0.972	0.993	0.978	0.943	0.866	0.862	0.689	0.851	0.714	0.501	0.394	0.398	0.331	0.369	0.295
T=750	1.000	1.000	1.000	0.999	1.000	1.000	0.997	0.982	0.981	0.887	0.980	0.921	0.802	0.594	0.584	0.441	0.563	0.440
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.996	0.978	0.996	0.986	0.969	0.764	0.755	0.572	0.744	0.585
MA(1) theta=0.5																		
T=250	0.999	0.998	0.997	0.979	0.996	0.995	0.961	0.885	0.876	0.714	0.857	0.777	0.480	0.406	0.399	0.351	0.346	0.275
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.989	1.000	0.998	0.985	0.813	0.784	0.611	0.756	0.644
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.976	0.966	0.851	0.961	0.909
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.996	0.958	0.996	0.985
AR(1) phi=0.3																		
T=250	0.911	0.859	0.858	0.753	0.850	0.707	0.641	0.561	0.552	0.433	0.522	0.346	0.235	0.245	0.256	0.257	0.210	0.148
T=500	1.000	1.000	1.000	0.986	1.000	0.987	0.977	0.926	0.919	0.809	0.915	0.755	0.614	0.475	0.474	0.383	0.447	0.301
T=750	1.000	1.000	1.000	0.998	1.000	0.998	0.999	0.993	0.992	0.952	0.992	0.930	0.925	0.689	0.675	0.544	0.662	0.461
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	1.000	0.995	0.996	0.856	0.845	0.709	0.841	0.643
AR(1) phi=0.4																		
T=250	0.997	0.993	0.992	0.967	0.991	0.960	0.956	0.872	0.859	0.752	0.851	0.651	0.486	0.393	0.381	0.340	0.321	0.209
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.990	0.998	0.975	0.972	0.809	0.792	0.679	0.771	0.552
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.964	0.957	0.889	0.953	0.799
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.997	0.982	0.996	0.950
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.951	0.997	0.950	0.910	0.679	0.631	0.527	0.581	0.370
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	0.980	0.912	0.978	0.867
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	1.000	0.992
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.22: empirical rejection rates when quantile q=0.90 and the white noise is t(10)

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.992	0.980	0.979	0.925	0.977	0.925	0.919	0.807	0.796	0.635	0.779	0.585	0.403	0.362	0.364	0.314	0.306	0.212
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.993	0.968	0.993	0.969	0.936	0.719	0.695	0.542	0.655	0.471
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.931	0.918	0.746	0.908	0.713
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.988	0.983	0.910	0.982	0.904
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.603	0.591	0.596	0.830	0.587	0.747	0.351	0.346	0.349	0.533	0.324	0.389	0.149	0.179	0.193	0.299	0.156	0.177
T=500	0.920	0.891	0.891	0.998	0.889	0.992	0.717	0.638	0.639	0.897	0.628	0.789	0.306	0.283	0.291	0.483	0.253	0.335
T=750	0.993	0.989	0.989	1.000	0.989	1.000	0.923	0.842	0.837	0.990	0.835	0.977	0.496	0.414	0.407	0.662	0.392	0.503
T=1000	0.999	0.996	0.996	1.000	0.996	1.000	0.985	0.943	0.939	1.000	0.936	0.998	0.731	0.534	0.533	0.832	0.519	0.701
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.932	0.902	0.899	0.854	0.897	0.790	0.714	0.591	0.593	0.558	0.565	0.425	0.253	0.258	0.263	0.308	0.218	0.189
T=500	1.000	0.998	0.998	1.000	0.998	0.995	0.989	0.950	0.944	0.920	0.944	0.837	0.736	0.530	0.518	0.508	0.483	0.366
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.994	0.997	0.978	0.955	0.751	0.732	0.705	0.715	0.548
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.923	0.910	0.862	0.902	0.750
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.795	0.752	0.746	0.957	0.739	0.932	0.584	0.514	0.513	0.806	0.495	0.654	0.221	0.215	0.222	0.415	0.179	0.224
T=500	0.986	0.979	0.979	1.000	0.979	1.000	0.903	0.838	0.828	0.991	0.821	0.970	0.549	0.444	0.446	0.731	0.416	0.548
T=750	0.999	0.999	0.999	1.000	0.999	1.000	0.991	0.970	0.969	1.000	0.968	0.999	0.846	0.654	0.646	0.929	0.629	0.802
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.992	0.992	1.000	0.992	1.000	0.951	0.779	0.767	0.980	0.760	0.932
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.978	0.958	0.959	0.986	0.957	0.962	0.837	0.761	0.755	0.833	0.740	0.663	0.401	0.356	0.352	0.446	0.293	0.224
T=500	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.994	0.992	0.999	0.992	0.983	0.926	0.714	0.697	0.803	0.673	0.584
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.921	0.907	0.960	0.902	0.840
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.987	0.983	0.996	0.981	0.962

Table 2.23: empirical rejection rates when quantile q=0.95 and the white noise is t(10)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.726	0.678	0.673	0.482	0.657	0.505	0.411	0.324	0.318	0.234	0.287	0.203	0.107	0.122	0.127	0.135	0.098	0.072
T=500	0.992	0.982	0.980	0.930	0.976	0.948	0.904	0.763	0.752	0.542	0.742	0.581	0.379	0.271	0.265	0.216	0.240	0.177
T=750	1.000	1.000	1.000	0.997	1.000	0.999	0.996	0.960	0.956	0.806	0.951	0.843	0.721	0.455	0.439	0.308	0.417	0.296
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.994	0.950	0.994	0.967	0.948	0.654	0.633	0.439	0.619	0.440
MA(1) theta=0.5																		
T=250	0.999	0.992	0.991	0.944	0.991	0.980	0.926	0.797	0.777	0.568	0.755	0.636	0.349	0.270	0.265	0.219	0.196	0.150
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997	0.962	0.997	0.990	0.969	0.704	0.657	0.468	0.628	0.502
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.950	0.929	0.751	0.917	0.837
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.992	0.907	0.990	0.968
AR(1) phi=0.3																		
T=250	0.836	0.785	0.781	0.635	0.774	0.578	0.534	0.405	0.404	0.297	0.366	0.232	0.149	0.157	0.149	0.142	0.108	0.071
T=500	0.999	0.994	0.993	0.969	0.993	0.970	0.958	0.860	0.851	0.708	0.843	0.629	0.503	0.336	0.325	0.253	0.302	0.184
T=750	1.000	0.999	0.999	0.997	0.999	0.998	0.999	0.981	0.976	0.915	0.976	0.885	0.865	0.560	0.543	0.394	0.521	0.315
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.990	0.998	0.983	0.985	0.768	0.740	0.588	0.723	0.496
AR(1) phi=0.4																		
T=250	0.990	0.984	0.983	0.948	0.980	0.919	0.930	0.796	0.773	0.649	0.753	0.508	0.356	0.262	0.241	0.220	0.193	0.109
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.995	0.972	0.995	0.955	0.953	0.727	0.678	0.537	0.651	0.404
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.931	0.913	0.812	0.905	0.665
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.989	0.957	0.988	0.888
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.990	0.905	0.986	0.895	0.844	0.552	0.488	0.375	0.421	0.225
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.979	0.949	0.832	0.936	0.778
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.984	1.000	0.978
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.23: empirical rejection rates when quantile q=0.95 and the white noise is t(10)

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.983	0.961	0.958	0.862	0.958	0.870	0.861	0.702	0.674	0.466	0.649	0.420	0.284	0.231	0.223	0.200	0.179	0.104
T=500	1.000	1.000	1.000	0.998	1.000	1.000	1.000	0.987	0.985	0.925	0.985	0.928	0.890	0.587	0.549	0.391	0.522	0.326
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	0.997	0.999	0.878	0.849	0.623	0.834	0.596
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.972	0.964	0.830	0.959	0.821
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.515	0.464	0.460	0.742	0.447	0.633	0.225	0.220	0.219	0.388	0.204	0.241	0.083	0.105	0.102	0.189	0.086	0.096
T=500	0.858	0.815	0.816	0.993	0.814	0.982	0.626	0.516	0.510	0.821	0.499	0.681	0.201	0.167	0.166	0.340	0.147	0.225
T=750	0.989	0.974	0.974	1.000	0.974	1.000	0.876	0.747	0.741	0.978	0.736	0.939	0.384	0.303	0.297	0.531	0.282	0.375
T=1000	0.997	0.993	0.993	1.000	0.993	1.000	0.972	0.899	0.894	1.000	0.893	0.989	0.634	0.410	0.405	0.735	0.390	0.547
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.896	0.842	0.835	0.788	0.827	0.685	0.602	0.467	0.448	0.424	0.429	0.279	0.146	0.172	0.171	0.201	0.133	0.097
T=500	0.999	0.998	0.998	0.995	0.997	0.989	0.974	0.910	0.904	0.860	0.900	0.737	0.612	0.393	0.374	0.370	0.348	0.240
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.989	0.986	0.982	0.985	0.959	0.922	0.638	0.624	0.579	0.609	0.405
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.992	0.852	0.833	0.779	0.822	0.604
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.728	0.662	0.655	0.931	0.648	0.876	0.471	0.393	0.379	0.712	0.358	0.510	0.124	0.144	0.143	0.272	0.117	0.116
T=500	0.974	0.960	0.961	1.000	0.960	1.000	0.861	0.755	0.748	0.981	0.738	0.943	0.421	0.327	0.311	0.605	0.291	0.393
T=750	0.999	0.997	0.996	1.000	0.996	1.000	0.982	0.943	0.941	0.999	0.940	0.999	0.781	0.516	0.496	0.870	0.475	0.693
T=1000	1.000	0.999	0.999	1.000	0.999	1.000	0.995	0.989	0.989	1.000	0.989	1.000	0.922	0.684	0.666	0.961	0.655	0.888
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.956	0.933	0.929	0.970	0.930	0.917	0.779	0.655	0.631	0.740	0.615	0.522	0.308	0.233	0.218	0.313	0.178	0.106
T=500	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.980	0.977	0.994	0.974	0.961	0.884	0.590	0.546	0.697	0.526	0.437
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.987	0.868	0.846	0.923	0.827	0.712
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.971	0.953	0.990	0.947	0.922

Table 2.24: empirical rejection rates when quantile q=0.99 and the white noise is t(10)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
MA(1) theta=0.3																		
T=250	0.516	0.400	0.379	0.228	0.368	0.248	0.206	0.138	0.125	0.057	0.113	0.044	0.031	0.040	0.031	0.028	0.019	0.008
T=500	0.967	0.933	0.929	0.761	0.927	0.823	0.781	0.535	0.505	0.296	0.494	0.303	0.158	0.103	0.091	0.070	0.082	0.052
T=750	0.999	0.998	0.998	0.977	0.998	0.986	0.979	0.847	0.826	0.596	0.819	0.649	0.524	0.232	0.211	0.109	0.206	0.092
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.976	0.970	0.844	0.968	0.891	0.844	0.409	0.389	0.209	0.374	0.201
MA(1) theta=0.5																		
T=250	0.989	0.957	0.949	0.798	0.946	0.916	0.806	0.572	0.501	0.285	0.478	0.331	0.161	0.108	0.084	0.065	0.064	0.025
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	0.986	0.871	0.984	0.958	0.911	0.455	0.381	0.226	0.347	0.230
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000	0.837	0.768	0.476	0.746	0.582
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.963	0.947	0.730	0.944	0.863
AR(1) phi=0.3																		
T=250	0.657	0.571	0.557	0.394	0.537	0.304	0.312	0.194	0.177	0.122	0.156	0.072	0.045	0.047	0.035	0.042	0.027	0.018
T=500	0.991	0.974	0.970	0.896	0.970	0.864	0.888	0.679	0.655	0.442	0.648	0.350	0.271	0.151	0.136	0.094	0.123	0.051
T=750	1.000	0.998	0.998	0.990	0.998	0.986	0.995	0.939	0.925	0.780	0.924	0.722	0.732	0.300	0.271	0.188	0.252	0.139
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.990	0.949	0.990	0.927	0.947	0.546	0.505	0.317	0.490	0.241
AR(1) phi=0.4																		
T=250	0.972	0.941	0.936	0.844	0.931	0.769	0.790	0.595	0.519	0.387	0.488	0.241	0.177	0.101	0.077	0.074	0.061	0.024
T=500	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.981	0.973	0.916	0.970	0.839	0.891	0.473	0.403	0.294	0.376	0.167
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	0.993	0.997	0.808	0.750	0.561	0.738	0.380
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.972	0.952	0.838	0.949	0.693
ARMA(1,1) theta=0.3 phi=0.3																		
T=250	1.000	1.000	1.000	0.992	1.000	0.997	1.000	0.965	0.928	0.743	0.921	0.698	0.685	0.314	0.219	0.158	0.170	0.085
T=500	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.910	0.829	0.613	0.809	0.492
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.941	0.998	0.921
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	0.995

Table 2.24: empirical rejection rates when quantile q=0.99 and the white noise is t(10)

(continued)

	p=5						p=10						p=25					
	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2	CCT	LRT	LBT1	LBT2	LMT1	LMT2
ARMA(1,1) theta=0.2 phi=0.2																		
T=250	0.948	0.886	0.871	0.680	0.867	0.683	0.687	0.438	0.392	0.233	0.365	0.183	0.121	0.079	0.068	0.057	0.053	0.023
T=500	1.000	1.000	1.000	0.996	1.000	0.994	0.999	0.960	0.949	0.792	0.946	0.769	0.762	0.335	0.292	0.162	0.267	0.122
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.975	0.998	0.981	0.993	0.679	0.610	0.370	0.597	0.312
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.917	0.873	0.631	0.862	0.602
MA(2) theta1=0.2 theta2=0.3																		
T=250	0.311	0.239	0.230	0.546	0.222	0.369	0.092	0.077	0.074	0.164	0.065	0.087	0.019	0.029	0.026	0.054	0.020	0.015
T=500	0.717	0.640	0.629	0.964	0.626	0.909	0.437	0.285	0.267	0.607	0.258	0.414	0.083	0.049	0.048	0.143	0.043	0.065
T=750	0.953	0.911	0.908	1.000	0.907	1.000	0.743	0.534	0.523	0.913	0.515	0.791	0.204	0.136	0.128	0.291	0.117	0.165
T=1000	0.991	0.973	0.971	1.000	0.971	1.000	0.916	0.769	0.766	0.987	0.760	0.963	0.446	0.201	0.183	0.479	0.174	0.300
MA(2) theta1=0.3 theta2=0.2																		
T=250	0.732	0.648	0.629	0.598	0.621	0.427	0.368	0.247	0.217	0.188	0.199	0.097	0.043	0.062	0.046	0.055	0.029	0.016
T=500	0.994	0.984	0.982	0.975	0.980	0.932	0.938	0.756	0.728	0.677	0.714	0.482	0.395	0.159	0.139	0.168	0.119	0.076
T=750	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.975	0.969	0.954	0.965	0.856	0.835	0.393	0.357	0.329	0.339	0.196
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.996	0.994	0.996	0.974	0.977	0.637	0.595	0.541	0.581	0.348
AR(2) phi1=0.2 phi2=0.3																		
T=250	0.533	0.465	0.449	0.837	0.443	0.730	0.272	0.192	0.176	0.471	0.162	0.242	0.049	0.058	0.040	0.089	0.031	0.033
T=500	0.935	0.895	0.892	1.000	0.890	0.998	0.757	0.570	0.546	0.939	0.541	0.833	0.260	0.130	0.114	0.357	0.103	0.155
T=750	0.992	0.987	0.987	1.000	0.987	1.000	0.956	0.838	0.826	0.998	0.821	0.993	0.597	0.279	0.252	0.681	0.237	0.429
T=1000	0.999	0.998	0.998	1.000	0.998	1.000	0.992	0.964	0.962	0.999	0.962	0.999	0.848	0.461	0.422	0.893	0.403	0.680
AR(2) phi1=0.3 phi2=0.2																		
T=250	0.906	0.838	0.826	0.909	0.816	0.795	0.624	0.441	0.379	0.530	0.357	0.266	0.143	0.085	0.066	0.108	0.051	0.029
T=500	1.000	0.996	0.996	1.000	0.996	1.000	0.987	0.928	0.913	0.970	0.907	0.878	0.744	0.359	0.301	0.457	0.269	0.178
T=750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.998	0.999	0.997	0.994	0.975	0.686	0.632	0.775	0.619	0.468
T=1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.888	0.863	0.952	0.853	0.759

2.7 Proofs

Proof of Proposition 1. This Proposition follows from Corollary 6.3.8 of Horn et al. (1990) . □

Proof of Lemma 1. This is a classical result and its justification, for instance, can be found in Section 6.3 of Bilodeau and Brenner (2008). □

Proof of Lemma 2. Note that the entries of $\sqrt{T}\boldsymbol{N}$ are of three types:

$$\frac{1}{\sqrt{T}}\sum_{i=2}^{T} X_{(i-1)k}X_{(i-1)l},\ \frac{1}{\sqrt{T}}\sum_{i=2}^{T} X_{(i-1)k}X_{il} \text{ and } \frac{1}{\sqrt{T}}\sum_{i=2}^{T} X_{ik}X_{il}$$

where k and l are indices ranging from $[1,\ldots,p]$. So a linear combination of all the entries of $\sqrt{T}\boldsymbol{N}$ can be expressed as

$$\sum_{i=2}^{T}\frac{1}{\sqrt{T}}\{\sum_{k,l\in[1,2,\ldots,p]} a_{kl}X_{(i-1)k}X_{(i-1)l}+\sum_{k,l\in[1,2,\ldots,p]} b_{kl}X_{(i-1)k}X_{il}+\sum_{k,l\in[1,2,\ldots,p]} c_{kl}X_{ik}X_{il}\} \tag{2.17}$$

where a_{kl}, b_{kl} and c_{kl} are coefficients.

Note that when $T\to\infty$, the three terms $\frac{1}{\sqrt{T}}\sum_{k,l} a_{kl}X_{1k}X_{1l}$, $\frac{1}{\sqrt{T}}\sum_{k,l} b_{kl}X_{(T-1)k}X_{Tl}$ and $\frac{1}{\sqrt{T}}\sum_{k,l} c_{kl}X_{Tk}X_{Tl}$ all tends to 0 in probability hence is negligible when studying the asymptotic behavior of the linear combination 2.17. In this sense, the linear combination (2.17) is essentially equivalent to the linear combination

$$\sum_{i=2}^{T}\frac{1}{\sqrt{T}}\{\sum_{k,l\in[1,2,\ldots,p]} \beta_{kl}X_{(i-1)k}X_{il}+\sum_{k,l\in[1,2,\ldots,p]} \alpha_{kl}X_{ik}X_{il}\} \tag{2.18}$$

where the coefficients $\beta_{kl}=b_{kl}$ and $\alpha_{kl}=a_{kl}+c_{kl}$.

Here we denote the term $\sum_{k,l\in[1,2,\ldots,p]}\beta_{kl}X_{(i-1)k}X_{il}+\sum_{k,l\in[1,2,\ldots,p]}\alpha_{kl}X_{ik}X_{il}$ by Δ_i and note that it does not depend on the number T.

Now we define

$$G_{Ti} \equiv \frac{1}{\sqrt{T}}(\Delta_i - \mathbb{E}(\Delta_i))$$

$$S_{Ti} \equiv \sum_{k=1}^{i} G_{Tk}$$

$$\mathcal{F}_{Ti} \equiv \sigma(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_i)$$

and note that the σ-field $\mathcal{F}_{Ti}$ actually is independent of T. Then the tuple $\{S_{Ti}, \mathcal{F}_{Ti}, 2 \leq i \leq T, T \geq 1\}$ is a zero mean square integrable martingale array with difference G_{Ti} and the σ-field $\mathcal{F}_{Ti}$ is clearly nested. Now we can apply Corollary 3.1 of Hall and Heyde (2014) on this martingale array to obtain the asymptotic normality of the random matrix $\sqrt{T}\boldsymbol{N}$.

First we verify the conditional Lindeberg condition. Note that under our setup, the conditional Lindeberg condition is equivalent to the Lindeberg condition. So here we verify the later since it is easier. For any fixed $\epsilon > 0$, we have

$$\begin{aligned}
&\sum_{i=2}^{T} \mathbb{E}[G_{Ti}^2 I(|G_{Ti}| > \epsilon)] \\
&= \sum_{i=1}^{T} \mathbb{E}[\frac{(\Delta_i - \mathbb{E}(\Delta_i))^2}{T} I(|\Delta_i - \mathbb{E}(\Delta_i)| > \epsilon\sqrt{T})] \\
&= \mathbb{E}[(\Delta_1 - \mathbb{E}(\Delta_1))^2 I(|\Delta_1 - \mathbb{E}(\Delta_1)| > \epsilon\sqrt{T})] \xrightarrow{p} 0
\end{aligned}$$

as T tends to ∞. Hence the conditional Lindeberg condition is satisfied.

Second we verify the conditional variance condition. Here the conditional variance

$$\sum_{i=2}^{T} \mathbb{E}[G_{Ti}^2 | \mathcal{F}_{T,i-1}]$$
$$= \sum_{i=2}^{T} \frac{\mathbb{E}[(\Delta_i - \mathbb{E}(\Delta_i))^2 | \mathcal{F}_{T,i-1}]}{T}$$
$$= \sum_{i=1}^{T-1} \frac{f(\boldsymbol{X}_i)}{T}$$

for some function $f(.)$ from $\mathbb{R}^p$ to $\mathbb{R}$ where

$$\begin{aligned}
f(\boldsymbol{X}_{i-1}) &= \mathbb{E}[(\sum_{k,l} \beta_{kl} X_{(i-1)k} X_{il} + \sum_{k,l} \alpha_{kl} X_{ik} X_{il} \\
&- \mathbb{E}(\sum_{k,l} \alpha_{kl} X_{ik} X_{il}))^2 | \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{i-1}] \\
&= \mathbb{E}[(\sum_{k,l} \beta_{kl} X_{(i-1)k} X_{il})(\sum_{k,l} \beta_{kl} X_{(i-1)k} X_{il}) | \boldsymbol{X}_{i-1}] \\
&+ 2\mathbb{E}[(\sum_{k,l} \beta_{kl} X_{(i-1)k} X_{il})(\sum_{k,l} \alpha_{kl} X_{ik} X_{il}) | \boldsymbol{X}_{i-1}] \\
&- (\mathbb{E}[\sum_{k,l} \alpha_{kl} X_{ik} X_{il}])^2 \\
&= \sum_{k,l,m,n} \mathbb{E}[\beta_{kl} \beta_{mn} X_{il} X_{in}] X_{(i-1)k} X_{(i-1)m} \\
&+ 2 \sum_{k,l,m,n} \mathbb{E}[\beta_{kl} \alpha_{mn} X_{il} X_{im} X_{in}] X_{(i-1)k} \\
&- (\mathbb{E}[\sum_{k,l} \alpha_{kl} X_{ik} X_{il}])^2 \\
&= \sum_{k,l,m,n} \beta_{kl} \beta_{mn} \sigma_{ln} X_{(i-1)k} X_{(i-1)m} - (\sum_{k,l} \alpha_{kl} \sigma_{kl})^2
\end{aligned}$$

by the moment properties of multivariate normal and the mean 0 property that we assume. Then by LLN, we have

$$\sum_{i=1}^{T-1} \frac{f(\boldsymbol{X}_i)}{T} \xrightarrow{p} \mathbb{E} f(\boldsymbol{X}_1).$$

Hence the conditional variance condition is verified.

So we have proved the asymptotic normality of the random matrix $\sqrt{T}\boldsymbol{N}$ and what remains is to characterize its covariance structure. Let us denote $\boldsymbol{A} \equiv \boldsymbol{\Sigma}_x^{1/2}$ which is chosen to be symmetric and $\boldsymbol{z}_i \equiv \boldsymbol{A}^{-1/2}\boldsymbol{X}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ be an i.i.d sequence of standard normal random vectors. Then $\boldsymbol{Y}_i$ can be expressed as $\boldsymbol{\Gamma}\boldsymbol{u}_i$ where

$$\boldsymbol{\Gamma} \equiv \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{A} \end{bmatrix}, \boldsymbol{u}_i \equiv \begin{bmatrix} \boldsymbol{z}_i \\ \boldsymbol{z}_{i+1} \end{bmatrix}.$$

Let us further denote $\boldsymbol{\eta}_i \equiv \boldsymbol{Y}_i\boldsymbol{Y}_i'$. Due to the lag-1 structure of the $\boldsymbol{Y}_i$ series, the asymptotic covariance of $\sqrt{T}\boldsymbol{N}$ is essentially determined by $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1)$ and $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. The computation result of the first term has actually been given in Lemma 1 so we will only need to calculate the second term here.

By properties of Kronecker product, we have

$$\begin{aligned} \mathrm{Cov}(\mathrm{vec}(\boldsymbol{\eta}_1), \mathrm{vec}(\boldsymbol{\eta}_2)) &= \mathrm{Cov}(\boldsymbol{Y}_1 \otimes \boldsymbol{Y}_1, \boldsymbol{Y}_2 \otimes \boldsymbol{Y}_2) \\ &= \mathrm{Cov}((\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})(\boldsymbol{u}_1 \otimes \boldsymbol{u_1}), (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})(\boldsymbol{u}_2 \otimes \boldsymbol{u_2})) \\ &= (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\,\mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2)(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})'. \end{aligned}$$

So all we need is the calculation of the middle term $\mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2)$. Here we note that this matrix term is of size $4p^2 \times 4p^2$ which can be represented via block matrix as follows:

$$\begin{aligned} &\mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2) \qquad (2.19) \\ = &\begin{bmatrix} \mathrm{Cov}(\boldsymbol{z}_1 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') & \mathrm{Cov}(\boldsymbol{z}_1 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_3 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \\ \mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') & \mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_3 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \end{bmatrix}. \end{aligned}$$

Under the i.i.d normal assumption, it is easy to check the only possible non-zero block of this matrix in the representation (2.27) is the lower left block $\mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')')$ which is of size $2p^2 \times 2p^2$.

From the definition of Kronecker product, this lower left block can be further partitioned into a $p \times p$ block matrix where its (i, j)-th block is represented by

$$\mathrm{Cov}(z_{2i} \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', z_{2j} \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \qquad (2.20)$$

and is of size $2p \times 2p$. Still from the i.i.d assumption, we can represent this (i, j)-th block(of size $2p \times 2p$) by a 2×2 block matrix where the only possible non-zero block is the lower left one and is of size $p \times p$. This $p \times p$ lower left block takes the form $\mathrm{Cov}(z_{2i}\boldsymbol{z}_2, z_{2j}\boldsymbol{z}_2)$ from Kronecker product definition and can be further simplified with the normal moment results. To be more precise, under the normal setup, we have

$$\mathrm{Cov}(z_{2i}\boldsymbol{z}_2, z_{2j}\boldsymbol{z}_2) = \begin{cases} \boldsymbol{e}_j\boldsymbol{e}_i' & \text{when } i \neq j, \\ \boldsymbol{I}_p + \boldsymbol{e}_i\boldsymbol{e}_i' & \text{when } i = j. \end{cases}$$

Thus we have specified the covariance structure $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ and hence the final covariance structure is equal to the sum of the this term, its transpose and the covariance structure appeared in Lemma 1.

□

Proof of Theorem 1. We are going to prove the first part of the theorem while the second part follows directly from Corollary 11.3.8 of Muirhead (2009).

First we prove the asymptotic distribution of the roots of the equation

$$|T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} - \lambda\boldsymbol{I}_p| = 0 \tag{2.21}$$

and the asymptotic distribution of the roots of the equation

$$|\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21} - \lambda\boldsymbol{I}_p| = 0 \tag{2.22}$$

are the same.

Denote $A \equiv \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21} = T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} + O_p(T^{-3/2})$ and $B \equiv T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21}$. From Proposition 1, we have

$$\sum_{i=1}^{p} |\lambda_{Ai} - \lambda_{Bi}|^2 \leq ||A - B||_2^2$$

where λ_{Ai} and λ_{Bi} are the i-the largest eigenvalue of A and B respectively and the matrix norm here is the Frobenius norm. Since each entry of $A - B$ is of

order $O_p(T^{-3/2})$ and here the dimension p is fixed, we have $||A-B||_2$ is of order $O_p(T^{-3/2})$, and this implies for each i,

$$T^{1/2}(\lambda_{Ai} - \lambda_{Bi}) \xrightarrow{p} 0. \tag{2.23}$$

Note that it is well known that under the independent case (2.5), the rate of multiplication for λ_{Ai} to have a non-trivial asymptotic distribution(which is normal) is $T^{1/2}$, so from (2.23) we have $T^{1/2}\lambda_{Ai}$ and $T^{1/2}\lambda_{Bi}$ have the same asymptotically normal distribution. Similar property also hold for any fixed linear combination of λ_{Ai}. Then applying the Cramer-Wold device argument, this first part is justified.

Second we prove the distribution of $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ are asymptotically the same. To be more precise, we prove

$$\mathrm{Cov}(\boldsymbol{N}_{12}) = \mathrm{Cov}(\boldsymbol{M}_{12}) = \frac{1}{T}\boldsymbol{\Sigma}_x \otimes \boldsymbol{\Sigma}_x.$$

Then based on the covariance structure, the random matrices $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ have the same asymptotically normal distribution. Hence it is also true for the pair $\boldsymbol{K}_{12}$ and $\boldsymbol{W}_{12}$.

Here we note that $\boldsymbol{N}_{12} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{X}_i\boldsymbol{X}'_{i+1}$ and a term like $\boldsymbol{X}_i\boldsymbol{X}'_{i+1}$ in the expression is clearly independent of a term like $\boldsymbol{X}_j\boldsymbol{X}'_{j+1}$ whenever $|i-j| > 1$. So the covariance structure of $\boldsymbol{N}_{12}$ is mainly determined by $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$ and $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_2)$ where $\boldsymbol{G}_i \equiv \boldsymbol{X}_i\boldsymbol{X}'_{i+1}$. Since $\mathrm{Cov}(\boldsymbol{X}_{1i}\boldsymbol{X}_{2j}, \boldsymbol{X}_{2l}\boldsymbol{X}_{3m}) = 0$ for any index i, j, l, m we pick from $[1, 2, 3, \ldots, p]$, this essentially justifies $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_2) = \boldsymbol{0}$.

From the above justifications, we know the covariance structure of $\boldsymbol{N}_{12}$ is completely determined by the term $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$ and more precisely we have $\mathrm{Cov}(\boldsymbol{N}_{12}) = \frac{1}{T}\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$. Since $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1) = \mathrm{Cov}(\boldsymbol{X}_1\boldsymbol{U}'_1, \boldsymbol{X}_1\boldsymbol{U}'_1)$, it is clear that the two random matrices $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ have the same covariance structure. So what remains is to calculate $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$.

To compute this, we have

$$\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1) \equiv \mathrm{var}(\mathrm{vec}(\boldsymbol{X}_1 \boldsymbol{X}_2')) = \mathrm{Cov}(\boldsymbol{X}_2 \otimes \boldsymbol{X}_1, \boldsymbol{X}_2 \otimes \boldsymbol{X}_1)$$

$$= (A \otimes A)\,\mathrm{Cov}(\boldsymbol{z}_2 \otimes \boldsymbol{z}_1, \boldsymbol{z}_2 \otimes \boldsymbol{z}_1)(A \otimes A)' = (A \otimes A)(A \otimes A)' = \boldsymbol{\Sigma}_x \otimes \boldsymbol{\Sigma}_x$$

where $A \equiv \boldsymbol{\Sigma}_x^{1/2}$ and $\{\boldsymbol{z}_i\}$ are the standard multivariate normal random vectors. Hence the covariance justification is completed. Then combining the covariance equality fact above with the previous two Lemmas 2 and 1 on the asymptotic normality of the random matrices $\boldsymbol{N}$ and $\boldsymbol{M}$, we know the random matrices $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ have the same asymptotically normal distribution.

At this point we are ready to prove the last part for the lag-1 case (2.4) which is of similar fashion of the first part: the asymptotic distribution of the roots of the equation

$$|T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.24}$$

and the asymptotic distribution of the roots of the equation

$$|\boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.25}$$

are the same.

But we note that here in advance that the arguments for the this part are not exactly the same comparing to the first part. For this part, we denote $A \equiv \boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} = T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} + O_p(T^{-3/2})$ and $B \equiv T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21}$. Then from Proposition 1, we have

$$\sum_{i=1}^{p} |\lambda_{Ai} - \lambda_{Bi}|^2 \leq ||A - B||_2^2$$

where λ_{Ai} and λ_{Bi} are the i-th largest eigenvalue of A and B respectively and the matrix norm here is the Frobenius norm. Since each entry of $A - B$ is of

order $O_p(T^{-3/2})$ and here the dimension p is fixed, we have $||A-B||_2$ is of order $O_p(T^{-3/2})$, and this implies for each i

$$T^{1/2}(\lambda_{Ai} - \lambda_{Bi}) \xrightarrow{p} 0. \tag{2.26}$$

From the second part, we know that $T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} = \boldsymbol{N}_{12}\boldsymbol{N}_{21}$ and $T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} = \boldsymbol{M}_{12}\boldsymbol{M}_{21}$ have the same asymptotic distribution. Then by the first part, we know this time that the rate of multiplication for λ_{Bi} to have a non-trivial asymptotic distribution(which is normal) is $T^{1/2}$, so from (2.26) we have $T^{1/2}\lambda_{Ai}$ and $T^{1/2}\lambda_{Bi}$ have the same asymptotically normal distribution. This time we note that similar property also hold for any fixed linear combination of λ_{Bi}. Then still applying the Cramer-Wold device argument, this part is justified.

Combing the results from the three parts, it follows that the asymptotic distributions of the sample canonical correlations of calculated from (2.4) and (2.5) are identical.

□

Proof of Lemma 3. The multivariate normal part justification can be found in Section 6.3 of Bilodeau and Brenner (2008). Here we present the calculation on its covariance structure.

Let $\boldsymbol{W} = \boldsymbol{V}_1\boldsymbol{V}_1\top$, then

$$\operatorname{var}\boldsymbol{W} = (\Lambda_y^{1/2}) \otimes (\Lambda_y^{1/2}) \operatorname{var}(\boldsymbol{z}\boldsymbol{z}\top)(\Lambda_y^{1/2}) \otimes (\Lambda_y^{1/2})$$

where $\boldsymbol{z} \sim E_{2p}(\boldsymbol{0}, \boldsymbol{I})$ is rotationally invariant. Using Proposition 13.2 of Bilodeau and Brenner (2008) (which was originally established in Tyler (1982)), $\operatorname{var}(\boldsymbol{z}\boldsymbol{z}\top)$ can be evaluated with

$$a_1 = \operatorname{var}(z_1 z_2) = \mathbb{E}(z_1^2 z_2^2) = \mu_{22},$$

$$a_2 = \operatorname{Cov}(z_1^2, z_2^2) = \mathbb{E}(z_1^2 z_2^2) - \mathbb{E}(z_1^2)\mathbb{E}(z_2^2) = \mu_{22} - \mu_2^2.$$

In terms of cumulants we have $a_1 = k_{22} + k_2^2$ and $a_2 = k_{22}$ where k_i and k_{ij} are the cumulants. And by differentiation on $\phi(.)$ (see the second property in Proposition 3) one can verify that

$$k_2 = -2\phi'(0) \equiv \alpha,$$

$$k_4 = 12(\phi''(0) - \phi'(0)^2),$$

$$k_{22} = 4(\phi''(0) - \phi'(0)^2).$$

And we have for z_1,

$$\frac{k_4}{k_2^2} = 3\frac{\phi''(0) - \phi'(0)^2}{\phi'(0)^2} \equiv 3\kappa,$$

where κ is introduced as a parameter to facilitate our expression and can be interpreted as an excess kurtosis. So we have $k_4 = 3\kappa\alpha^2$ and $k_{22} = \kappa\alpha^2$. And finally we obtain $a_1 = (1+\kappa)\alpha^2$ and $a_2 = \kappa\alpha^2$ from which

$$\mathrm{var}(\boldsymbol{z}\boldsymbol{z}\top) = \alpha^2(1+\kappa)(\boldsymbol{I} + \boldsymbol{K}_{2p}) + \alpha^2\kappa\,\mathrm{vec}(\boldsymbol{I})[\mathrm{vec}(\boldsymbol{I})]\top$$

and

$$\begin{aligned}\mathrm{var}\,\boldsymbol{W} &= \alpha^2(1+\kappa)(\boldsymbol{I} + \boldsymbol{K}_{2p})(\boldsymbol{\Lambda}_y \otimes \boldsymbol{\Lambda}_y) + \alpha^2\kappa\,\mathrm{vec}(\boldsymbol{\Lambda}_y)[\mathrm{vec}(\boldsymbol{\Lambda}_y)]\top \\ &= (1+\kappa)(\boldsymbol{I} + \boldsymbol{K}_{2p})(\boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_y) + \kappa\,\mathrm{vec}(\boldsymbol{\Sigma}_y)[\mathrm{vec}(\boldsymbol{\Sigma}_y)]\top.\end{aligned}$$

Via Taylor's expansion and differentiation of $\phi(.)$, we can obtain $\phi'(0) = -\frac{1}{2}\mathbb{E}z_1^2$ and $\phi''(0) = \frac{\mathbb{E}z_1^4}{12}$. So we have

$$\kappa = \frac{\phi''(0) - \phi'(0)^2}{\phi'(0)^2} = \frac{\mathbb{E}z_1^4 - 3\mathbb{E}z_1^2}{3\mathbb{E}z_1^2}.$$

And this completes the characterization of the covariance structure. □

Proof of Lemma 4. The proof of the asymptotic normality part is essentially the same as Lemma 2, so we do not repeat this part here and go directly into establishing the covariance structure instead. Let us denote $\boldsymbol{A} \equiv \boldsymbol{\Lambda}_x^{1/2}$ which is chosen

to be symmetric and $\boldsymbol{z}_i \sim E_p(0, \boldsymbol{I}_p)$ be and i.i.d sequence of rotationally invariant random vectors. Then $\boldsymbol{Y}_i$ can be expressed as $\boldsymbol{\Gamma}\boldsymbol{u}_i$ where

$$\boldsymbol{\Gamma} \equiv \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{A} \end{bmatrix}, \boldsymbol{u}_i \equiv \begin{bmatrix} \boldsymbol{z}_i \\ \boldsymbol{z}_{i+1} \end{bmatrix}.$$

Let us further denote $\boldsymbol{\eta}_i \equiv \boldsymbol{Y}_i \boldsymbol{Y}_i'$. Due the lag-1 structure of the $\boldsymbol{Y}_i$ series, the asymptotic covariance of $\sqrt{T}\boldsymbol{N}$ is essentially determined by $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_1)$ and $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. The computation of the first term has actually been given in Lemma 3 so we will only need to calculate the second term here.

By properties of Kronecker product, we have

$$\begin{aligned} \mathrm{Cov}(\mathrm{vec}(\boldsymbol{\eta}_1), \mathrm{vec}(\boldsymbol{\eta}_2)) &= \mathrm{Cov}(\boldsymbol{Y}_1 \otimes \boldsymbol{Y}_1, \boldsymbol{Y}_2 \otimes \boldsymbol{Y}_2) \\ &= \mathrm{Cov}((\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})(\boldsymbol{u}_1 \otimes \boldsymbol{u_1}), (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})(\boldsymbol{u}_2 \otimes \boldsymbol{u_2})) \\ &= (\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\, \mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2)(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})'. \end{aligned}$$

So all we need is the calculation of the middle term $\mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2)$. Here we note that this matrix term is of size $4p^2 \times 4p^2$ which can be represented via block matrix as follows:

$$\mathrm{Cov}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1, \boldsymbol{u}_2 \otimes \boldsymbol{u}_2) \tag{2.27}$$

$$= \begin{bmatrix} \mathrm{Cov}(\boldsymbol{z}_1 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') & \mathrm{Cov}(\boldsymbol{z}_1 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_3 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \\ \mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') & \mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_3 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \end{bmatrix}.$$

Under the i.i.d elliptical assumption, it is easy to check the only possible non-zero block of this matrix in the representation (2.27) is the lower left block $\mathrm{Cov}(\boldsymbol{z}_2 \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', \boldsymbol{z}_2 \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')')$ which is of size $2p^2 \times 2p^2$.

From the definition of Kronecker product, this lower left block can be further partitioned into a $p \times p$ block matrix where its (i, j)-th block is represented by

$$\mathrm{Cov}(z_{2i} \otimes (\boldsymbol{z}_1', \boldsymbol{z}_2')', z_{2j} \otimes (\boldsymbol{z}_2', \boldsymbol{z}_3')') \tag{2.28}$$

and is of size $2p \times 2p$. Still from the i.i.d assumption, we can represent this (i, j)-th block(of size $2p \times 2p$) by a 2×2 block matrix where the only possible non-zero

block is the lower left one and is of size $p \times p$. This $p \times p$ lower left block takes the form $\mathrm{Cov}(z_{2i}\boldsymbol{z}_2, z_{2j}\boldsymbol{z}_2)$ from Kronecker product definition. Thus we have specified the covariance structure $\mathrm{Cov}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ and hence the final covariance structure is equal to the sum of the this term, its transpose and the covariance structure appeared in Lemma 3. □

Proof of Theorem 2. The proof for the first part of the theorem is essentially the same as the normal case with some minor modifications. First we prove the asymptotic distribution of the roots of the equation

$$|T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.29}$$

and the asymptotic distribution of the roots of the equation

$$|\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.30}$$

are the same.

Denote $A \equiv \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21} = T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} + O_p(T^{-3/2})$ and $B \equiv T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21}$. From Proposition 1, we have

$$\sum_{i=1}^{p} |\lambda_{Ai} - \lambda_{Bi}|^2 \leq ||A - B||_2^2$$

where λ_{Ai} and λ_{Bi} are the i-the largest eigenvalue of A and B respectively and the matrix norm here is the Frobenius norm. Since each entry of $A - B$ is of order $O_p(T^{-3/2})$ and here the dimension p is fixed, we have $||A - B||_2$ is of order $O_p(T^{-3/2})$, and this implies for each i

$$T^{1/2}(\lambda_{Ai} - \lambda_{Bi}) \xrightarrow{p} 0. \tag{2.31}$$

Note that it is well known that under the independent case (2.5), the rate of multiplication for λ_{Ai} to have a non-trivial asymptotic distribution(which is normal) is $T^{1/2}$, so from (2.31) we have $T^{1/2}\lambda_{Ai}$ and $T^{1/2}\lambda_{Bi}$ have the same asymptotically normal distribution. Similar property also hold for any fixed linear combination of λ_{Ai}. Then applying the Cramer-Wold device argument, this first part is justified.

Second we prove that the distribution of the off-diagonal blocks $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ are asymptotically the same. Or equivalently, we have $\mathrm{Cov}(\boldsymbol{N}_{12}) = \mathrm{Cov}(\boldsymbol{M}_{12})$. Since $\boldsymbol{N}_{12} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{X}_i \boldsymbol{X}'_{i+1}$ and a term like $\boldsymbol{X}_i \boldsymbol{X}'_{i+1}$ in the expression is clearly independent of a term like $\boldsymbol{X}_j \boldsymbol{X}'_{j+1}$ whenever $|i-j| > 1$, the covariance structure of $\boldsymbol{N}_{12}$ is mainly determined by $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$ and $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_2)$ where $\boldsymbol{G}_i \equiv \boldsymbol{X}_i \boldsymbol{X}'_{i+1}$. Note that $\mathrm{Cov}(\boldsymbol{X}_{1i}\boldsymbol{X}_{2j}, \boldsymbol{X}_{2l}\boldsymbol{X}_{3m}) = 0$ for any index i, j, l, m we pick from $[1, 2, 3, \ldots, p]$, this essentially justifies $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_2) = \boldsymbol{0}$. So the covariance structure of $\boldsymbol{N}_{12}$ is completely determined by the term $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$ and more precisely we have $\mathrm{Cov}(\boldsymbol{N}_{12}) = \frac{1}{T}\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1)$. From definitions, we know $\mathrm{Cov}(\boldsymbol{G}_1, \boldsymbol{G}_1) = \mathrm{Cov}(\boldsymbol{X}_1\boldsymbol{U}'_1, \boldsymbol{X}_1\boldsymbol{U}'_1)$, so it is clear that the two random matrices $\boldsymbol{N}_{12}$ and $\boldsymbol{M}_{12}$ have the same covariance structure.

At this point we are ready to prove the last part for the lag-1 case (2.4) which is of similar fashion of the first part: The asymptotic distribution of the roots of the equation

$$|T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.32}$$

and the asymptotic distribution of the roots of the equation

$$|\boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} - \lambda \boldsymbol{I}_p| = 0 \tag{2.33}$$

are the same.

But we note that here in advance that the arguments for the this part are not exactly the same comparing to the first part. For this part, we denote $A \equiv \boldsymbol{N}_{11}^{-1}\boldsymbol{N}_{12}\boldsymbol{N}_{22}^{-1}\boldsymbol{N}_{21} = T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} + O_p(T^{-3/2})$ and $B \equiv T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21}$. Then from Proposition 1, we have

$$\sum_{i=1}^{p} |\lambda_{Ai} - \lambda_{Bi}|^2 \leq ||A - B||_2^2$$

where λ_{Ai} and λ_{Bi} are the i-th largest eigenvalue of A and B respectively and the matrix norm here is the Frobenius norm. Since each entry of $A - B$ is of order $O_p(T^{-3/2})$ and here the dimension p is fixed, we have $||A - B||_2$ is of order $O_p(T^{-3/2})$, and this implies for each i

$$T^{1/2}(\lambda_{Ai} - \lambda_{Bi}) \xrightarrow{p} 0. \tag{2.34}$$

From the second part, we know that $T^{-1}\boldsymbol{W}_{12}\boldsymbol{W}_{21} = \boldsymbol{N}_{12}\boldsymbol{N}_{21}$ and $T^{-1}\boldsymbol{K}_{12}\boldsymbol{K}_{21} = \boldsymbol{M}_{12}\boldsymbol{M}_{21}$ have the same asymptotic distribution. Then by the first part, we know this time that the rate of multiplication for λ_{Bi} to have a non-trivial asymptotic distribution(which is normal) is $T^{1/2}$, so from (2.34) we have $T^{1/2}\lambda_{Ai}$ and $T^{1/2}\lambda_{Bi}$ have the same asymptotically normal distribution. This time we note that similar property also hold for any fixed linear combination of λ_{Bi}. Then still applying the Cramer-Wold device argument, this part is justified.

Combing the results from the above three parts of argument, it follows that the asymptotic distributions of the sample canonical correlations of calculated from (2.4) and (2.5) are identical so the first part of the theorem is justified.

Now what remains is to justify the second part of the theorem which presents the explicit density form of the joint distribution. Note that from the part we have just proved, we only need to obtain the asymptotic joint density of the canonical correlations for the independent case (2.5) under the elliptical assumption.

Applying Theorem 5.2 and Proposition 5.3 of Eaton and Tyler (1994) to our elliptical context, we have the asymptotic joint distribution of $\sqrt{T}\hat{r}_1, \sqrt{T}\hat{r}_2, ..., \sqrt{T}\hat{r}_p$ is the same as the joint distribution of the ordered singular values of the random matrix $\boldsymbol{Z}$, hence the asymptotic joint distribution of $T\hat{r}_1^2, ..., T\hat{r}_p^2$ is the same as the joint distribution of the ordered eigenvalues of $\boldsymbol{Z}\boldsymbol{Z}'$. Also from Proposition 5.3 of Eaton and Tyler (1994), we have the random matrix $\boldsymbol{Z}\boldsymbol{Z}'$ follows a $Wishart_p(p, (1+\gamma)\boldsymbol{I}_p)$ distribution. Then applying Corollary 9.4.2 of Muirhead (2009) to the matrix $\boldsymbol{Z}\boldsymbol{Z}'$ via a simple change of variable $w_i = nl_i$(here n and l_i bear the same meaning as in the Corollary 9.4.2 of Muirhead (2009)), we obtained the joint density presented in the second part of this theorem.

□

Proof of Proposition 4. Let us denote $B_n \equiv \mathbb{E}\{X_n^4 I(X_n^2 \geq \epsilon\sqrt{n})\}$. First we note that if B_n does not converge to 0, then the event $a_n^2 \geq \epsilon\sqrt{n}$ must occur infinitely often, or equivalent, we say there exists an infinite subsequence of $\{a_i\}_{i=1}^{\infty}$, denote by $\{a_{k_i}\}_{i=1}^{\infty}$, such that for every i we have $a_{k_i}^2 \geq \epsilon\sqrt{k_i}$. Conversely, if such an infinite subsequence exists, then by the following calculation:

$$B_{k_i} \equiv \mathbb{E}\{X_{k_i}^4 I(X_{k_i}^2 \geq \epsilon\sqrt{k_i})\} \geq \frac{(\epsilon\sqrt{k_i})^2}{k_i} = \epsilon^2,$$

B_n must NOT converge to 0.

Now suppose such an infinite subsequence $\{a_{k_i}\}_{i=1}^{\infty}$ exists, we have

$$A_{k_i} - A_{k_i-1} = \frac{a_{k_i}^4}{k_i} + (\frac{k_i - 1}{k_i} - 1)A_{k_i-1} \geq \epsilon^2 + (\frac{k_i - 1}{k_i} - 1)A_{k_i-1}.$$

This indicates $\{A_n\}_{n=1}^{\infty}$ can not be a Cauchy sequence which contradicts the convergence condition of $\{A_n\}_{n=1}^{\infty}$. Hence B_n must converge to 0.

□

Proof of Proposition 5. It is easy to see that $\mathbb{E}(X_n^2) = \frac{\sum_{i=1}^n (a_i - \bar{a}_n)^2}{n}$ under our setup. For this centralized case, we denote $B_n \equiv \mathbb{E}\{(X_n^2 - \mathbb{E}(X_n^2))^2 I(|X_n^2 - \mathbb{E}(X_n^2)| \geq \epsilon\sqrt{n})\}$. Suppose B_n does not converge to 0, then since $\mathbb{E}(X_n^2) = \frac{\sum_{i=1}^n (a_i - \bar{a}_n)^2}{n}$ is bounded for all n, this implies the event $(a_n - \bar{a}_n)^2 \geq \epsilon\sqrt{n}$ must occur infinitely often. And since $\bar{a}_n$ is also bounded for all n, this again implies that the event $a_n^2 \geq \epsilon\sqrt{n}$ must occur infinitely often, or equivalent, there exists an infinite subsequence of $\{a_i\}_{i=1}^{\infty}$, denote by $\{a_{k_i}\}_{i=1}^{\infty}$, such that for every i we have $a_{k_i}^2 \geq \epsilon\sqrt{k_i}$.

Now apply similar Cauchy sequence arguments on $\{A_n\}_{n=1}^{\infty}$ as in Proposition 4, we get a contradiction and the proof is completed.

□

Proof of Proposition 6. Use similar arguments as in Proposition 5. □

Proof of Proposition 7. Use similar arguments as in Proposition 5. □

Proof of Lemma 5. We prove using the similar arguments which appear in Lemma 2. Without ambiguity, we suppress the global conditioning notation $(.|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T, \ldots)$ in this proof.

Note that the entries of $\sqrt{T}\boldsymbol{S}^{(b)}$ are of two types:

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T} X^*_{Ti,k}X^*_{Ti,l} \text{ and } \frac{1}{\sqrt{T}}\sum_{i=2}^{T} X^*_{T(i-1),k}X^*_{Ti,l}$$

where k and l are indices ranging from $[1, \ldots, p]$. So a linear combination of all the entries of $\sqrt{T}\boldsymbol{S}^{(\boldsymbol{b})}$ can be expressed as

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\sum_{k,l\in[1,2,\ldots,p]} a_{kl}X^*_{Ti,k}X^*_{Ti,l} + \frac{1}{\sqrt{T}}\sum_{i=2}^{T}\sum_{k,l\in[1,2,\ldots,p]} b_{kl}X^*_{T(i-1),k}X^*_{Ti,l} \tag{2.35}$$

where a_{kl} and b_{kl} are coefficients.

Under mild conditions on the sample path, when $T \to \infty$, the term

$$\frac{1}{\sqrt{T}}\sum_{k,l\in[1,2,\ldots,p]} a_{kl}X^*_{T1,k}X^*_{T1,l}$$

would be negligible for studying the asymptotic behavior of the linear combination (2.35). In this sense, the linear combination (2.35) is essentially equivalent to the linear combination

$$\sum_{i=2}^{T}\frac{1}{\sqrt{T}}\{\sum_{k,l\in[1,2,\ldots,p]} b_{kl}X^*_{T(i-1),k}X^*_{Ti,l} + \sum_{k,l\in[1,2,\ldots,p]} a_{kl}X^*_{Ti,k}X^*_{Ti,l}\}. \tag{2.36}$$

Here we define the following terms and note that these all depend on the number T:

$$\Delta_{Ti} \equiv \sum_{k,l\in[1,2,\ldots,p]} b_{kl}X^*_{T(i-1),k}X^*_{Ti,l} + \sum_{k,l\in[1,2,\ldots,p]} a_{kl}X^*_{Ti,k}X^*_{Ti,l},$$

$$G_{Ti} \equiv \frac{1}{\sqrt{T}}(\Delta_{Ti} - \mathbb{E}(\Delta_{Ti})),$$

$$S_{Ti} \equiv \sum_{k=1}^{i} G_{Tk},$$
$$\mathcal{F}_{Ti} \equiv \sigma(\boldsymbol{X}^*_{T1}, \boldsymbol{X}^*_{T2}, \ldots, \boldsymbol{X}^*_{Ti}).$$

Then the tuple $\{S_{Ti}, \mathcal{F}_{Ti}, 2 \leq i \leq T, T \geq 1\}$ is a zero mean square integrable martingale array with difference G_{Ti} and the σ-field $\mathcal{F}_{Ti}$ is nested. Now we investigate under what conditions on the sample path can we apply Corollary 3.1 of Hall and Heyde (2014) on this martingale array to obtain the conditional asymptotic normality of the random matrix $\sqrt{T}\boldsymbol{S}^{(b)}$ given $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T$.

First we investigate the conditional Lindeberg condition. As noted before, here the conditional Lindeberg condition is equivalent to the Lindeberg condition. So we study the later instead. To be more specific, we need to check for what conditions on the sample path, we could have for any fixed $\epsilon > 0$,

$$\begin{aligned}
&\sum_{i=2}^{T} \mathbb{E}[G^2_{Ti} I(|G_{Ti}| > \epsilon)] \\
&= \sum_{i=2}^{T} \mathbb{E}[\frac{(\Delta_{Ti} - \mathbb{E}(\Delta_{Ti}))^2}{T} I(|\Delta_{Ti} - \mathbb{E}(\Delta_{Ti})| > \epsilon\sqrt{T})] \\
&= \frac{T-1}{T} \mathbb{E}[(\Delta_{T2} - \mathbb{E}(\Delta_{T2}))^2 I(|\Delta_{T2} - \mathbb{E}(\Delta_{T2})| > \epsilon\sqrt{T})] \xrightarrow{p} 0
\end{aligned}$$

as T tends to ∞.

Note that there are three types of terms in Δ_{Ti}:

1. $X^*_{T(i-1),k} X^*_{Ti,l}$,
2. $X^*_{Ti,k} X^*_{Ti,l}$ where $k \neq l$,
3. $(X^*_{Ti,k})^2$.

When $\mathbb{E}(\Delta_{T2})$ is bounded for all T, it is easy to see that if

$$\mathbb{E}[(\Delta_{T2} - \mathbb{E}(\Delta_{T2}))^2 I(|\Delta_{T2} - \mathbb{E}(\Delta_{T2})| > \epsilon\sqrt{T})] \xrightarrow{p} 0 \tag{2.37}$$

does not hold for general linear combination, then it must **NOT** hold for at least one the three special cases:

1. $\Delta_{T2} = X^*_{T1,k} X^*_{T2,l}$,

2. $\Delta_{T2} = X^*_{T2,k} X^*_{T2,l}$ where $k \neq l$,

3. $\Delta_{T2} = (X^*_{T2,k})^2$.

Then when the following conditions on the infinite sample path are satisfied:

1. $\frac{\sum_{i=1}^T X_{i,k}^4}{T}$ converges for all $k \in [1, \ldots, p]$,

2. $\frac{\sum_{i=1}^T X_{i,k}^2 X_{i,l}^2}{T}$ converges for all $k \neq l$ and $k, l \in [1, \ldots, p]$,

3. $\bar{X}_{T,k} \equiv \frac{\sum_{i=1}^T X_{i,k}}{T}$ is bounded for all $k \in [1, \ldots, p]$,

4. $\frac{\sum_{i=1}^T (X_{i,k} - \bar{X}_{T,k})^2}{T}$ is bounded for all $k \in [1, \ldots, p]$,

5. $\frac{\sum_{i=1}^T (X_{i,k} - \bar{X}_{T,k})(X_{i,l} - \bar{X}_{T,l})}{T}$ is bounded for all $k \neq l$ and $k, l \in [1, \ldots, p]$,

we can apply Proposition 6 to the special case (2), Proposition 7 to special case (3) and Proposition 5 to special case (1) to conclude that (2.37) holds given the infinite sample path $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_T, \cdots$. We also note that $\mathbb{E}(\Delta_{T2})$ is bounded is actually implied by the conditions above.

Recall that we have assumed the finite fourth moment condition throughout this section, and this implies all the above conditions would be satisfied almost surely. Hence we have justified that the conditional Lindeberg condition holds almost surely.

Second we investigate the conditional variance condition. Here the conditional variance

$$\sum_{i=2}^{T} \mathbb{E}[G_{Ti}^2 | \mathcal{F}_{T,i-1}] = \sum_{i=2}^{T} \frac{\mathbb{E}[(\Delta_{Ti} - \mathbb{E}(\Delta_{Ti}))^2 | \mathcal{F}_{T,i-1}]}{T}. \tag{2.38}$$

First we check for the special case (1) when $\Delta_{Ti} = X^*_{T(i-1),k} X^*_{Ti,l}$. we have

$$\begin{aligned}
&\sum_{i=2}^{T} \frac{\mathbb{E}[(\Delta_{Ti} - \mathbb{E}(\Delta_{Ti}))^2 | \mathcal{F}_{T,i-1}]}{T} \\
&= (\frac{1}{T}\sum_{i=1}^{T} X^2_{i,l})(\frac{1}{T}\sum_{i=1}^{T}(X^*_{Ti,k})^2) \\
&\xrightarrow{p} \sigma_k^2 \sigma_l^2
\end{aligned}$$

for almost all sample paths. Note that the Markov inequality is used above.

Other cases can be similarly justified and this condition also holds almost surely. Combining all the above we have proved the asymptotic normality of the random matrix $\sqrt{T}\boldsymbol{S}^{(b)}$ given $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T$.

□

Proof of Theorem 3. Note that from the multivariate normality established in Lemma 5, essentially we have,

$$n(\boldsymbol{D}_n - \boldsymbol{0}) \xrightarrow{d} \boldsymbol{D} \tag{2.39}$$

and

$$n(\boldsymbol{D}^*_n - \boldsymbol{0}) \xrightarrow{d} \boldsymbol{D} \quad a.s. \tag{2.40}$$

for some distribution $\boldsymbol{D}$ where

$$\boldsymbol{D}_n \equiv \boldsymbol{S}_{11}^{-1/2} \boldsymbol{S}_{12} \boldsymbol{S}_{22}^{-1} \boldsymbol{S}_{21} \boldsymbol{S}_{11}^{-1/2}$$

and

$$\boldsymbol{D}^*_n \equiv (\boldsymbol{S}^{(b)}_{11})^{-1/2} \boldsymbol{S}^{(b)}_{12} (\boldsymbol{S}^{(b)}_{22})^{-1} \boldsymbol{S}^{(b)}_{21} (\boldsymbol{S}^{(b)}_{11})^{-1/2}.$$

Then by the continuous mapping theorem (see for example Theorem 2.3 in Van der Vaart (2000)), we have for any fixed $c \in \mathbb{R}$, along almost all the sample sequences $\omega = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T, \ldots)$,

$$\lim_{T\to\infty} \mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \le c | \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T) - \mathbb{P}(\sqrt{T}\hat{r_1} \le c) \xrightarrow{a.s.} 0.$$

Denote $F_T^{\omega}(c) \equiv \mathbb{P}(\sqrt{T}\hat{r_1}^{(b)} \leq c|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T)$, $G_T(c) \equiv \mathbb{P}(\sqrt{T}\hat{r_1} \leq c)$, $C = \{c_j, j \geq 1\} \equiv$ set of rational numbers enlarged by any irrational discontinuity points of $\{G_T(.)\}$ and sets

$$A_1 = \{\omega : F_T^{\omega}(c_j) - G_T(c_j) \to 0, j \geq 1\},$$

$$A_2 = \{\omega : F_T^{\omega}(c) \textbf{ is a probability distribution fucntion}, T \geq 1\}.$$

It follows that $\mathbb{P}A_1 = 1$ and $\mathbb{P}A_2 = 1$. hence $\mathbb{P}A_1A_2 = 1$. Then from Corollay 1 in Section 8.2 of Chow and Teicher (2012), we have $F_T^{\omega} - G_T \xrightarrow{c} 0$ for $\omega \in A_1A_2$, then by Lemma 3 in the same section of Chow and Teicher (2012), $F_T^{\omega}(c) - G_T(c)$ converges uniformly to 0 over $-\infty < c < \infty$ for $\omega \in A_1A_2$ and this completes the proof. □

Chapter 3

Weighted Least Square Estimation of Autoregressive Models for Matrix-Valued Time Series

3.1 Introduction

Matrix and tensor type data are becoming more prevalent recently, and have been extensively studied for independent samples. On the other hand, matrix/tensor observations generated through the time usually exhibit temporal dependence, and require new tools and analysis to understand their dynamics and make predictions. In a pioneering work, Chen et al. (2020) proposed an autoregressive model for matrix-valued time series, and studied three estimators: projection estimator, least squares estimator (LSE), and maximum likelihood estimator (MLE). The MLE is derived under an additional assumption that the covariance tensor of the error matrix has a special form of a Kronecker outer product. It has been found that the MLE usually leads to the smallest estimation and prediction error, comparing with the other two estimators, even if that additional assumption fails to hold. A possible explanation of this phenomenon is that the MLE approach amounts to estimating the error convariance tensor by a Kronecker product. Although this can be a biased estimator, it is nevertheless a better one than the scalar tensor, and also can be viewed as a regularized version of the sample covariance tensor. Similar phenomenon has been observed in the covariance matrix estimation context, see for example Bickel and Levina (2004), Bickel et al. (2008) and Ledoit and Wolf (2003).

Motivated by these works, we propose a weighted least squares estimator (WLSE), where the weights correspond with the marginal variances of the elements of the error matrix. Essentially this is incorporating a covariance tensor estimator, given by the scalar tensor consisting of all marginal sample variances of the residual matrices, into the least squares estimation. Again, although this estimator can be biased, it nevertheless leads to a consistently estimator that can be more efficient than the direct LSE. We establish the central limit theorem of the WLSE, and demonstrate its performances through numerical studies and an example on economic indicators, comparing with other estimators.

The rest of this chapter is organized as follows. In Section 3.2 we define the matrix-valued autoregressive time series model and consider a weighted least square approach for the parameter estimation which leads to our WLSE. Then We prove a central limit theorem for the WLSE. In Section 3.3 and 3.4, we carried out some simulation studies and a real data analysis to show the performance of the WLSE. The proof are relegated in Section 3.5.

3.2 Weighted Least Square Estimator

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T$ be a matrix-valued time series of length T where each observation $\boldsymbol{X}_i$ is of size $m \times n$. Let vec(.) be the vectorization of a matrix by stacking the columns. After vectorization, we can directly apply the traditional vector autoregressive model (VAR) of order 1 to $\operatorname{vec} \boldsymbol{X}_t$ and have the following VAR(1) representation:

$$\operatorname{vec}(\boldsymbol{X}_t) = \Phi \operatorname{vec}(\boldsymbol{X}_{t-1}) + \operatorname{vec}(\boldsymbol{E}_t) \tag{3.1}$$

where Φ is the coefficient matrix of size $mn \times mn$ and $\boldsymbol{E}_t$ is the matrix innovation of size $m \times n$.

From this VAR(1) representation, it can be readily observed that the roles of

rows and columns are mixed, the matrix structure of the original series has not been fully utilized and the interpretation of the coefficient matrix Φ might be a difficult task.

To overcome these drawbacks, Chen et al. (2020) propose the *matrix autoregressive model (of order 1)*, denoted by MAR(1), which takes the form

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}' + \boldsymbol{E}_t, \tag{3.2}$$

where $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are $m \times m$ and $n \times n$ autoregressive coefficient matrices, and $\boldsymbol{E}_t = (e_{t,ij})$ is a $m \times n$ matrix white noise.

We note that after taking the vec(.) operation on both sides of (3.2), the MAR(1) model can be represented in the form of VAR(1)

$$\text{vec}(\boldsymbol{X}_t) = (\boldsymbol{B} \otimes \boldsymbol{A})\,\text{vec}(\boldsymbol{X}_{t-1}) + \text{vec}(\boldsymbol{E}_t), \tag{3.3}$$

where $\otimes$ denotes the matrix Kronecker product. So the model MAR(1) can be viewed as a special case of the model VAR(1) with coefficient matrix in Kronecker product form.

It is also worth pointing out that there is an identifiability issue regarding the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the MAR(1) model in (3.2) as the model remains the same if we divide A and multiply B by the same nonzero constant. To avoid this identifiability issue, we use the convention that $\boldsymbol{A}$ is normalized with Frobenius norm 1.

In (3.2), the innovation series $\{\boldsymbol{E}_t\}$ is assumed to be a matrix white noise. This means $\text{Cov}(\text{vec}(\boldsymbol{E}_t), \text{vec}(\boldsymbol{E}_s)) = \boldsymbol{0}$ as long as $t \neq s$ while at the same time $\Sigma \equiv \text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is not necessarily diagonal.

It is interesting to note that when solving the MLE for a multivariate regression or a vector autoregressive (VAR) estimation problem, the covariance structure of the innovation is irrelevant. However this is not correct for the matrix-valued autoregressive time series model like (3.2) as both the left and

right matrix multiplication are introduced in the model specification. Though the LSE approach is applicable for general covariance structure, it fails to incorporate in the covariance structure information of the matrix innovation. So it is important to ask whether we can extend the LSE procedure with this type of covariance information integrated. One natural idea is to consider a modified minimization problem with the covariance structure involved and solve it.

Given the observations $\{\boldsymbol{X}_t\}_{t=1}^T$, our weighted least square approach is aiming to solve the following optimization problem:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{t=2}^{T} \|(\boldsymbol{X}_t - \boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}') \circ \boldsymbol{W}\|_F^2 \tag{3.4}$$

where $\boldsymbol{W} \equiv (w_{ij}) \in \mathbb{R}^{mn}$ is a given weight matrix with entries $w_{ij} = \frac{1}{\sqrt{\text{var}(e_{t,ij})}}$. In practice, as the true values of these variances are not available, we use the sample estimates of these marginal variances (for example, estimated from the residuals after fitting a VAR(1) model to the data) instead to form a sample version weight matrix $\hat{\boldsymbol{W}}$ for the computation of our WLSE.

To solve the WLS problem (3.4), we can take the partial derivatives of the objective function with respect to the entries of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively to obtain the gradient condition for the problem. But first we need to introduce some notations to facilitate our expression.

Let us denote $f(\boldsymbol{A}, \boldsymbol{B}) \equiv \sum_{t=2}^{T} \|(\boldsymbol{X}_t - \boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}') \circ \boldsymbol{W}\|_F^2$. From matrix calculus, the trace function enjoys the following three differentiation properties:

- $\frac{\partial \text{Tr}[XY]}{\partial X} = 2Y'$,
- $\frac{\partial \text{Tr}[XAX']}{\partial X} = XA' + XA$,
- $\text{Tr}[(A \circ W)B] = \text{Tr}[A(W' \circ B)]$.

Thus we can apply these properties to take partial derivatives of $f(.)$ with respect to the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and set them to $\boldsymbol{0}$ to obtain the following gradient condition for the WLS minimization problem (3.4):

$$\begin{aligned}
&\sum_{t=2}^{T}(\boldsymbol{W}\circ\boldsymbol{W}\circ(\boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}'))\boldsymbol{B}\boldsymbol{X}_{t-1}'-\sum_{t=2}^{T}(\boldsymbol{W}\circ\boldsymbol{W}\circ\boldsymbol{X}_t)\boldsymbol{B}\boldsymbol{X}_{t-1}'=0\\
&\sum_{t=2}^{T}(\boldsymbol{W}'\circ\boldsymbol{W}'\circ(\boldsymbol{B}\boldsymbol{X}_{t-1}'\boldsymbol{A}'))\boldsymbol{A}\boldsymbol{X}_{t-1}-\sum_{t=2}^{T}(\boldsymbol{W}'\circ\boldsymbol{W}'\circ\boldsymbol{X}_t')\boldsymbol{A}\boldsymbol{X}_{t-1}=0.
\end{aligned}\tag{3.5}$$

Note that we have a total number of m^2+n^2 equations here in system (3.5).

The function $\sum_t\|(\boldsymbol{X}_t-\boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}')\circ\boldsymbol{W}\|_F^2$ is guaranteed to have at least one global minimum, hence the solutions to (3.5) always exist. On the other hand, if two pairs of matrices $(\hat{\boldsymbol{A}},\hat{\boldsymbol{B}})$ and $(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}})$ both satisfy the gradient condition and $\hat{\boldsymbol{B}}\otimes\hat{\boldsymbol{A}}=\tilde{\boldsymbol{B}}\otimes\tilde{\boldsymbol{A}}$, we say they are the same solution of (3.5).

Similar as the least square procedure, to solve the WLS problem (3.4), starting with some initial values, we iteratively update the two matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ by updating the entries of one of them in the gradient condition (3.5) by solving system of linear equations while holding the entries of the other one fixed.

Now we are ready to state the asymptotic result for the WLS estimator.

Theorem 4 (Asymptotics of the WLSE)**.** *Define* $\boldsymbol{\alpha}\equiv\mathrm{vec}(\boldsymbol{A})$, $\boldsymbol{\beta}\equiv\mathrm{vec}(\boldsymbol{B}')$, $\boldsymbol{W}_t'\equiv[(\boldsymbol{B}\boldsymbol{X}_t')\otimes\boldsymbol{I}:\boldsymbol{I}\otimes(\boldsymbol{A}\boldsymbol{X}_t)]$, $\boldsymbol{M}\equiv\mathrm{diag}(\mathrm{vec}(\boldsymbol{W}\circ\boldsymbol{W}))$ *and* $\boldsymbol{H}\equiv\mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\boldsymbol{W}_t')+\boldsymbol{\gamma}\boldsymbol{\gamma}'$ *where* $\boldsymbol{\gamma}\equiv(\boldsymbol{\alpha}',\boldsymbol{0}')'\in\mathbb{R}^{m^2+n^2}$. *Let* $\Xi_4\equiv\boldsymbol{H}^{-1}\mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\Sigma\boldsymbol{M}\boldsymbol{W}_t')\boldsymbol{H}^{-1}$ *and* $\boldsymbol{V}\equiv[\boldsymbol{\beta}\otimes\boldsymbol{I},\boldsymbol{I}\otimes\boldsymbol{\alpha}]$. *Assume that the innovations* $\boldsymbol{E}_1,\ldots,\boldsymbol{E}_T$ *are i.i.d with mean zero, finite second moment and is absolutely continuous with respect to the Lebesgue measure. Also assume the causality condition* $\rho(\boldsymbol{A})\cdot\rho(\boldsymbol{B})<1$ *and* $\boldsymbol{A},\boldsymbol{B}$ *and* Σ *are nonsingular. Then for the WLSE, it holds that*

$$\sqrt{T}\begin{pmatrix}\mathrm{vec}(\hat{\boldsymbol{A}}-\boldsymbol{A})\\ \mathrm{vec}(\hat{\boldsymbol{B}}'-\boldsymbol{B}')\end{pmatrix}\Rightarrow N(\boldsymbol{0},\Xi_4);$$

and

$$\sqrt{T}[\mathrm{vec}(\hat{\boldsymbol{B}}')\otimes\mathrm{vec}(\hat{\boldsymbol{A}})-\mathrm{vec}(\boldsymbol{B}')\otimes\mathrm{vec}(\boldsymbol{A})]\Rightarrow N(\boldsymbol{0},\boldsymbol{V}\Xi_4\boldsymbol{V}').$$

Corollary 1. *The asymptotic results in Theorem 4 still hold when we replace the weight matrix* $\boldsymbol{W}$ *by its sample version* $\hat{\boldsymbol{W}}$.

3.3 Simulation Results

In this section, we compare the performances of the WLSE with the LSE and MLE introduced in Chen et al. (2020) under different settings for various choices of dimension pair (m, n) and sample size T.

Given a dimension pair (m, n), we simulate $\boldsymbol{X}_t$ according to model (3.2), and the entries of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are randomly generated and rescaled to guarantee the causality condition and the constraint $\|\boldsymbol{A}\|_F = 1$ is satisfied. The coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are fixed for a given specified simulation setting with multiple replications.

First we show the finite sample performances under five different settings of the innovation covariance structure $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$.

- Setting I: The covariance matrix $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is set to $\Sigma = \boldsymbol{I}$.
- Setting II: The covariance matrix $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is a mixture of random covariance matrix and a random diagonal matrix.
- Setting III: The covariance matrix $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is set to $\Sigma_c \otimes \Sigma_r$, the Kronecker product of two random covariance matrices Σ_c and Σ_r.
- Setting IV: The covariance matrix $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is a random diagonal marix.
- Setting V: The covariance matrix $\text{Cov}(\text{vec}(\boldsymbol{E}_t))$ is a mixture of random diagonal matrix and Kronecker product matrix generated similarly as in Setting III.

For each setting, we repeat the simulation 100 times, and show a box plot of

$$\log(\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2).$$

The simulation results for relatively small sample sizes are shown in Figures 3.1 to 3.5. In these figures, the dimensions m and n increase from top to bottom,

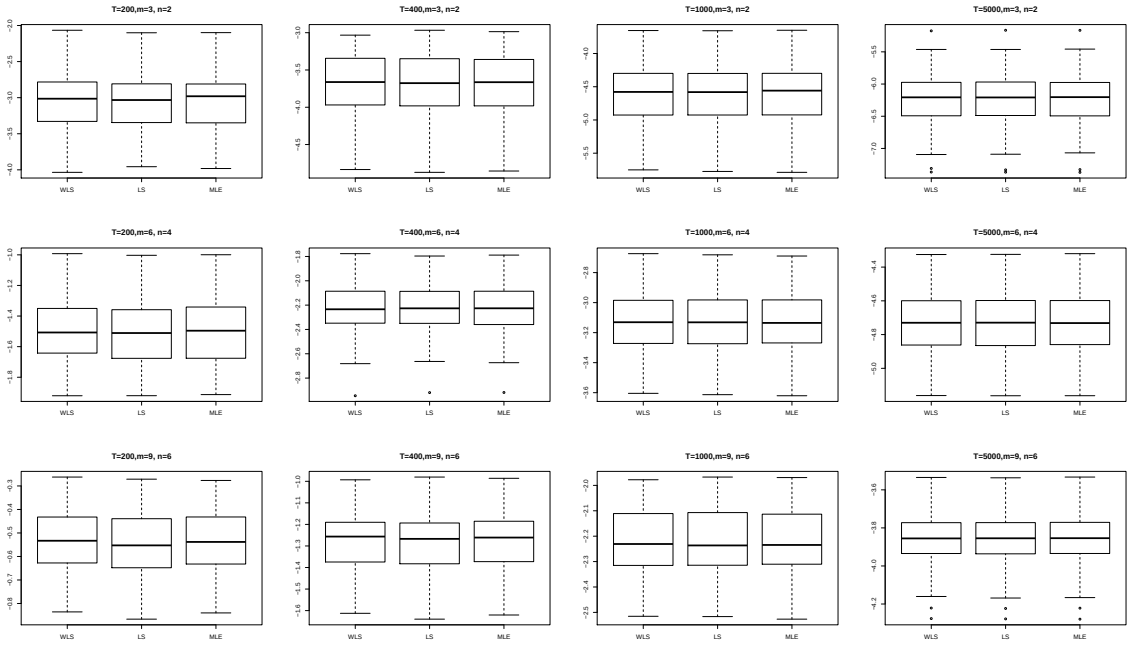

Figure 3.1: Comparison of three estimators, LS, WLS and MLE, under Setting I. The three rows correspond to $(m, n) = (3, 2)$, $(6, 4)$ and $(9, 6)$ respectively, and the four columns $T = 200$, 400, 1000 and 5000 respectively.

with values in $(m, n) = (3, 2), (6, 4), (9, 6)$. The sample size T increases from left to right at $T = 200, 400, 1000$ and 5000 respectively.

Under Setting I, we can observe from Figure 3.1 that LSE is the best estimator when the covariance matrix is identity. This is reasonable as it is the maximum likelihood estimator under this setting. And we note that the performance of MLE and WLSE are very close but slightly worse than the LSE. This is reasonable as both the MLEs and WLSEs need to estimate some additional parameters. Figures 3.2 to 3.5 show that both the MLE and WLSE outperform the LSE under Setting II to V. And from Figures 3.3 to 3.4, we see that the MLE outperforms WLSE under Setting III and the WLSE outperforms MLE under Setting IV which are as expected. When is innovation covariance structure is of the Kronecker plus diagonal mixture type, we observe from Figure 3.4 that the WLSE and MLE have comparable performance as long as the matrices used to form the mixture

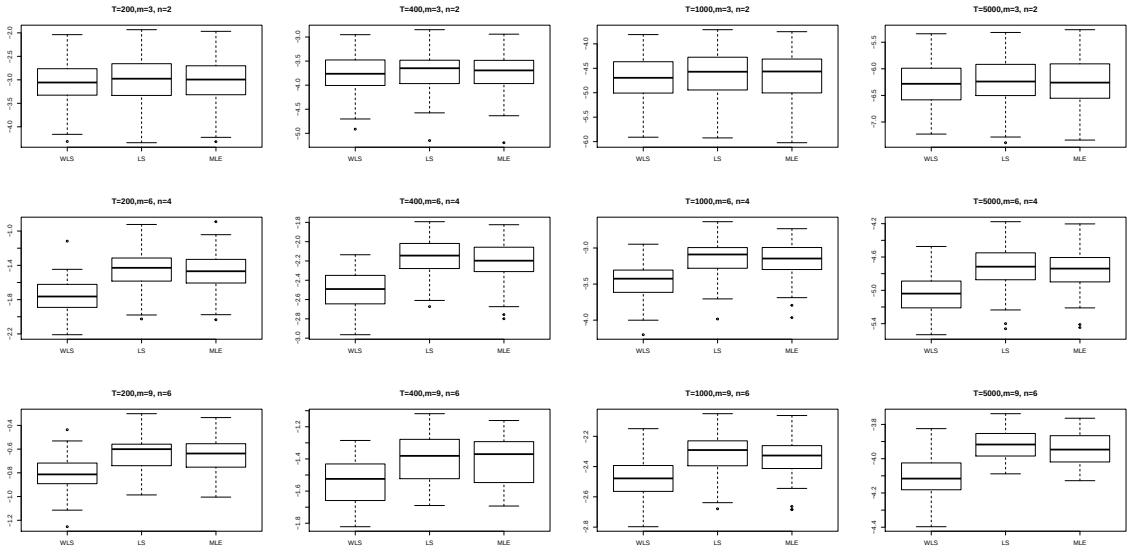

Figure 3.2: Comparison of three estimators, LS, WLS and MLE, under Setting II. The three rows correspond to $(m, n) = (3, 2)$, $(6, 4)$ and $(9, 6)$ respectively, and the four columns $T = 200$, 400, 1000 and 5000 respectively.

covariance structure have similar magnitude. Lastly, from Figure 3.2, we note that WLSE outperforms MLE when the mixture type is Σ_1 plus Σ_2 where Σ_1 is a randomly generated covariance matrix and Σ_2 is a randomly generated diagonal matrix.

Then we compare the asymptotic efficiencies of these three estimators by letting $T \to \infty$ under the five different settings. We fix the dimension pair $(m, n) = (3, 2)$ for all the settings under this experiment. We show the results in Figures 3.6 to 3.15. And we observe from these plots that the WLSE outperforms MLE and LS under Settings II, IV and V while the MLE outperforms WLSE and LSE under Settings III. All the estimators have similar performance under Setting I which is as expected.

Lastly, we perform an experiment to show the finite-sample performance of the asymptotic covariance matrix. We fix the dimension pair $(m, n) = (3, 2)$ as

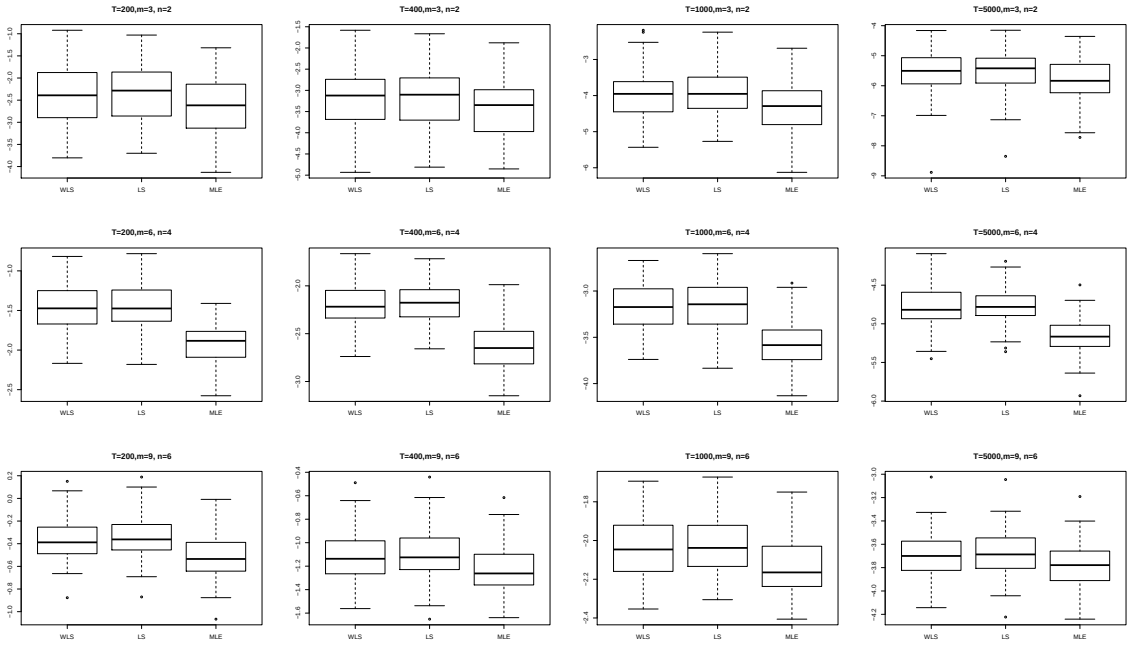

Figure 3.3: Comparison of three estimators, LS, WLS and MLE, under Setting III. The three rows correspond to $(m, n) = (3, 2)$, $(6, 4)$ and $(9, 6)$ respectively, and the four columns $T = 200$, 400, 1000 and 5000 respectively.

results for larger dimensions are similar. Under each of the five settings, we create 95% confidence intervals of all parameters based on the asymptotic normality distribution. Two types of confidence intervals are constructed: one for the entries of $\boldsymbol{A}$ and $\boldsymbol{B}$ separately, and the other for the entries of $\text{vec}(\boldsymbol{B}) \otimes \text{vec}(\boldsymbol{A})$. We replicate the experiment 1000 times for all the settings. Tables 3.1 to 3.10 shows the percentage that the true parameter falls within the marginal confidence interval of each parameter under for the three estimators under various scenarios. We observe that the coverage is quite good, especially when T is large.

3.4 Real Data Analysis

We apply our WLS approach on a time series data of four economic indicators (3-month interbank interest rate, GDP growth, total manufacture production

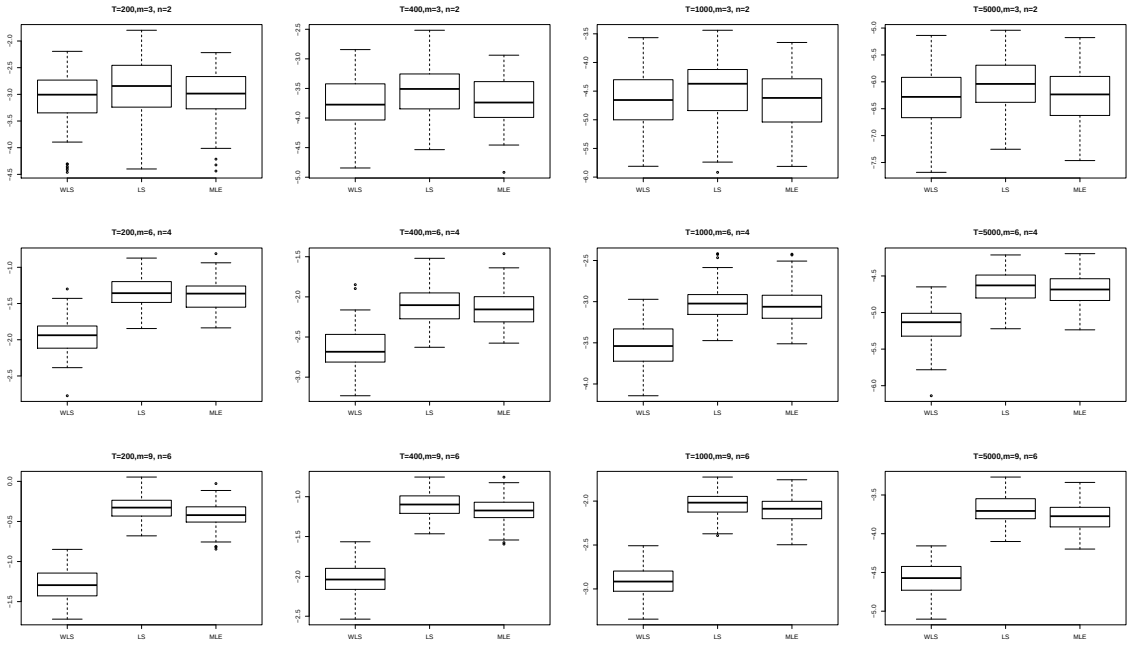

Figure 3.4: Comparison of three estimators, LS, WLS and MLE, under Setting IV. The three rows correspond to $(m, n) = (3, 2)$, $(6, 4)$ and $(9, 6)$ respectively, and the four columns $T = 200$, 400, 1000 and 5000 respectively.

growth and total consumer price index) from five countries (Canada, France, Germany, United Kingdom and United States) which contain observations in 4×5 matrix format. This dataset is presented in Section 5.2 of Chen et al. (2020) and downloaded from Organization for Economic Co-operation and Development (OECD) at https://data.oecd.org/. We fit a MAR(1) model to the data using four different estimation approaches: the three estimators proposed in Chen et al. (2020) (Projection, LSE and MLE) and the WLS estimator proposed in this Chapter. We also fit a stacked VAR(1) model and univariate AR(1) and AR(2) models for each individual time series. The residual sum of squares of each model and the sum of squares of the original data are reported in Table 3.11.

Table 3.12 and 3.13 show the estimated coefficients and their corresponding standard error (in the parentheses) of $\boldsymbol{A}$ and $\boldsymbol{B}$ using the WLS approach. The sign of significance for these coefficients are reported in Table 3.14 and 3.14.

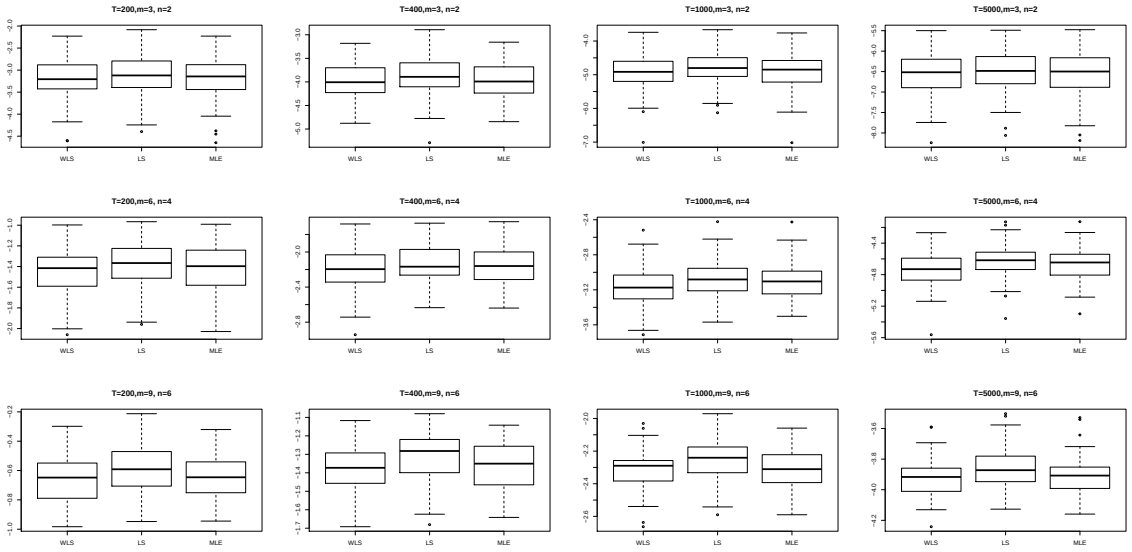

Figure 3.5: Comparison of three estimators, LS, WLS and MLE, under Setting V. The three rows correspond to $(m, n) = (3, 2)$, $(6, 4)$ and $(9, 6)$ respectively, and the four columns $T = 200$, 400, 1000 and 5000 respectively.

We also compare the out-of-sample rolling forecast performance between the four approaches of the MAR(1) model with the univariate AR(1) and AR(2) and the stacked VAR(1) model. We fit these models using all available data at time $t-1$ and obtained the one step ahead prediction $\hat{\boldsymbol{X}}_{t-1}(1)$ for $\boldsymbol{X}_t$ at time t. Table 3.16 shows the sum of prediction error squares $\|\hat{\boldsymbol{X}}_{t-1}(1) - \boldsymbol{X}_t\|_F^2$ of all methods.

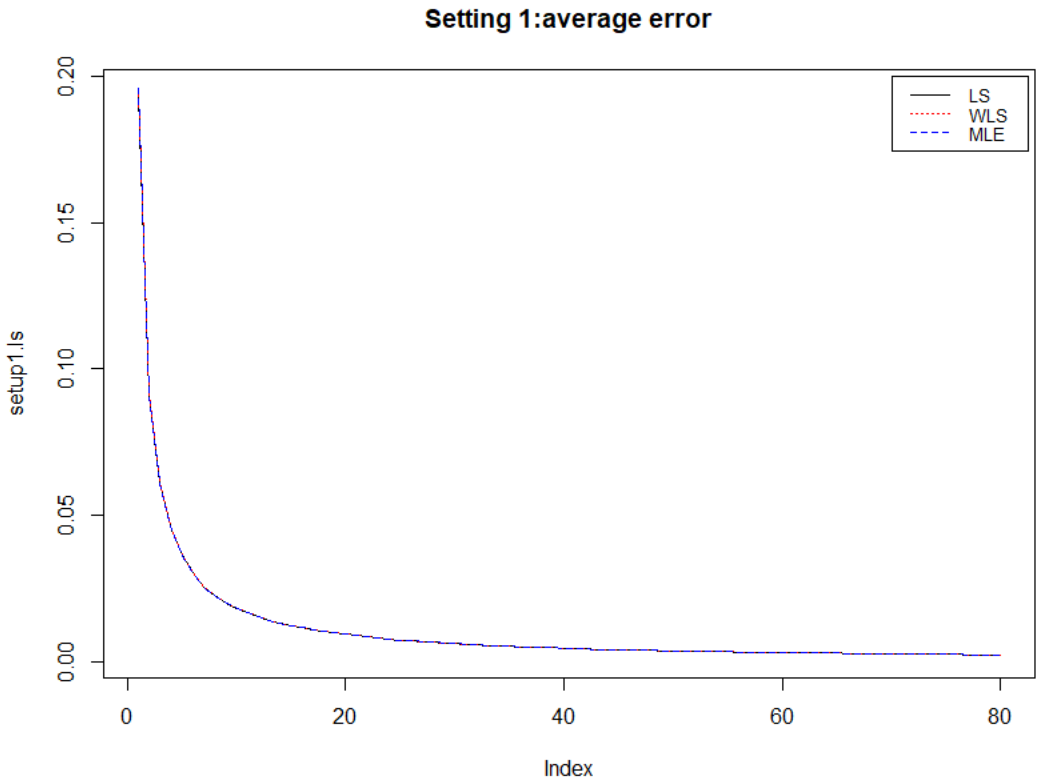

Figure 3.6: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting I (identity covariance structure), shows the average error over 100 repetitions for $\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9425	0.9443	0.9374
T=200	0.9466	0.9467	0.9451
T=500	0.9485	0.9494	0.9485
T=1000	0.9477	0.9485	0.9473

Table 3.1: Percentage of coverage of 95% confidence intervals for estimated $(\text{vec}'(\hat{\boldsymbol{A}}), \text{vec}'(\hat{\boldsymbol{B}}))'$ under setting I

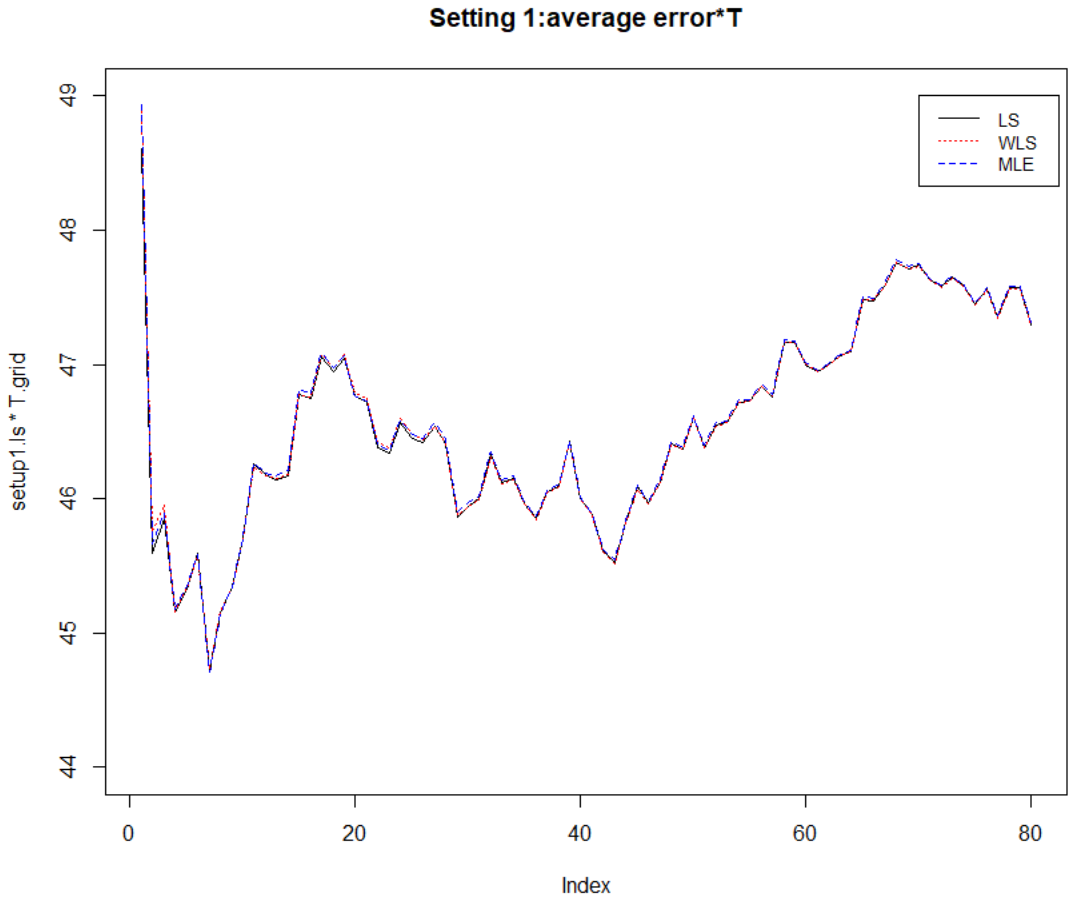

Figure 3.7: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting I (identity covariance structure), shows the average error over 100 repetitions for $T \times \|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9456	0.9478	0.9414
T=200	0.9451	0.9457	0.9431
T=500	0.9499	0.9503	0.9490
T=1000	0.9461	0.9464	0.9459

Table 3.2: Percentage of coverage of 95% confidence intervals for estimated $\text{vec}(\hat{\boldsymbol{B}}) \otimes \text{vec}(\hat{\boldsymbol{A}})$ under setting I

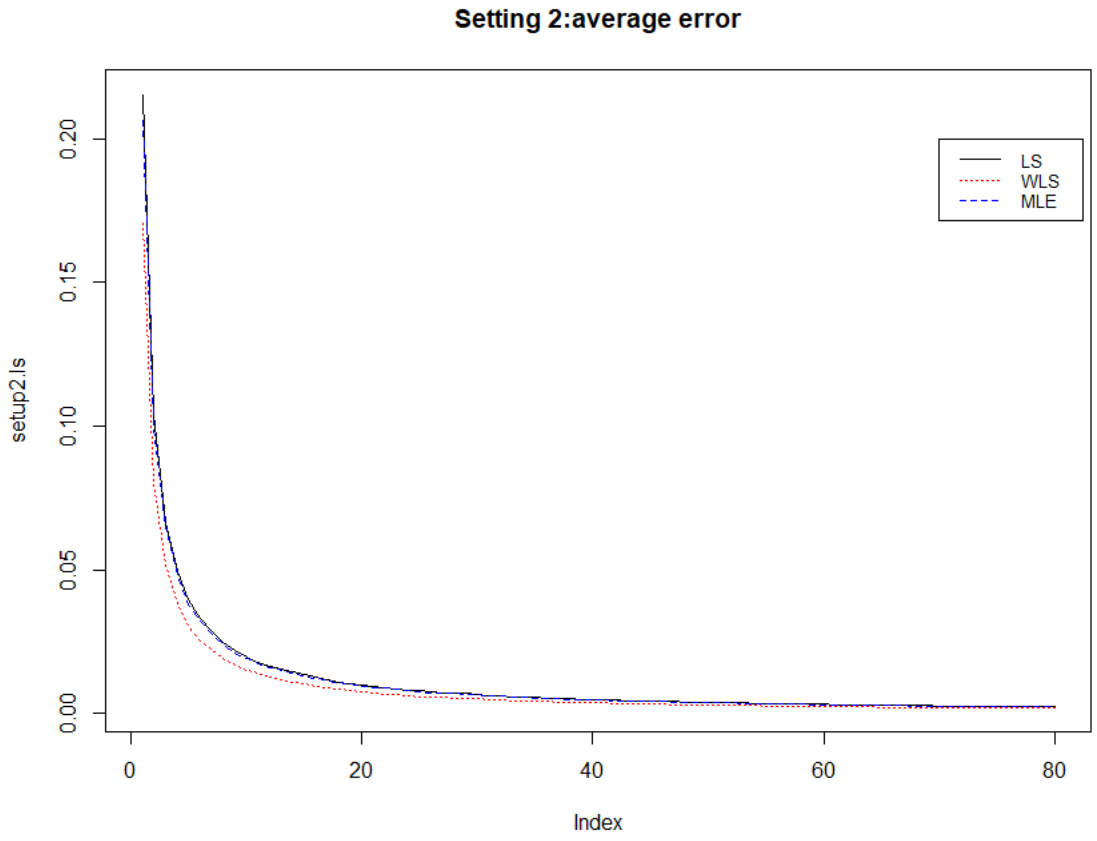

Figure 3.8: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting II (diagonal covariance structure + random covariance structure with eigenvalues generated from standard normal), shows the average error over 100 repetitions for $\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9404	0.9382	0.8986
T=200	0.9465	0.9478	0.9155
T=500	0.9457	0.9471	0.9158
T=1000	0.9532	0.9548	0.9261

Table 3.3: Percentage of coverage of 95% confidence intervals for estimated $(\text{vec}'(\hat{\boldsymbol{A}}), \text{vec}'(\hat{\boldsymbol{B}}))'$ under setting II

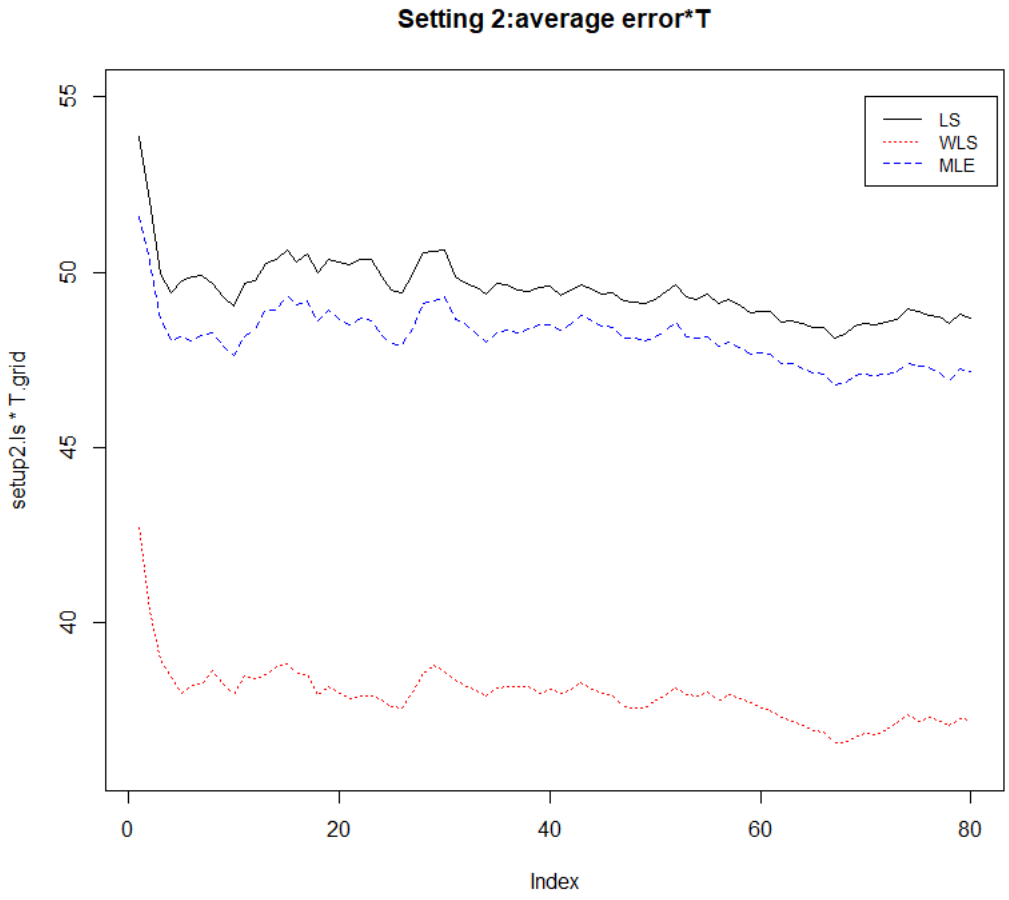

Figure 3.9: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting II (diagonal covariance structure + random covariance structure with eigenvalues generated from standard normal), shows the average error over 100 repetitions for $T \times \|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9411	0.9415	0.8964
T=200	0.9466	0.9500	0.9145
T=500	0.9477	0.9483	0.9133
T=1000	0.9526	0.9537	0.9235

Table 3.4: Percentage of coverage of 95% confidence intervals for estimated $\text{vec}(\hat{\boldsymbol{B}}) \otimes \text{vec}(\hat{\boldsymbol{A}})$ under setting II

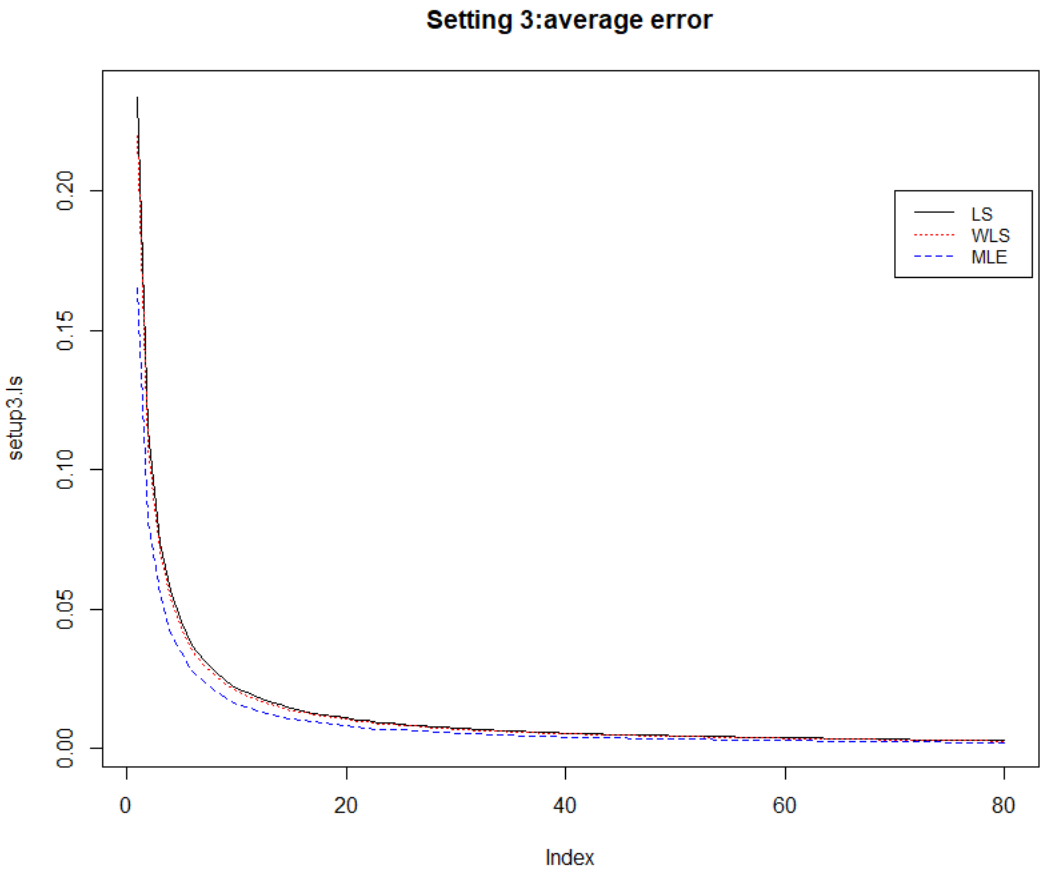

Figure 3.10: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting III (Kronecker covariance structure), shows the average error over 100 repetitions for $\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9395	0.9407	0.9339
T=200	0.9472	0.9480	0.9452
T=500	0.9505	0.9502	0.9488
T=1000	0.9483	0.9483	0.9485

Table 3.5: Percentage of coverage of 95% confidence intervals for estimated $(\text{vec}'(\hat{\boldsymbol{A}}), \text{vec}'(\hat{\boldsymbol{B}}))'$ under setting III

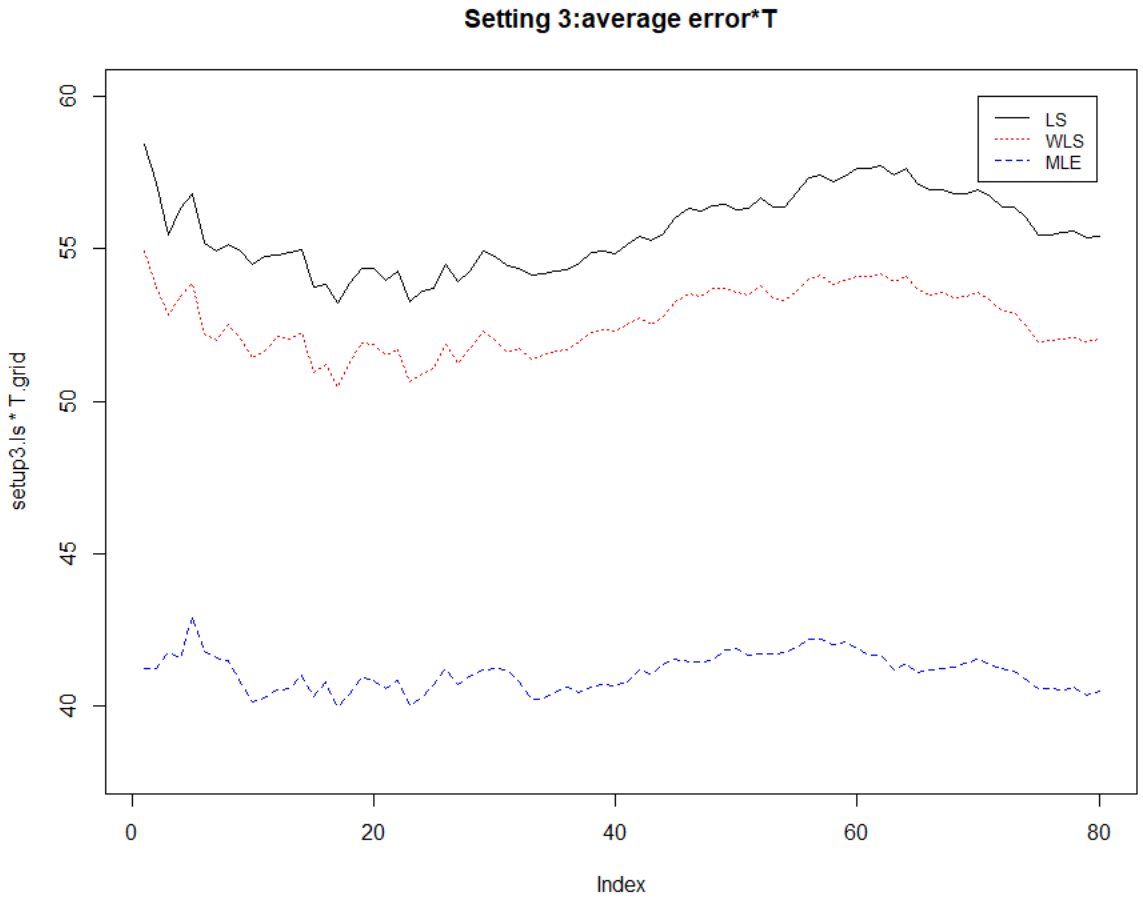

Figure 3.11: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting III (Kronecker covariance structure), shows the average error over 100 repetitions for $T \times \|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9427	0.9456	0.9385
T=200	0.9499	0.9502	0.9481
T=500	0.9483	0.9493	0.9483
T=1000	0.9473	0.9475	0.9469

Table 3.6: Percentage of coverage of 95% confidence intervals for estimated $\text{vec}(\hat{\boldsymbol{B}}) \otimes \text{vec}(\hat{\boldsymbol{A}})$ under setting III

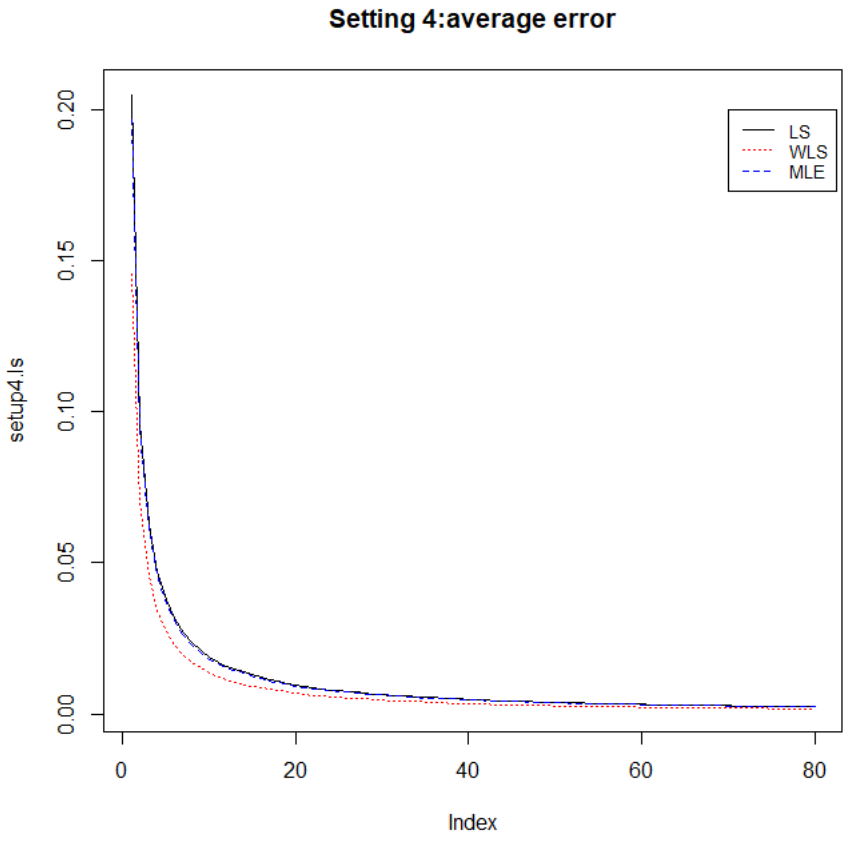

Figure 3.12: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting IV (diagonal covariance structure), shows the average error over 100 repetitions for $\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9403	0.9428	0.9074
T=200	0.9471	0.9473	0.9176
T=500	0.9522	0.9521	0.9245
T=1000	0.9486	0.9484	0.9227

Table 3.7: Percentage of coverage of 95% confidence intervals for estimated $(\text{vec}'(\hat{\boldsymbol{A}}), \text{vec}'(\hat{\boldsymbol{B}}))'$ under setting IV

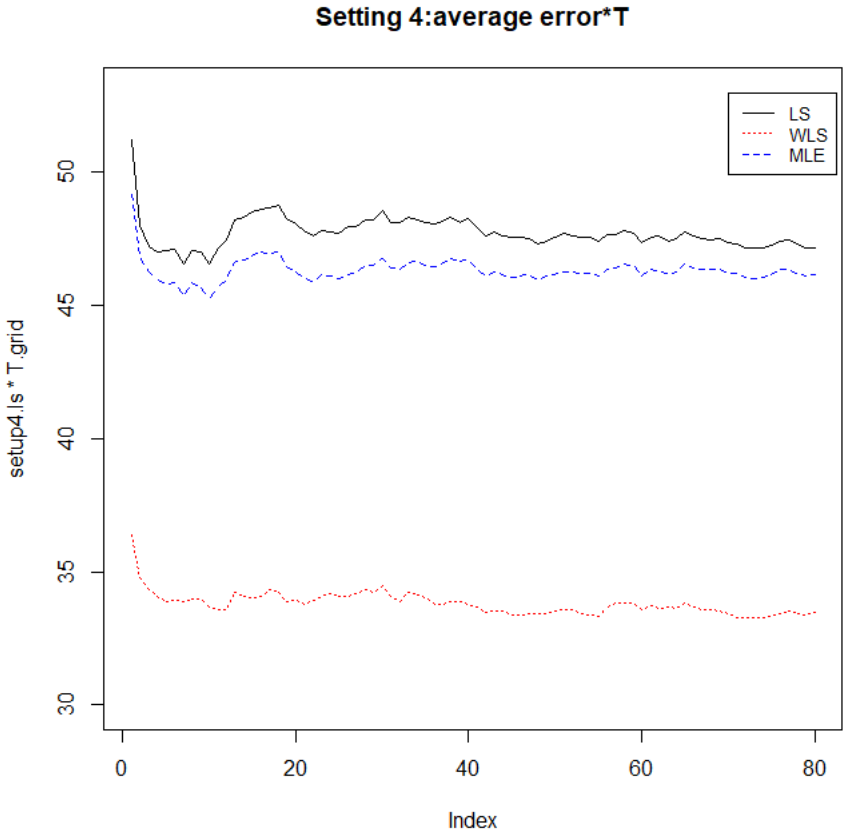

Figure 3.13: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting IV (diagonal covariance structure), shows the average error over 100 repetitions for $T \times \|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9423	0.9473	0.9083
T=200	0.9499	0.9512	0.9144
T=500	0.9512	0.9515	0.9177
T=1000	0.9498	0.9492	0.9179

Table 3.8: Percentage of coverage of 95% confidence intervals for estimated $\text{vec}(\hat{\boldsymbol{B}}) \otimes \text{vec}(\hat{\boldsymbol{A}})$ under setting IV

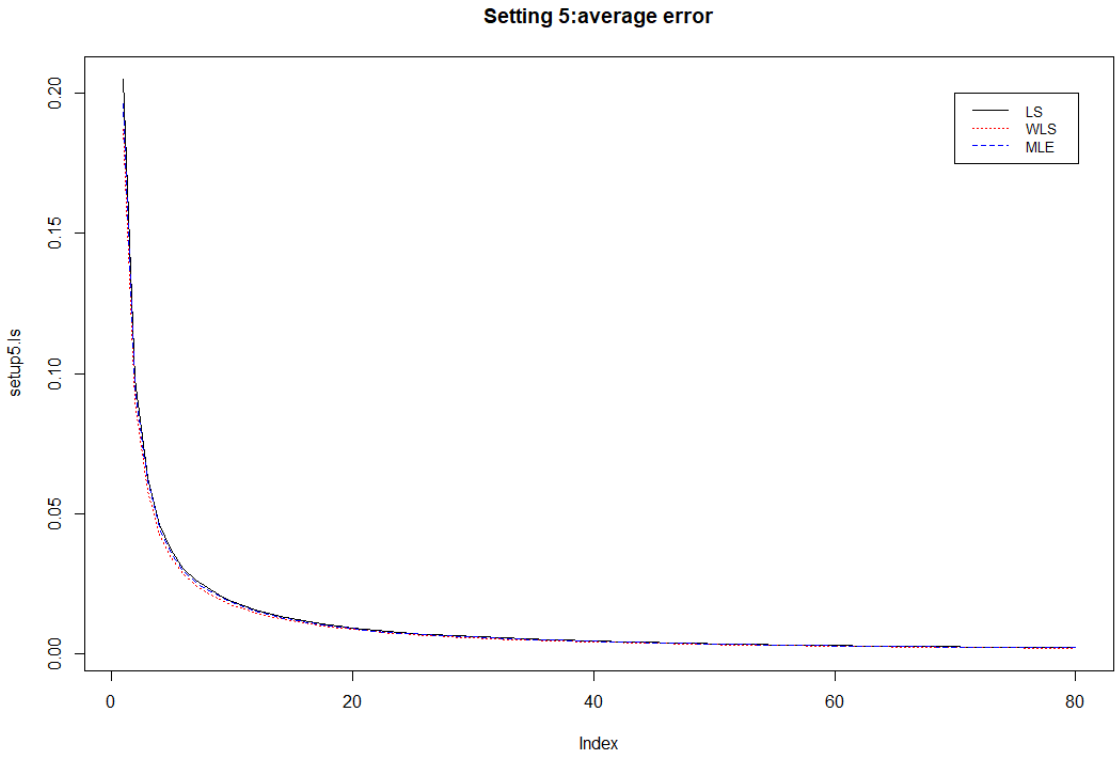

Figure 3.14: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting V (diagonal+Kronecker covariance structure), shows the average error over 100 repetitions for $\|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9411	0.9424	0.9278
T=200	0.9468	0.9470	0.9343
T=500	0.9487	0.9496	0.9409
T=1000	0.9475	0.9483	0.9398

Table 3.9: Percentage of coverage of 95% confidence intervals for estimated $(\text{vec}'(\hat{\boldsymbol{A}}), \text{vec}'(\hat{\boldsymbol{B}}))'$ under setting V

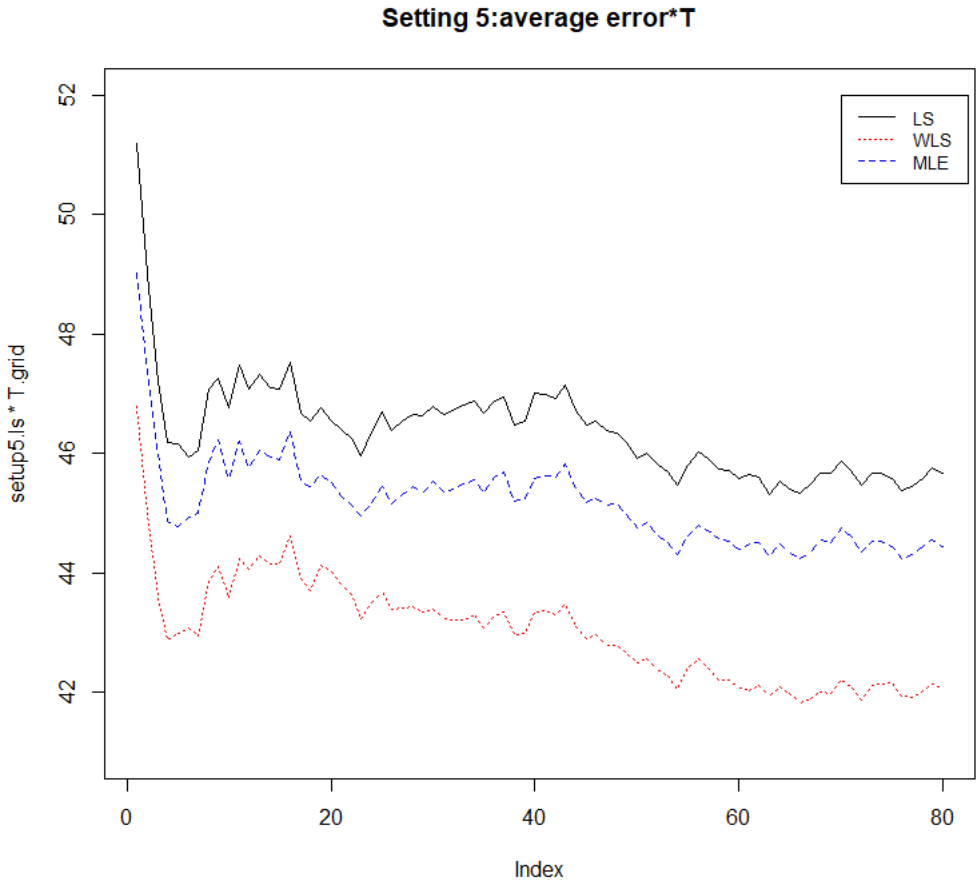

Figure 3.15: Comparison of asymptotic efficiencies of three estimators,LS,WLS,and MLE, under Setting V (diagonal+Kronecker covariance structure), shows the average error over 100 repetitions for $T \times \|\hat{\boldsymbol{B}} \otimes \hat{\boldsymbol{A}} - \boldsymbol{B} \otimes \boldsymbol{A}\|_F^2$.

	WLS	LS	MLE
T=100	0.9444	0.9468	0.9302
T=200	0.9453	0.9460	0.9335
T=500	0.9495	0.9504	0.9402
T=1000	0.9483	0.9492	0.9401

Table 3.10: Percentage of coverage of 95% confidence intervals for estimated $\text{vec}(\hat{\boldsymbol{B}}) \otimes \text{vec}(\hat{\boldsymbol{A}})$ under setting V

MAR(1).PROJ	MAR(1).LSE	MAR(1).MLE	MAR(1).WLS	VAR(1)	iAR(1)	iAR(2)	original
1959.26	1412.26	1436.83	1417.40	1076.27	1656.15	1539.44	2076.00

Table 3.11: Residual sum of squares of MAR(1) model using four different estimators and the stacked VAR(1) estimator, and the total residual sum of squares of fitting univariate AR(1) and AR(2) to each individual time series, and the total sum of squares of the original (normalized) data.

	Int	GDP	Prod	CPI
Int	0.26 (0.06)	0.28 (0.08)	0.06 (0.09)	0.06 (0.05)
GDP	-0.18 (0.05)	0.45 (0.07)	0.34 (0.09)	-0.07 (0.04)
Prod	-0.21 (0.05)	0.37 (0.07)	0.46 (0.09)	-0.01 (0.04)
CPI	-0.17 (0.07)	0.12 (0.10)	0.04 (0.12)	0.25 (0.06)

Table 3.12: Estimated left coefficient matrix $\boldsymbol{A}$ of MAR(1) using WLS method. Standard errors are shown in the parenthesis.

	USA	DEU	FRA	GBR	CAN
USA	0.76 (0.11)	-0.06 (0.14)	0.13 (0.11)	0.43 (0.11)	-0.06 (0.13)
DEU	0.34 (0.07)	0.13 (0.11)	0.61 (0.08)	0.48 (0.07)	-0.23 (0.09)
FRA	0.41 (0.10)	0.03 (0.15)	0.32 (0.12)	0.26 (0.11)	0.03 (0.14)
GBR	0.48 (0.10)	-0.07 (0.13)	0.10 (0.10)	0.57 (0.10)	-0.03 (0.12)
CAN	0.50 (0.08)	0.05 (0.11)	0.06 (0.09)	0.52 (0.08)	0.22 (0.10)

Table 3.13: Estimated right coefficient matrix $\boldsymbol{B}$ of MAR(1) using WLS method. Standard errors are shown in the parenthesis.

	Int	GDP	Prod	CPI
Int	+	+	0	0
GDP	−	+	+	0
Prod	−	+	+	0
CPI	−	0	0	+

Table 3.14: Sign of significance for the entries of matrix $\boldsymbol{A}$ at 5% level. The symbols $(+, -, 0)$ indicate positively significant, negatively significant and insignificant respectively.

	USA	DEU	FRA	GBR	CAN
USA	+	0	0	+	0
DEU	+	0	+	+	−
FRA	+	0	+	+	0
GBR	+	0	0	+	0
CAN	+	0	0	+	+

Table 3.15: Sign of significance for the entries of matrix $\boldsymbol{B}$ at 5% level. The symbols $(+, -, 0)$ indicate positively significant, negatively significant and insignificant respectively.

MAR(1).PROJ	MAR(1).LSE	MAR(1).MLE	MAR(1).WLS	VAR(1)	iAR(1)	iAR(2)
159.99	147.36	139.59	149.13	172.45	158.04	166.57

Table 3.16: Sum of out-of-sample prediction error squares of MAR(1) model using four different estimators and the stacked VAR(1) estimator, and the total sum of out-of-sample prediction error squares of fitting univariate AR(1) and AR(2) to each individual time series.

3.5 Proofs

Proof of Theorem 4. For the weighted least square estimators, applying Lemma 8 similarly as in the proof of Theorem 3 in Chen et al. (2020), we have

$$\hat{\boldsymbol{A}} = \boldsymbol{A} + O_p(T^{-1/2}), \quad \textit{and} \quad \hat{\boldsymbol{B}} = \boldsymbol{B} + O_p(T^{-1/2}).$$

Now repeating the gradient condition (3.5) in the sample fashion, we have:

$$\begin{aligned} \sum_{t=2}^{T}(\boldsymbol{W} \circ \boldsymbol{W} \circ (\hat{\boldsymbol{A}}\boldsymbol{X}_{t-1}\hat{\boldsymbol{B}}'))\hat{\boldsymbol{B}}\boldsymbol{X}_{t-1}' - \sum_{t=2}^{T}(\boldsymbol{W} \circ \boldsymbol{W} \circ \boldsymbol{X}_t)\hat{\boldsymbol{B}}\boldsymbol{X}_{t-1}' = 0 \\ \sum_{t=2}^{T}(\boldsymbol{W}' \circ \boldsymbol{W}' \circ (\hat{\boldsymbol{B}}\boldsymbol{X}_{t-1}'\hat{\boldsymbol{A}}'))\hat{\boldsymbol{A}}\boldsymbol{X}_{t-1} - \sum_{t=2}^{T}(\boldsymbol{W}' \circ \boldsymbol{W}' \circ \boldsymbol{X}_t')\hat{\boldsymbol{A}}\boldsymbol{X}_{t-1} = 0. \end{aligned} \tag{3.6}$$

Replacing each $\boldsymbol{X}_t$ by $\boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}' + \boldsymbol{E}_t$ in (3.6) and using the trick $\hat{\boldsymbol{A}}\boldsymbol{X}_{t-1}\hat{\boldsymbol{B}}' - \boldsymbol{A}\boldsymbol{X}_{t-1}\boldsymbol{B}' = (\hat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{X}_{t-1}\hat{\boldsymbol{B}}' + \boldsymbol{A}\boldsymbol{X}_{t-1}(\hat{\boldsymbol{B}}' - \boldsymbol{B}')$ to simply, we can obtain

$$\begin{aligned} &\sum_{t=2}^{T}(\boldsymbol{W} \circ \boldsymbol{W} \circ ((\hat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{X}_{t-1}\boldsymbol{B}'))\boldsymbol{B}\boldsymbol{X}_{t-1}' \\ &+ \sum_{t=2}^{T}(\boldsymbol{W} \circ \boldsymbol{W} \circ (\boldsymbol{A}\boldsymbol{X}_{t-1}(\hat{\boldsymbol{B}}' - \boldsymbol{B}')))\boldsymbol{B}\boldsymbol{X}_{t-1}' \\ &= \sum_{t=2}^{T}(\boldsymbol{W} \circ \boldsymbol{W} \circ \boldsymbol{E}_t)\boldsymbol{B}\boldsymbol{X}_{t-1}' + o_p(\sqrt{T}), \\ &\sum_{t=2}^{T}\boldsymbol{X}_{t-1}'\boldsymbol{A}'(\boldsymbol{W} \circ \boldsymbol{W} \circ ((\hat{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{X}_{t-1}\boldsymbol{B}')) \\ &+ \sum_{t=2}^{T}\boldsymbol{X}_{t-1}'\boldsymbol{A}'(\boldsymbol{W} \circ \boldsymbol{W} \circ (\boldsymbol{A}\boldsymbol{X}_{t-1}(\hat{\boldsymbol{B}}' - \boldsymbol{B}'))) \\ &= \sum_{t=2}^{T}\boldsymbol{X}_{t-1}'\boldsymbol{A}'(\boldsymbol{W} \circ \boldsymbol{W} \circ \boldsymbol{E}_t) + o_p(\sqrt{T}). \end{aligned}$$

Taking vectorization on both sides, we have

$$\begin{pmatrix} \sum_t((\boldsymbol{X}_{t-1}\boldsymbol{B}') \otimes \boldsymbol{I})\boldsymbol{M}((\boldsymbol{B}\boldsymbol{X}_{t-1}') \otimes \boldsymbol{I}) & \sum_t((\boldsymbol{X}_{t-1}\boldsymbol{B}') \otimes \boldsymbol{I})\boldsymbol{M}(\boldsymbol{I} \otimes (\boldsymbol{A}\boldsymbol{X}_{t-1})) \\ \sum_t(\boldsymbol{I} \otimes (\boldsymbol{X}_{t-1}'\boldsymbol{A}'))\boldsymbol{M}((\boldsymbol{B}\boldsymbol{X}_{t-1}') \otimes \boldsymbol{I}) & \sum_t(\boldsymbol{I} \otimes (\boldsymbol{X}_{t-1}'\boldsymbol{A}'))\boldsymbol{M}(\boldsymbol{I} \otimes (\boldsymbol{A}\boldsymbol{X}_{t-1})) \end{pmatrix}$$

$$\times \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \mathrm{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix}$$
$$= \sum_t \begin{pmatrix} ((\boldsymbol{X}_{t-1}\boldsymbol{B}') \otimes \boldsymbol{I})\boldsymbol{M} \\ (\boldsymbol{I} \otimes (\boldsymbol{X}'_{t-1}\boldsymbol{A}'))\boldsymbol{M} \end{pmatrix} \mathrm{vec}(\boldsymbol{E}_t) + o_p(\sqrt{T}),$$

which can be rewritten as

$$(\sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\boldsymbol{W}'_{t-1}) \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \mathrm{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix} = \sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\,\mathrm{vec}(\boldsymbol{E}_t) + o_p(\sqrt{T}). \quad (3.7)$$

Since $\boldsymbol{X}_t$ is strictly stationary with i.i.d. innovations, by the ergodic theorem, we have

$$\frac{1}{T}\sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\boldsymbol{W}'_{t-1} \to \mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\boldsymbol{W}_t) \quad a.s.$$

Note that here the matrix $\mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\boldsymbol{W}_t)$ is not full rank, but adding a term $\boldsymbol{\gamma}\boldsymbol{\gamma}'$ would make it full rank. On the other hand, since we require the Frobenius norm of $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$ to be 1, it holds that $\boldsymbol{\alpha}'(\mathrm{vec}(\hat{\boldsymbol{A}}) - \boldsymbol{\alpha}) = o_p(T^{-1/2})$. Hence we have

$$(\mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\boldsymbol{W}_t) + \boldsymbol{\gamma}\boldsymbol{\gamma}') \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \mathrm{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix} = \frac{1}{T}\sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\,\mathrm{vec}(\boldsymbol{E}_t) + o_p(T^{-1/2}). \quad (3.8)$$

By martingale CLT(see Hall and Heyde (2014)), the term on the RHS satisfies

$$\frac{1}{\sqrt{T}}\sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\,\mathrm{vec}(\boldsymbol{E}_t) \Rightarrow N(\boldsymbol{0}, \mathbb{E}(\boldsymbol{W}_t\boldsymbol{M}\Sigma\boldsymbol{M}\boldsymbol{W}'_t)).$$

So multiplying $\sqrt{T}$ on both sides of (3.8), we have

$$\sqrt{T}\boldsymbol{H} \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \mathrm{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix} = \frac{1}{\sqrt{T}}\sum_t \boldsymbol{W}_{t-1}\boldsymbol{M}\,\mathrm{vec}(\boldsymbol{E}_t) + o_p(1),$$

and it follows that

$$\sqrt{T} \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \mathrm{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix} \Rightarrow N(\boldsymbol{0}, \Xi_4).$$

For the last part, noting that

$$\begin{aligned}
&\operatorname{vec}(\hat{\boldsymbol{B}}') \otimes \operatorname{vec}(\hat{\boldsymbol{A}}) - \operatorname{vec}(\boldsymbol{B}') \otimes \operatorname{vec}(\boldsymbol{A}) \\
=&\operatorname{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \otimes \boldsymbol{\alpha} + \boldsymbol{\beta} \otimes \operatorname{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) + \operatorname{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \otimes \operatorname{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\
=&\operatorname{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \otimes \boldsymbol{\alpha} + \boldsymbol{\beta} \otimes \operatorname{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) + O_p(T^{-1}) \\
=&\boldsymbol{V} \begin{pmatrix} \operatorname{vec}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \\ \operatorname{vec}(\hat{\boldsymbol{B}}' - \boldsymbol{B}') \end{pmatrix} + o_p(T^{-1/2}),
\end{aligned}$$

where $\boldsymbol{V} \equiv [\boldsymbol{\beta} \otimes \boldsymbol{I}, \boldsymbol{I} \otimes \boldsymbol{\alpha}]$, the asymptotic statement follows and the proof is complete.

□

Proof of Corollary 1. This can be proved by applying Slusky's Theorem in the proof of Theorem 4. □

Chapter 4
Some Results on Constrained LASSO

4.1 Introduction

During the 1950s, Markowitz (1952) laid the foundation for the analysis of mean variance efficient portfolios. His original approach involves solving a quadratic programming problem with linear constraints.

Suppose we have p assets with return vector $\boldsymbol{R} \equiv (R_1, R_2, \ldots, R_p)' \in \mathbb{R}^p$ where R_i is the return of the i-th asset. Define

$$\boldsymbol{\mu} \equiv \mathbb{E}(\boldsymbol{R}), \Sigma \equiv \text{var}(\boldsymbol{R})$$

to be the mean and covariance structure of the return respectively and assume Σ to be nonsingular. Then a portfolio with weight vector $\boldsymbol{\beta} \in \mathbb{R}^p$ has the expected return $\boldsymbol{\beta}'\boldsymbol{\mu}$ and risk $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$.

Markowitz considers the problem of selecting the portfolio which maximizes the tradeoff of the expected return and risk. This can be formulated mathematically as follows.

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\beta}'\mathbf{1}=1, \boldsymbol{\beta}'\boldsymbol{\mu}=\tau}{\arg\min} \quad \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} \tag{4.1}$$

where τ is a given expected return.

Problem (4.1) can be solved directly using orthogonal space decomposition and the optimal solution $\boldsymbol{\beta}^{opt}$ is a linear combination of $\Sigma^{-1}\boldsymbol{\mu}$ and $\Sigma^{-1}\mathbf{1}$. This is usually referred as the classic mean-variance approach of Markowitz.

When trying to involve a large number of assets, it has been shown that the performance of the classic approach can be poor under general large dimensional setups due to the fact that the solution of the classical Markowitz problem depends sensitively on the input mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ while at the same time it is often a challenging task to estimate such a large dimensional covariance matrix consistently.

To alleviate these issues under large dimensional setup, it is natural to consider the regularized approach. For example, Brodie et al. (2009) modified the classic Markowitz problem (4.1) by adding a penalty proportional to the L_1-norm of the portfolio weight vector $\boldsymbol{\beta}$ to the objective function and the new problem is expressed mathematically as follows:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\beta}'\mathbf{1}=1,\boldsymbol{\beta}'\boldsymbol{\mu}=\tau}{\arg\min} \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \tag{4.2}$$

where λ is the tuning parameter. We call this L_1-norm $\|\boldsymbol{\beta}\|_1$ the *gross exposure* of the portfolio with weight vector $\boldsymbol{\beta}$. This concept of imposing gross exposure constraint was also considered and introduced independently in DeMiguel et al. (2009) and Fan et al. (2012c).

Another motivation for considering regularized optimization problems like (4.2) comes from classification. Suppose we have two p-dimensional normal distributions $N_p(\boldsymbol{\mu}_1, \Sigma)$ (with label variable $Y = 1$) and $N_p(\boldsymbol{\mu}_2, \Sigma)$ (with label variable $Y = 2$). Let $\boldsymbol{X}$ be a random vector which is drawn from either one of these two distributions with equal probabilities. We consider the problem of determining which class $\boldsymbol{X}$ is drawn.

For any linear discriminant classifier $\delta_{\boldsymbol{\beta}}(\boldsymbol{X}) = \mathcal{I}\{\boldsymbol{\beta}'(\boldsymbol{X} - \bar{\boldsymbol{\mu}}) > 0\}$, where $\bar{\boldsymbol{\mu}} \equiv \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a prescribed direction and $\mathcal{I}(.)$ is the indicator function with value 1 corresponds to assigning $\boldsymbol{X}$ to label $Y = 2$ and 0 corresponds to label $Y = 1$. The theoretical misclassification probability of this classifier with direction $\boldsymbol{\beta}$ is $1-\boldsymbol{\Phi}\{\boldsymbol{\beta}'\boldsymbol{\mu}_d/(\boldsymbol{\beta}'\Sigma\boldsymbol{\beta})^{1/2}\}$ where $\boldsymbol{\mu}_d = \frac{\boldsymbol{\mu}_2-\boldsymbol{\mu}_1}{2}$. To minimize this misclassification

probability, it is equivalent to minimize $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}$ subject to $\boldsymbol{\beta}'\boldsymbol{\mu}_d = 1$. And this essentially leads us to the Fisher discriminant with direction $\boldsymbol{\beta}_{Fisher} = \Sigma^{-1}\boldsymbol{\mu}_d$.

For real data applications, we can perform Fisher discriminant analysis using the sample version of the direction $\boldsymbol{\beta}_{Fisher}$ and the performance is usually good when p is small. But when the dimension p grows larger, it is well-known that simply perform the Fisher discriminant analysis might produce poor results. This again is due to the accumulation of noises when estimating Σ and $\boldsymbol{\mu}_d$. Like in the portfolio optimization context, here it is natural to consider regularizing the problem. And this leads us to the following regularized optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\beta}'\boldsymbol{\mu}_d=1}{\arg\min} \quad \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \tag{4.3}$$

where λ is the tuning parameter.

So we have seen that regularized optimization problems like (4.2) and (4.3) are important in the portfolio optimization and classification context. There are some approximate algorithms proposed to solve such constrained LASSO problems in literature, see for example Fan et al. (2012a) and James et al. (2012). But very few work has been done for solving these constrained LASSO problems exactly. In this chapter, we focus on problem (4.3) and propose an exact algorithm for solving this constrained LASSO problem. Furthermore, following the work of Mairal and Yu (2012), we also investigate the complexity of the constrained LASSO, and prove that the complexity is exponential in p.

The rest of this chapter is organized as follows. In Section 4.2, we study some properties of our constrained LASSO problem and propose an exact algorithm to compute its solution path based on these properties. In Section 4.3, we investigate the complexity nature of our problem and provide several plots for illustration purpose. The proof are relegated in Section 4.4.

4.2 Exact Algorithm for One Constraint LASSO problem

In this section, we first investigate the solution path of the following simpler regularized optimization problem

$$\boldsymbol{\beta}^{\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\beta}'\mathbf{1}=1}{\arg\min}\ \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \tag{4.4}$$

and extend our results to more general one constraint case later. In the process, we propose an algorithm to compute the whole solution path of this problem. We assume Σ to be a symmetric positive definite matrix throughout this chapter.

4.2.1 Basic Properties of the Problem

First we need the following result from convex analysis to establish the relationship between the tuning parameter λ and the optimal solution $\boldsymbol{\beta}^{\lambda}$:

Proposition 8. *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be convex. Then $\boldsymbol{x}$ minimizes f over a convex set $X \subset \mathbb{R}^p$ if and only if there exists a subgradient $\boldsymbol{d} \in \partial f(\boldsymbol{x})$ such that:*

$$\boldsymbol{d}'(\boldsymbol{z} - \boldsymbol{x}) \geq 0, \qquad \forall \boldsymbol{z} \in X.$$

With this proposition, we can prove the following result illustrating the basic properties of the optimization problem (4.4):

Proposition 9. *For the regularized optimization problem (4.4), the optimal solution $\boldsymbol{\beta}^{\lambda}$ is continuous in λ and the following equality*

$$2\Sigma\boldsymbol{\beta}^{\lambda} + \lambda\boldsymbol{d}^{\lambda} = c\mathbf{1} \tag{4.5}$$

holds for some constant c and subgradient $\boldsymbol{d}^{\lambda}$ of the L_1 norm function $\|.\|_1$ at $\boldsymbol{\beta}^{\lambda}$. Conversely, if equality (4.5) holds for some $\boldsymbol{\beta}$, $\boldsymbol{d} \in \partial\|\boldsymbol{\beta}\|_1$ and c at λ, then such $\boldsymbol{\beta}$ must coincide with the optimal solution $\boldsymbol{\beta}^{\lambda}$. Furthermore, when $\lambda > 0$, the subgradient $\boldsymbol{d}^{\lambda}$ which satisfies equality (4.5) is unique and continuous in λ.

Based on the above proposition, we can prove a stronger result for the optimization problem. The idea of the proof is based on an observation of the solution path between two point λ_1 and λ_2 where the optimal solution at these two points have the same sign pattern and is originally due to Mairal and Yu (2012).

We state the result in the following theorem.

Theorem 5. (Finite Piecewise Linearity of the Path) *Let* $\mathbf{1} \in \mathbb{R}^p$ *be the vector with all entries 1, and* Σ *be a positive definite matrix of dimension* $p \times p$. *Let*

$$\boldsymbol{\beta}^{\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}'\mathbf{1}=1}{\arg\min} \ \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

then $\boldsymbol{\beta}^{\lambda}$ *is a continuous piecewise linear function of finite many pieces in* λ.

4.2.2 The Exact Algorithm

With the finite piecewise linearity property established for the regularized optimization problem (4.4), we are ready to present an exact algorithm for computing its whole solution path. But first we need to introduce some conventions and assumptions for our discussion.

When the optimal solution $\boldsymbol{\beta}^{\lambda}$ evolves with λ, we say *the optimal solution component* β_k *hits zero at* λ if there exists a sequence $\{\lambda_n\}$ tending to λ from the **left** such that $\beta_k \neq 0$ at all λ_n while $\beta_k = 0$ at λ, and we say the *subgradient compoent* d_k *hits* $1(-1)$ *at* λ if there exists a sequence $\{\lambda_n\}$ tending to λ from the **left** such that $d_k \in (-1, 1)$ at all λ_n while $d_k = 1(-1)$ at λ. We say *the optimal solution component* β_k *leaves zero at* λ if there exists a sequence $\{\lambda'_n\}$ tending to λ from the **right** such that $\beta_k \neq 0$ at all λ'_n while $\beta_k = 0$ at λ, and we say the *subgradient compoent* d_k *leaves the boundary* ± 1 *at* λ for the similar meaning.

Then we define the notion of smoothness for λ in $[0, \lambda_{\max}]$. Note that here $\lambda_{\max} = \min\left\{\lambda | \boldsymbol{\beta}_j^{\lambda} \geqslant 0 \ \text{ for all } j\right\}$ and beyond this point(or say for $\lambda > \lambda_{\max}$) $\boldsymbol{\beta}^{\lambda}$ would no longer change.

Definition 2. *A point λ in $[0, \lambda_{\max}]$ is **smooth** if at this particular point, for any index j, either $\boldsymbol{\beta}_j^\lambda \neq 0$ or $\boldsymbol{\beta}_j^\lambda = 0$ but with corresponding subgradient $\boldsymbol{d}_j^\lambda$ lying in the open interval $(-1, 1)$.*

Lastly, we introduce two assumptions and make some remarks.

Assumption 1. When $\lambda = 0$, the initial optimal solution $\boldsymbol{\beta}^0$ has no zero component.

Remark: With this assumption, we can define $\boldsymbol{d}_j^\lambda$ to be $\text{sign}(\boldsymbol{\beta}_j^\lambda)$ for each j at $\lambda = 0$. We adopt this convention in the rest of this chapter.

Assumption 2. If$\lambda \in [0, \lambda_{\max}]$ is not a smooth point, then only one index j satisfies $\beta_j^\lambda = 0$ and $|\boldsymbol{d}_j^\lambda| = 1$.

Remark: Note that for continuous data, these two assumptions are usually satisfied.

With these preparations, now we can present a result which is not only interesting by itself but also helpful for understanding the motivation of our algorithm.

Proposition 10. *Under the above assumptions, for $\lambda \in (0, \lambda_{\max})$, if an optimal component β_k hits zero at λ(scenario 1), it would stay at zero in $[\lambda, \lambda + \epsilon)$ for some positive ϵ and $-1 < d_k < 1$ in the open interval $(\lambda, \lambda + \epsilon)$. Conversely, if a d_k hits 1 or -1 at λ(scenario 2), then there exists some positive ϵ such that $\beta_k \neq 0$ in the open interval $(\lambda, \lambda + \epsilon)$.*

Proof. For both scenarios, it is clear that λ is a non-smooth point. Then by Proposition 9 and our assumptions, there exists a small interval $\mathcal{B}_\lambda = (\lambda - \delta, \lambda + \delta)$ around λ, such that the conditions in the definition of smooth point are not violated for any index $j \neq k$.

Scenario 1:

We define the following set:

$$\mathcal{A} \triangleq \{k\} \cup \{j | \beta_j \neq 0 \text{ at } \lambda\}.$$

Then it is easy to see $\mathcal{A}^c = \{j|\beta_j = 0 \text{ but } d_j \text{ lies in } (-1,1)\}$.

Suppose the set $\mathcal{A}$ has cardinality q, Then WLOG we can partition Σ and $\boldsymbol{d}$ according to set $\mathcal{A}$ like:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}; \boldsymbol{d} = \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{z} \end{pmatrix}$$

where Σ_{11} and $\boldsymbol{s}$ correspond to $\mathcal{A}$, Σ_{22} and $\boldsymbol{z}$ correspond to $\mathcal{A}^c$.

Recall how we define the sequence $\{\lambda_n\}$ when we make conventions of "β_k *hits zero*" and "d_k *hits* $1(-1)$" in the previous section. WLOG, we can assume β_k is positive at all λ_n and $\{\lambda_n\} \subset \mathcal{B}_\lambda$.

Then we argue by contradiction. If the corresponding d_k would **not** be away from the boundary 1 and -1 in an open interval $(\lambda, \lambda + \epsilon)$ for some $\epsilon > 0$, then there exists a $\kappa \in \mathcal{B}_\lambda$ and $\kappa > \lambda$ such that d_k remains at 1 at κ.

Now for the tuning parameter, at λ, we have

$$2\Sigma\boldsymbol{\beta} + \lambda \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{z} \end{pmatrix} = c\mathbf{1}, \tag{4.6}$$

and for any $\Delta\lambda$ such that $\lambda + \Delta\lambda$ equals to κ or any λ_n, we would have

$$2\Sigma(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) + (\lambda + \Delta\lambda) \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{z} + \Delta\boldsymbol{z} \end{pmatrix} = (c + \Delta c)\mathbf{1}, \tag{4.7}$$

where $\Delta\boldsymbol{\beta} = (\Delta\boldsymbol{\beta}^1, \Delta\boldsymbol{\beta}^2)'$ and $\Delta\boldsymbol{\beta}^2 = \mathbf{0}$ corresponds to the update of the set $\mathcal{A}^c$.

Then we take the difference of the two equalities (4.7) and (4.6), and simplify the first q and last $p - q$ equations separately. This leads to the following two equalities:

$$2\Sigma_{11}\Delta\boldsymbol{\beta}^1 + \Delta\lambda\boldsymbol{s} = \Delta c\mathbf{1}_q,$$

$$2\Sigma_{21}\Delta\boldsymbol{\beta}^1 + \Delta\lambda\Delta\boldsymbol{z} + \Delta\lambda\boldsymbol{z} + \lambda\Delta\boldsymbol{z} = \Delta c\mathbf{1}_{p-q}.$$

Using the fact that $\mathbf{1}_q'\Delta\boldsymbol{\beta}^1 = 0$, we can solve

$$\Delta\boldsymbol{\beta}^1 = \frac{1}{2}\Delta\lambda\boldsymbol{t}, \tag{4.8}$$

$$\Delta \boldsymbol{z} = \frac{\Delta\lambda \boldsymbol{l}}{\Delta\lambda + \lambda}, \tag{4.9}$$

Here

$$\boldsymbol{t} = \frac{\alpha}{\gamma}\boldsymbol{b} - \boldsymbol{a}$$

and

$$\boldsymbol{l} = \frac{\alpha}{\gamma}\mathbf{1}_{p-q} - \Sigma_{21}\boldsymbol{t} - \boldsymbol{z}$$

where $\alpha = \mathbf{1}_q'\Sigma_{11}^{-1}\boldsymbol{s}$, $\gamma = \mathbf{1}_q'\Sigma_{11}^{-1}\mathbf{1}_q$, $\boldsymbol{a} = \Sigma_{11}^{-1}\boldsymbol{s}$, $\boldsymbol{b} = \Sigma_{11}^{-1}\mathbf{1}_q$.

Note that we always have $\text{sign}(\beta_k) = -\text{sign}(t_k) \neq 0$ at $\{\lambda_n\}$, where t_k is the k-th component of the fixed vector $\boldsymbol{t}$ in equation (4.8). This would imply $\text{sign}(\beta_k) = \text{sign}(t_k) \neq 0$ at κ which is impossible since we can choose κ to be as close to λ as possible and β_k can not have a sudden sign change. Thus we have proved the first part of the proposition.

Scenario 2:

We still argue by contradiction so that assume there exists a $\kappa \in \mathcal{B}_\lambda$ and $\kappa > \lambda$ such that β_k remains at 0 at κ. In this scenario we define set $\mathcal{A}$ as follows:

$$\mathcal{A} \triangleq \{j | \beta_j \neq 0 \text{ at } \lambda\}.$$

Then under this scenario, it is easy to see that the compliment set

$$\mathcal{A}^c = \{k\} \cup \{j | \beta_j = 0 \text{ but } d_j \text{ lies in } (-1, 1)\}.$$

Now this time we partition the matrix and vectors according to the new $\mathcal{A}$ and apply the same arguements as in the previous scenario, this time we look at the equation (4.9). This time we always have $\text{sign}(\Delta \boldsymbol{z}_k) = -\text{sign}(l_k) \neq 0$ at $\{\lambda_n\}$, where l_k is the k-th component of the fixed vector $\boldsymbol{l}$ in equation (4.9). This implies $\text{sign}(\Delta \boldsymbol{z}_k) = \text{sign}(l_k) \neq 0$ at κ which is impossible since this would force d_k to cross the subgradient boundary value at λ.

□

We will continue to use the notations $\boldsymbol{t}$, $\boldsymbol{l}$, α, γ, $\boldsymbol{a}$, $\boldsymbol{b}$ which appear in the proof of Proposition 10 throughout the rest of this chapter.

Under the assumptions we have, we note that the occurrence of the scenario 2 is equivalent to say β_k leaves 0 at λ. Then it is not hard to see that the two scenarios mentioned in Proposition 10 **can not** happen together at a point $\lambda \in [0, \lambda_{\max}]$ and any non-smooth point in $[0, \lambda_{\max}]$ would correpond to the occurrence of one the scenarios. The occurrences of the scenarios through out $[0, \lambda_{\max}]$ must be finite by Theorem 5 thus there are finite numbers of non-smooth point in $[0, \lambda_{\max}]$.

We futher note that the equations (4.6) and (4.7) would be valid for a partition via a set $\mathcal{A}$ as long as the subgradient components $\boldsymbol{s}$ which corresponds to $\mathcal{A}$ is stable at λ and $\lambda + \Delta\lambda$ while the update $\Delta\boldsymbol{\beta}^2$ which corresponds to $\mathcal{A}^\rfloor$ equals to zero.

Now we start from $\lambda = 0$ and gradually increase it to $\lambda_{\max}$ for our explanation of the algorithm. At $\lambda_0 = 0$, we can solve the optimal solution $\boldsymbol{\beta}^0$ explicitly and from the assumptions the next occurrence of the scenarios(or say the next non-smooth point) must be scenario 1. The question is how to find the step size $\Delta\lambda$ to achieve the occurrence. We start with a set $\mathcal{A}_0 = \{1, 2, \ldots, p\}$ for the partition. Since the equations (4.6), (4.7) will hold continuously when we increase $\Delta\lambda$ from λ untill we reach the next non-smooth point, reaching the next non-smooth point would essentially mean one of the equations among (4.8) and (4.9) has reach the limit that beyond that point it can not be true(For (4.8) it means letting a β_k accross 0 and for (4.9) it means letting a z_k accross the boundary so be outside $[-1, 1]$) and vice versa. So at the current stage $\lambda_0 = 0$, we only need to find the smallest $\Delta\lambda$ which makes a nonzero β_j^λ hit zero via equation (4.8). Or more precisely find

$$\lambda_1 \triangleq \lambda_0 + \min_j \left\{ \frac{2\beta_j^0}{-t_j} | j \,\text{such that}\, \operatorname{sgn}(\beta_j^0) = -\operatorname{sgn}(t_j) \right\}$$

and the index m which achieves the minimum in the above equation. The corresponding $\boldsymbol{\beta}^1$, $\boldsymbol{d}^1$ can be updated via (4.8) and (4.9) easily.

So now we are at λ_1, the question would be what set $\mathcal{A}$ shall we use to continue the search for the next non-smooth point. From Proposition 10, we set $\mathcal{A}_1$=$\mathcal{A}_0/\{m\}$, it would work for our purpose and from the new set $\mathcal{A}$ we shall update $\boldsymbol{t}$, $\boldsymbol{l}$, α, γ, $\boldsymbol{a}$, $\boldsymbol{b}$ correspondingly. To search the next non-smooth point this time we also need to consider the possibility of scenario 2, thus we need to utilize (4.8) and (4.9) together to find the smallest step size $\Delta\lambda$ hitting the next non-smooth point. So this time we find

$$\lambda_2 \triangleq \lambda_1 + \tau$$

where

$$\tau = \min\{\tau_1, \tau_2\}$$

Here

$$\tau_1 \triangleq \min_j \left\{-2\beta_j^1/t_j | j \in \mathcal{A}_1 \text{ and } \operatorname{sgn}(\beta_j^1) = -\operatorname{sgn}(t_j)\right\}$$

$$\tau_2 \triangleq \min_{i,j} \left\{ \frac{\lambda_1}{\frac{l_i}{1-z_i} - 1}, \frac{\lambda_1}{\frac{l_j}{1+z_j} - 1} | i \in \mathcal{C}_1, j \in \mathcal{C}_2 \right\}$$

$$\mathcal{C}_1 \triangleq \{i | i \in \mathcal{A}_1^c \text{ and } l_i > 1 - z_i\}$$

$$\mathcal{C}_2 \triangleq \{j | j \in \mathcal{A}_1^c \text{ and } l_j < -(1 + z_j)\}$$

Also we here find the index m which achieves the minimum τ.

Similarly we can update $\boldsymbol{\beta}^2$, $\boldsymbol{d}^2$. The question remaining is how to update the $\mathcal{A}$ for the next incremental search. Again from Proposition 10, what we need to do is as follows. For the current index m, if scenario 1 happens(nonzero β_j^1 hit zero), then update $\mathcal{A}_2 = \mathcal{A}_1/\{m\}$, otherwise update $\mathcal{A}_2 = \mathcal{A}_1 \cup \{m\}$.

Now we can repeat the above procedures iteratively and by Theorem 5 and Proposition 9 we know it would terminate in finitely many steps. From λ_k, we find

$$\lambda_{k+1} \triangleq \lambda_k + \tau$$

where

$$\tau = \min\{\tau_1, \tau_2\}$$

$$\tau_1 \triangleq \min_j \left\{ -2\beta_j^k / t_j | j \in \mathcal{A}_k \text{ and } \operatorname{sgn}(\beta_j^k) = -\operatorname{sgn}(t_j) \right\}$$

$$\tau_2 \triangleq \min_{i,j} \left\{ \frac{\lambda_k}{\frac{l_i}{1-z_i} - 1}, \frac{\lambda_k}{\frac{l_j}{1+z_j} - 1} | i \in \mathcal{C}_1, j \in \mathcal{C}_2 \right\}$$

$$\mathcal{C}_1 \triangleq \{ i | i \in \mathcal{A}_k^c \text{ and } l_i > 1 - z_i \}$$

$$\mathcal{C}_2 \triangleq \{ j | j \in \mathcal{A}_k^c \text{ and } l_j < -(1 + z_j) \}.$$

Also we can find the index m for the current stage and update $\mathcal{A}_{k+1}$ according to the rule: if scenario 1 happens (nonzero β_j^k hit zero), then update $\mathcal{A}_{k+1} = \mathcal{A}_k / \{m\}$, otherwise update $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{m\}$. The $\boldsymbol{\beta}^{k+1}$, $\boldsymbol{d}^{k+1}$ is again updated via (4.8) and (4.9) from the previous $\boldsymbol{\beta}^k$, $\boldsymbol{d}^k$ with the τ calculated at the current stage. And the $\boldsymbol{t}$, $\boldsymbol{l}$, α, γ, $\boldsymbol{a}$, $\boldsymbol{b}$ for the current stage could be calculated since we have obtained the partition set $\mathcal{A}_{k+1}$.

With the above explanation, we summarize our algorithm for the regularized optimization problem (4.4) as follows.

Step 1. Solve the initial optimal solution $\boldsymbol{\beta}^0$ when $\lambda_0 = 0$. Set the initial active set $\mathcal{A}_0 = \{1, 2, \ldots, p\}$.

Step 2. Suppose we have already obtain $\boldsymbol{\beta}^k$, $\boldsymbol{d}^k$, λ_k , $\mathcal{A}_k$.

Compute $\alpha = \mathbf{1}_q' \Sigma_{11}^{-1} \boldsymbol{s}$, $\gamma = \mathbf{1}_q' \Sigma_{11}^{-1} \mathbf{1}_q$, $\boldsymbol{a} = \Sigma_{11}^{-1} \boldsymbol{s}$, $\boldsymbol{b} = \Sigma_{11}^{-1} \mathbf{1}_q$, $\boldsymbol{t} = \frac{\alpha}{\gamma} \boldsymbol{b} - \boldsymbol{a}$, $\boldsymbol{l} = \frac{\alpha}{\gamma} \mathbf{1}_{p-q} - \Sigma_{21} \boldsymbol{t} - \boldsymbol{z}$

with the partition set $\mathcal{A}_k$ and current $\boldsymbol{s}$ and $\boldsymbol{z}$ in $\boldsymbol{d}^k$.

Compute $\tau = \min\{\tau_1, \tau_2\}$ and denote m to be the index achieving such τ.

Update $\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \frac{1}{2} \tau \boldsymbol{t}$, $\boldsymbol{d}^{k+1} = \boldsymbol{d}^k + \Delta \boldsymbol{z}$, here $\Delta z_j = \frac{\tau \boldsymbol{l}}{\tau + \lambda_k}$ for $j \in \mathcal{A}_k^c$, $\Delta z_j = 0$ otherwise.

For the partition set, if for index m, scenario 1 happens, update $\mathcal{A}_{k+1} = \mathcal{A}_k / \{m\}$, otherwise update $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{m\}$.

Step 3. Repeat step 2 untill it reaches $\lambda_k = \lambda_{\max}$ where all $\beta_j \geqslant 0$.

4.2.3 The Weighted Case

In some circumstances, we might be more interested in the following weighted problem:

$$\boldsymbol{\beta}^{\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\beta}'\mathbf{1}=1}{\arg\min}\ \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\sum_{i=1}^{p}\sigma_{ii}^{1/2}|\beta_i|$$

where σ_{ii} is the i-th diagonal element of the matrix Σ.

This problem could be transformed into the form:

$$\boldsymbol{\beta}^{\lambda}_{\text{new}} = \underset{\boldsymbol{\beta}_{\mathbf{new}}\in\mathbb{R}^p, \boldsymbol{\beta}'_{\text{new}}\boldsymbol{u}=1}{\arg\min}\ \boldsymbol{\beta}'_{\text{new}}R\boldsymbol{\beta}_{\text{new}} + \lambda\|\boldsymbol{\beta}_{\text{new}}\|_1$$

via a change of variable $\boldsymbol{\beta}_{\text{new}} = \text{diag}\{\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2}\}\boldsymbol{\beta}$. Here

$$R = \text{diag}\{\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2}\}^{-1}\Sigma\,\text{diag}\{\sigma_{11}^{1/2}, \ldots, \sigma_{pp}^{1/2}\}^{-1}, \boldsymbol{u} = (\sigma_{11}^{-1/2}, \ldots, \sigma_{pp}^{-1/2})'.$$

This leads us to study the slightly more general case:

$$\boldsymbol{\beta}^{\lambda} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\beta}'\boldsymbol{u}=1}{\arg\min}\ \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 \tag{4.10}$$

where Σ is a symmetric positive definite matrix and $\boldsymbol{u}$ is a general nonzero vector in $\mathbb{R}^p$.

Like previous section, we can establish similar results for problem (4.10) and the proof is essentially the same. These results are stated as follows:

Proposition 11. *For the regularized optimization problem (4.10), the optimal solution $\boldsymbol{\beta}^{\lambda}$ is continuous in λ and the following equality*

$$2\Sigma\boldsymbol{\beta}^{\lambda} + \lambda\boldsymbol{d}^{\lambda} = c\boldsymbol{u} \tag{4.11}$$

holds for some constant c and subgradient $\boldsymbol{d}^{\lambda}$ of the L_1 norm function $\|.\|_1$ at $\boldsymbol{\beta}^{\lambda}$. Conversely, if equality (4.11) holds for some $\boldsymbol{\beta}$, $\boldsymbol{d} \in \partial\|\boldsymbol{\beta}\|_1$ and c at λ, then such $\boldsymbol{\beta}$ must coincide with the optimal solution $\boldsymbol{\beta}^{\lambda}$. Furthermore, when $\lambda > 0$, the subgradient $\boldsymbol{d}^{\lambda}$ which satisfies equality (4.11) is unique and continuous in λ.

Theorem 6. (Finite Piecewise Linearity of the Path) *Let $\boldsymbol{u} \in \mathbb{R}^p$ be a nonzero vector, and Σ be a positive definite matrix of dimension $p \times p$. Let*

$$\boldsymbol{\beta}^{\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}'\boldsymbol{u}=1}{\arg\min} \ \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

then $\boldsymbol{\beta}^{\lambda}$ is a continuous piecewise linear function of finite many pieces in λ.

Assume the two assumptions in the previous section, we note that Proposition 10 also holds for problem (4.10). Using the fact that $\boldsymbol{u}'_{\mathcal{A}}\Delta\boldsymbol{\beta}^1 = 0$ ($\mathcal{A}$ is the partition set at point λ, as defined in the proof of Proposition 10), we obtain the updating equalities for problem (4.10):

$$\Delta\boldsymbol{\beta}^1 = \frac{1}{2}\Delta\lambda\boldsymbol{t}, \tag{4.12}$$

$$\Delta\boldsymbol{z} = \frac{\Delta\lambda\boldsymbol{l}}{\Delta\lambda + \lambda}, \tag{4.13}$$

Here

$$\boldsymbol{t} = \frac{\alpha}{\gamma}\boldsymbol{b} - \boldsymbol{a}$$

and

$$\boldsymbol{l} = \frac{\alpha}{\gamma}\boldsymbol{u}_{\mathcal{A}^c} - \Sigma_{21}\boldsymbol{t} - \boldsymbol{z}$$

where $\alpha = \boldsymbol{u}'_{\mathcal{A}}\Sigma_{11}^{-1}\boldsymbol{s}$, $\gamma = \boldsymbol{u}'_{\mathcal{A}}\Sigma_{11}^{-1}\boldsymbol{u}_{\mathcal{A}}$, $\boldsymbol{a} = \Sigma_{11}^{-1}\boldsymbol{s}$, $\boldsymbol{b} = \Sigma_{11}^{-1}\boldsymbol{u}_{\mathcal{A}}$.

Then the algorithm we summarized in the previous section can be implemented to solve optimization problem (4.10) with the newly defined vector $\boldsymbol{t}$ and $\boldsymbol{l}$ in equations (4.12) and (4.13). Here we need to be careful about the stopping rule since γ might be zero with a zero $\boldsymbol{u}_{\mathcal{A}}$. But for our weighted case, as all the entries in $\boldsymbol{u}$ are positive, the stopping rule would be the same as problem (4.4). So we adopt all the definitions in the previous section and summarize the algorithm for the regularized problem (4.10) as follows.

Step 1. Solve the initial optimal solution $\boldsymbol{\beta}^0$ when $\lambda_0 = 0$. Set the initial active set $\mathcal{A}_0 = \{1, 2, \ldots, p\}$.

Step 2. Suppose we have already obtain $\boldsymbol{\beta}^k$, $\boldsymbol{d}^k$, λ_k , $\mathcal{A}_k$.

Compute $\alpha = \boldsymbol{u}'_{\mathcal{A}}\Sigma_{11}^{-1}\boldsymbol{s}$, $\gamma = \boldsymbol{u}'_{\mathcal{A}}\Sigma_{11}^{-1}\boldsymbol{u}_{\mathcal{A}}$, $\boldsymbol{a} = \Sigma_{11}^{-1}\boldsymbol{s}$, $\boldsymbol{b} = \Sigma_{11}^{-1}\boldsymbol{u}_{\mathcal{A}}$, $\boldsymbol{t} = \frac{\alpha}{\gamma}\boldsymbol{b} - \boldsymbol{a}$, $\boldsymbol{l} = \frac{\alpha}{\gamma}\boldsymbol{u}_{\mathcal{A}^c} - \Sigma_{21}\boldsymbol{t} - \boldsymbol{z}$

with the partition set $\mathcal{A}_k$ and current $\boldsymbol{s}$ and $\boldsymbol{z}$ in $\boldsymbol{d}^k$.

Compute $\tau = \min\{\tau_1, \tau_2\}$ and denote m to be the index achieving such τ.

Update $\boldsymbol{\beta}^{k+1}$=$\boldsymbol{\beta}^k$+$\frac{1}{2}\tau\boldsymbol{t}$, $\boldsymbol{d}^{k+1} = \boldsymbol{d}^k + \Delta\boldsymbol{z}$, here $\Delta z_j = \frac{\tau \boldsymbol{l}}{\tau+\lambda_k}$ for $j \in \mathcal{A}_k^c$, $\Delta z_j = 0$ otherwise.

For the partition set, if for index m, scenario 1 happens, update $\mathcal{A}_{k+1} = \mathcal{A}_k/\{m\}$, otherwise update $\mathcal{A}_{k+1} = \mathcal{A}_k \cup \{m\}$.

Step 3. Repeat step 2 untill it reaches $\lambda_k = \lambda_{\max}$ where all $\beta_j = 0$ except for the index j which has the largest corresponding $|\mu_j|$.

4.3 Complexity of the Constrained Lasso

In this section we are going to establish a complexity result for the regularized problem (4.10) with intuitions from Mairal and Yu (2012). First we present a lemma concerning the sign patterns of two related problems:

Lemma 6. *Assume $X \in \mathbb{R}^{n\times p}$ is a full column rank matrix and $\boldsymbol{y} \in \mathbb{R}^n$ is a vector. Then there exists a symmetric positive definite matrix Σ_0, a constraint vector $\boldsymbol{u} \in \mathbb{R}^p$ such that after excluding the zero sign pattern, the sign pattern set of the optimization problem*

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{4.14}$$

is completely included in the sign pattern set of the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{w}'\boldsymbol{u}=1} \frac{1}{2}\boldsymbol{w}'\Sigma_0\boldsymbol{w} + \lambda\|\boldsymbol{w}\|_1. \tag{4.15}$$

Now we can give a lemma on the upper bound of the number of linear segments in the solution path of problem (4.15):

Lemma 7. *Assume $\Sigma \in \mathbb{R}^{p\times p}$ is positive definite, then the number of linear segments in the regulariztion path of problem (4.15) is less than $(3^p + 1)/2$.*

In Mairal and Yu (2012), the authors proved the upper bound $(3^p + 1)/2$ actually can be achieved for problem (4.14), combining their result with Lemma 6 and 7, we can obtain a similar result for problem (4.10)(or equivalently, problem (4.15)) as stated in the following theorem.

Theorem 7. *Suppose we are free to choose the symmetric positive definite matrix* $\Sigma \in \mathbb{R}^{p\times p}$ *and the constraint vector* $\boldsymbol{u} \in \mathbb{R}^p$ *in problem (4.10). Then in the worst scenario, the regulariztion path of this problem has exactly* $(3^p - 1)/2$ *linear segments.*

We provide some pathological solution path plots here to show that the upper exponential bound in Theorem 7 are indeed achievable. In Figures 4.1 and 4.2 , we observe that there are 4 an 13 linear segments in the solution path respectively which coincides with the value of the identity $(3^p - 1)/2$ when $p = 2$ and 3.

4.4 Proofs

Proof of Proposition 8. See the appendix of Bertsekas (2008). □

Proof of Proposition 9. The continuity of $\boldsymbol{\beta}^\lambda$ in λ can be argued by contradiction.

Suppose $\boldsymbol{\beta}^\lambda$ is not continuous at λ_0, then there exists a sequence $\lambda_n \to \lambda_0 (n \in \mathbb{Z}^+)$ and an $\epsilon > 0$, such that $\boldsymbol{\beta}^{\lambda_n} \notin \mathcal{B}_\epsilon(\boldsymbol{\beta}^{\lambda_0})$ for all $n \in \mathbb{Z}^+$. Define $f_\lambda(\boldsymbol{\beta}) \triangleq \boldsymbol{\beta}'\Sigma\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1$, then we can find an $\boldsymbol{x} \in \mathcal{B}_\epsilon(\boldsymbol{\beta}^{\lambda_0})$ such that $f_{\lambda_0}(\boldsymbol{x}) < f_{\lambda_0}(\boldsymbol{\beta}^{\lambda_n}) - C$ for all $n \in \mathbb{Z}^+$ where C is a positive constant. So at λ_n, we have

$$f_{\lambda_n}(\boldsymbol{x}) - f_{\lambda_n}(\boldsymbol{\beta}^{\lambda_n}) = f_{\lambda_0}(\boldsymbol{x}) - f_{\lambda_0}(\boldsymbol{\beta}^{\lambda_n}) + (\lambda_n - \lambda_0)(\|\boldsymbol{x}\|_1 - \|\boldsymbol{\beta}^{\lambda_n}\|_1)$$

$$< -C + (\lambda_n - \lambda_0)(\|\boldsymbol{x}\|_1 - \|\boldsymbol{\beta}^{\lambda_n}\|_1).$$

WLOG, we can assume $\|\boldsymbol{\beta}^{\lambda_n}\|_1$ is bounded for all $n \in \mathbb{Z}^+$ so as $n \to \infty$ the second term in the above inequality would tend to 0. Hence we have for sufficiently large n, $f_{\lambda_n}(\boldsymbol{x}) < f_{\lambda_n}(\boldsymbol{\beta}^{\lambda_n})$ and this leads to a contradiction.

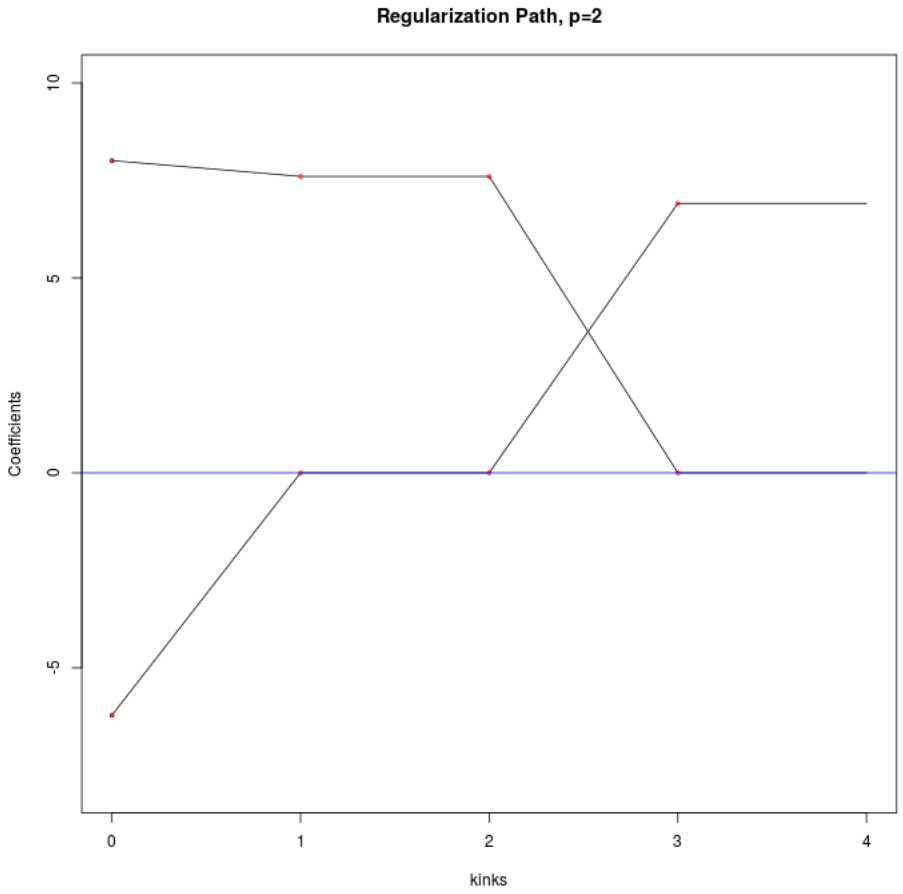

Figure 4.1: Pathological regularization path with $p = 2$ variables and $(3^2-1)/2 = 4$ kinks. The curves represent the values of the coefficients at every kink of the path. We use a non-linear(log) scale for the coefficients.

To prove the equality, we apply the previous proposition directly. Here the cost function $f(\boldsymbol{\beta}) = \boldsymbol{\beta}'\Sigma\boldsymbol{\beta}+\lambda\|\boldsymbol{\beta}\|_1$ and the set X is the hyperplane which satisfies $\boldsymbol{\beta}'\mathbf{1} = 1$. Note that here $\partial f(\boldsymbol{\beta}) = 2\Sigma\boldsymbol{\beta} + \lambda\boldsymbol{d}$ where $\boldsymbol{d} \in \partial\|\boldsymbol{\beta}\|_1$, the subdifferential set of the L_1 norm at $\boldsymbol{\beta}$.

When $\boldsymbol{\beta}$ ranges over X, $\boldsymbol{\beta}-\boldsymbol{\beta}^\lambda$ essentially gives us the orthogonal compliment of the space spanned by the $\mathbf{1}$ vector, which has dimension $p-1$. By the structure of such a linear space, the inequality in the proposition now turns into the equality $(2\Sigma\boldsymbol{\beta}^\lambda + \lambda\boldsymbol{d}^\lambda)'(\boldsymbol{\beta} - \boldsymbol{\beta}^\lambda) = 0$. This indicates $2\Sigma\boldsymbol{\beta}^\lambda + \lambda\boldsymbol{d}^\lambda = c\mathbf{1}$ for some constant

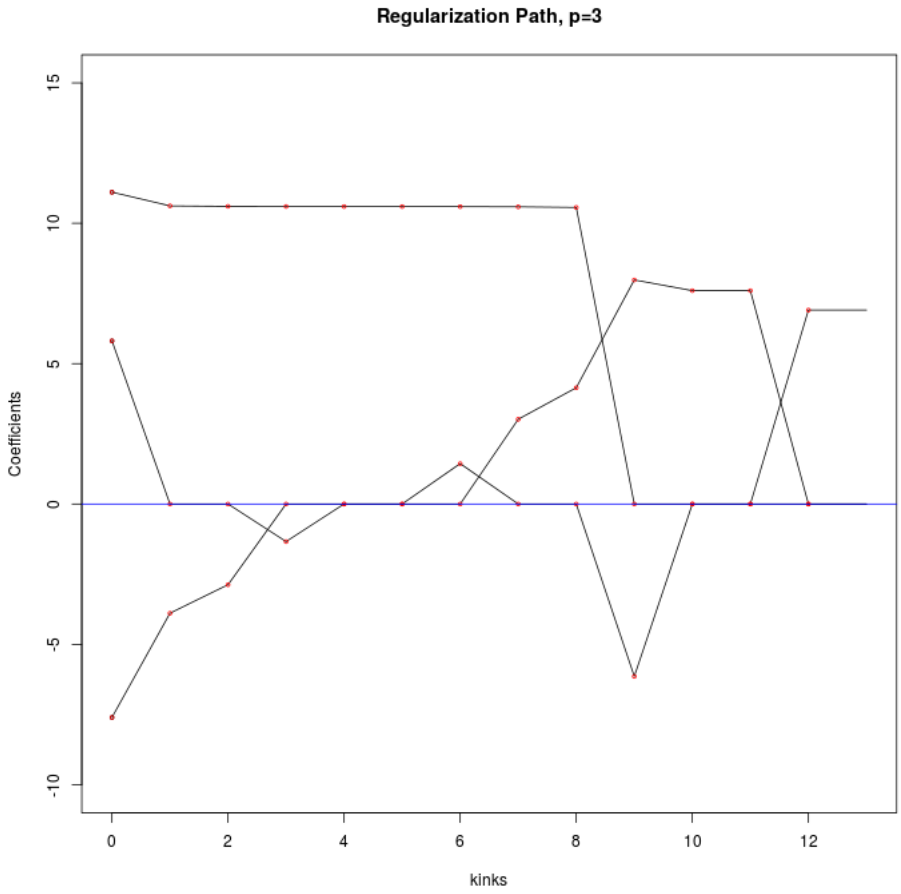

Figure 4.2: Pathological regularization path with $p = 3$ variables and $(3^3-1)/2 = 13$ kinks. The curves represent the values of the coefficients at every kink of the path. We use a non-linear(log) scale for the coefficients.

c. Then if and only if assertion in Proposition 2 implies the converse is true.

For uniqueness, we argue by contradiction. If not, then there exist another subgradient $\boldsymbol{d}_1^\lambda \neq \boldsymbol{d}^\lambda$ and another constant $c_1 \neq c$ such that equality (4.5) holds for for $\boldsymbol{d}_1^\lambda$ and c_1. Take the difference of the two equalities, we have

$$\lambda(\boldsymbol{d}_1^\lambda - \boldsymbol{d}^\lambda) = (c_1 - c)\mathbf{1}.$$

Since $\boldsymbol{\beta}'\mathbf{1} = 1$, we know $\beta_j \neq 0$ for some j at λ, this indicates the jth component of $\boldsymbol{d}_1^\lambda - \boldsymbol{d}^\lambda$ is 0. This is impossible since the jth component in the right hand side is nonzero.

At last it remains to prove the continuity of $\boldsymbol{d}^{\lambda}$ in λ. Note that we only need to prove the constant c is continuous in λ. From now on, let us use $c(\lambda)$ to denote the unique constant value at λ. If $c(\lambda)$ is not continuous at the point λ_0, then there exists some $\epsilon > 0$ and a sequence $\{\lambda_n\}$ satisfying $\lim_{n\longrightarrow\infty} |\lambda_n - \lambda_0| = 0$ such that $|c(\lambda_n) - c(\lambda_0)| > \epsilon$ for any n. Evaluate the equality (4.5) at λ_0 and λ_n, take the difference, we have

$$2\Sigma(\boldsymbol{\beta}^{\lambda_n} - \boldsymbol{\beta}^{\lambda_0}) + (\lambda_n \boldsymbol{d}^{\lambda_n} - \lambda_0 \boldsymbol{d}^{\lambda_0}) = (c(\lambda_n) - c(\lambda_0))\mathbf{1}. \tag{4.16}$$

Like when we prove the uniqueness, we know $\beta_j \neq 0$ for some index j in a neighborhood of λ_0. Now we look at the jth component of equality (4.16). Since $\boldsymbol{\beta}^{\lambda}$ is continuous in λ and $d_j^{\lambda_n} = d_j^{\lambda_0}$ for sufficiently large n, we know the jth component of the left hand side will converge to 0. But the jth component of the right hand side clearly wouldn't converge to 0. So we have a contradiction and this completes the proof. □

Proof of Theorem 5. First we define $\eta(\lambda) \triangleq \text{sign}(\beta^{\lambda})$ to be the sign vector at λ for any $\lambda > 0$. At any λ_1, λ_2 where $0 < \lambda_1 < \lambda_2$, we have $2\Sigma\boldsymbol{\beta}^{\lambda_1} + \lambda_1 \boldsymbol{d}^{\lambda_1} = c_1\mathbf{1}$ and $2\Sigma\boldsymbol{\beta}^{\lambda_2} + \lambda_2 \boldsymbol{d}^{\lambda_2} = c_2\mathbf{1}$ for some c_1 and c_2 by the subgradient optimality condition we proved in propositon 2. For any $\alpha \in [0, 1]$, multiply the first above equality by α and the second by $1-\alpha$ then add them up, it is easy to see that $\alpha\boldsymbol{\beta}^{\lambda_1} + (1-\alpha)\boldsymbol{\beta}^{\lambda_2}$ satisfies the subgradient optimality condition at $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2$. So we have $\boldsymbol{\beta}^{\alpha\lambda_1+(1-\alpha)\lambda_2} = \alpha\boldsymbol{\beta}^{\lambda_1} + (1-\alpha)\boldsymbol{\beta}^{\lambda_2} \triangleq \boldsymbol{\beta}_{\alpha}^{\lambda_1,\lambda_2}$.

This implies whenever two optimal solutions $\boldsymbol{\beta}^{\lambda_1}$ and $\boldsymbol{\beta}^{\lambda_2}$ have the same signs, the solution path between λ_1 and λ_2 is a linear segment. This together with the fact that the sign patterns of $\boldsymbol{\beta}^{\lambda}$ as λ varies over $(0, +\infty)$ is at most 3^p implies the solution path $\{\boldsymbol{\beta}^{\lambda} | \lambda \geqslant 0\}$ is continuous and piecewise linear with finite many pieces. □

Proof of Lemma 6. First we define

$$\boldsymbol{u} \triangleq \frac{X'\boldsymbol{y}}{\eta} \qquad \Sigma \triangleq X'X - \frac{X'\boldsymbol{y}\boldsymbol{y}'X}{\eta} \tag{4.17}$$

where η is a positive real number chosen to be large enough to make Σ positive definite.

Then from (4.17) we can solve the following relationships:

$$\Sigma + \eta\boldsymbol{u}\boldsymbol{u}' = X'X \qquad \eta\boldsymbol{u}' = \boldsymbol{y}'X. \tag{4.18}$$

Now we note that after we expand the quadratic term in problem (4.14), it takes the form

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2}\boldsymbol{w}'X'X\boldsymbol{w} + \lambda\|\boldsymbol{w}\|_1 - \boldsymbol{y}'X\boldsymbol{w}. \tag{4.19}$$

Substituting $X'X$ and $\boldsymbol{y}'X$ from (4.18), we have the problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2}\boldsymbol{w}'(\Sigma + \eta\boldsymbol{u}\boldsymbol{u}')\boldsymbol{w} + \lambda\|\boldsymbol{w}\|_1 - \eta\boldsymbol{w}'\boldsymbol{u} \tag{4.20}$$

which has an equivalent form

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda\|\boldsymbol{w}\|_1 + \frac{1}{2}\eta(\boldsymbol{w}'\boldsymbol{u} - 1)^2. \tag{4.21}$$

Now given λ, we denote $\boldsymbol{w}^*$ to be the solution of problem (4.20) and c_λ to be the value of $\boldsymbol{w}^{*\prime}\boldsymbol{u}$. Then when $c_\lambda \neq 0$(note that from the equivalent form (4.21) it is not hard to see $0 \leqslant c_\lambda < 1$), with the variable transformation $\boldsymbol{w}_0 = \boldsymbol{w}/c_\lambda$, the solution of the problem (4.20) is proportional to the solution of the following optimization problem

$$\min_{\boldsymbol{w}_0\in\mathbb{R}^p, \boldsymbol{w}_0'\boldsymbol{u}=1} \frac{1}{2}c_\lambda^2\boldsymbol{w}_0'(\Sigma + \eta\boldsymbol{u}\boldsymbol{u}')\boldsymbol{w}_0 + \lambda * c_\lambda * \|\boldsymbol{w}_0\|_1$$

which is in turn equivalent to

$$\min_{\boldsymbol{w}_0\in\mathbb{R}^p, \boldsymbol{w}_0'\boldsymbol{u}=1} \frac{1}{2}\boldsymbol{w}_0'(\Sigma + \eta\boldsymbol{u}\boldsymbol{u}')\boldsymbol{w}_0 + \frac{\lambda}{c_\lambda}\|\boldsymbol{w}_0\|_1. \tag{4.22}$$

So we have shown that after excluding the zero sign pattern, the sign pattern set of problem (4.20) is included in the sign pattern set of problem (4.21). Now

we see problem (4.21) is essentially a form of problem (4.15) and to complete the proof, we only need to choose $\Sigma_0 = \Sigma + \eta \boldsymbol{u}\boldsymbol{u}'$.

□

Proof of Lemma 7. This follows by applying a similar procedure as in Mairal and Yu (2012). □

Proof of Theorem 7. Note that Lemma 6 essentially relates problem (4.15) with problem (4.14). So we can simply choose the the design of X and $\boldsymbol{y}$ for the worst-case complexity in Mairal and Yu (2012) for problem (4.14) to construct the corresponding Σ and $\boldsymbol{u}$ for our need in problem (4.15), since the zero sign pattern can not be included in problem (4.14), we essentially have a construction of Σ and $\boldsymbol{u}$ with $\frac{3^p+1}{2} - 1 = \frac{3^p-1}{2}$ linear segments. □

Bibliography

T. Anderson. An introduction to multivariate statistical analysis, 2003.

D. Bertsekas. Nonlinear programming, 3rd printing with corrections. *Athena Scientific*, 2008.

P. J. Bickel and E. Levina. Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004.

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

M. Bilodeau and D. Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 2008.

G. E. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.

T. S. Breusch. Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31):334–355, 1978.

J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

R. Chen, H. Xiao, and D. Yang. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 2020.

Y. S. Chow and H. Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2012.

V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management science*, 55(5):798–812, 2009.

M. L. Eaton and D. Tyler. The asymptotic distribution of singular-values with applications to canonical correlations and correspondence analysis. *Journal of Multivariate Analysis*, 50(2):238–264, 1994.

M. L. Eaton and D. E. Tyler. On wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *The Annals of Statistics*, pages 260–271, 1991.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Fan, Y. Feng, and X. Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012a.

J. Fan, Y. Li, and K. Yu. Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428, 2012b.

J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012c.

L. G. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica: Journal of the Econometric Society*, pages 1293–1301, 1978.

P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic press, 2014.

R. A. Horn, R. A. Horn, and C. R. Johnson. *Matrix analysis*. Cambridge university press, 1990.

J. Hosking. Lagrange-multiplier tests of multivariate time-series models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):219–230, 1981.

J. R. Hosking. The multivariate portmanteau statistic. *Journal of the American Statistical Association*, 75(371):602–608, 1980.

G. M. James, C. Paulson, and P. Rusmevichientong. The constrained lasso. In *Refereed Conference Proceedings*, volume 31, pages 4945–4950. Citeseer, 2012.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

W. Li and A. McLeod. Distribution of the residual autocorrelations in multivariate arma time series models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):231–239, 1981.

G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

R. J. Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

J. Shao, Y. Wang, X. Deng, S. Wang, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2): 1241–1265, 2011.

D. E. Tyler. Radial estimates and the test for sphericity. *Biometrika*, 69(2): 429–436, 1982.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.