SOME RESULTS IN HIGH-DIMENSIONAL STATISTICS

A Thesis
Presented to
The Academic Faculty

by

Yi Xiao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2015

SOME RESULTS IN HIGH-DIMENSIONAL STATISTICS

Approved by:

Professor Ming Yuan, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor C. F. Jeff Wu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Liang Peng
J. Mack Robinson College of Business
Georgia State University

Professor Roshan Vengazhiyil
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 11 May 2015

To my parents and grandparents.

ACKNOWLEDGEMENTS

First of all, I want to express my deepest gratitude for my advisor, Professor Ming Yuan. His enthusiasm and knowledge led me into the world of high-dimensional statistics. During my study, he was always supportive and respectful of my research ideas. He guided me through the most difficult times and provided necessary help while allowing me to develop independent research skills. It was an honor to have had the chance to work with Professor Ming Yuan.

I am also thankful to my thesis committee members: Dr. Jeff Wu, Dr. Roshan Vengazhiyil, Dr. Jye-Chyi Lu, Dr. Yao Xie and Dr. Liang Peng for their kindness in evaluating my thesis and their constructive feedback. They helped refine my thesis and encouraged me to achieve better results. I want to thank Dr. Alan Erera, Dr. Paul Kvam and Dr. R. Gary Parker for their dedicated service as the current and former Associate Chair for Graduate Studies of the department of Industrial and Systems Engineering.

I also appreciate all the help from my friends at Georgia Tech. Special thanks to Qianyi Wang, as your company was an indispensable part of the exciting journey of my Ph.D. studies. The enjoyable and unforgettable time we spent together made me a better person.

Finally, I want to thank my parents and grandparents, without whom I would not be the man I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

High-dimensional statistics is one of the most active research topics in modern statistics. It also has applications in many fields, such as computer science, biology and economics. Recent advancements in computer technology enable large volumes of data to be collected and stored relatively easily. Combining this with the advancements in processing and analytical capabilities of computers, we see an even faster growth in research and technology, making our daily lives much easier than ever before. At the same time, the complexity of data both in size and structure brings new challenges to statisticians, to be able to differentiate useful information from noise in an efficient and accurate manner.

A common problem in high-dimensional statistics is when the number of covariates exceeds the sample size. Most classical approaches would either be inapplicable or produce unsatisfactory results in such problems., although extensive research efforts have been made to overcome these difficulties. One of the more popular approaches to tackle the lack of degrees of freedom is to introduce additional assumptions on the data structure to reduce model complexity, such as sparsity of coefficients for linear regression models and sparsity of inverse covariance matrix for Gaussian graphical models. It is shown that under certain assumptions and with proper regularization on the parameters we can obtain reasonably good estimates for these models even if the sample size is limited. However, it is still unclear how to justify these assumptions in certain scenarios.

The purpose of this thesis is to narrow the gap between theory and practice in the field of high-dimensional statistics by studying some of the more widely adopted assumptions in literature and by introducing new testing procedures. To be more specific, we

will cover l_1-regularized estimations for time series and testing for the sparse Gaussian graphical model.

In the first chapter we explore the applications of l_1-regularized regression methods for Gaussian vector autoregressive processes. We decompose the classical regression model into smaller submodels and obtain sparse solutions by applying l_1-penalties. We show that under mild conditions the design matrices corresponding to the submodels are actually generated from some α-mixing processes. Therefore, a more general problem is to study the performance of the l_1-regularized methods for a linear model with a random design matrix that is generated by an α-mixing Gaussian process with exponential decay rate. Our main result verifies the restricted eigenvalue assumption for the mixing random design based on the generic chaining technique, and derives the l_p error bound for the Lasso and Dantzig selectors. We also study the sufficient conditions for a VAR(p) model to guarantee a tight error bound of the solutions and discuss how to select the order of the model. Finally, we illustrate the variable selection and estimation performance of Lasso by several sets of simulation.

In the second chapter, we propose a new statistic to test the decomposable structure of a Gaussian graphical model in the high-dimensional setting. It is based on the eigenvalues of the sample covariance matrix. In the case when the null hypothesis corresponds to a group independence structure, we derive the asymptotic distribution of the proposed statistic and show that it is invariant under non-singular linear transformations within each group. When testing an arbitrary decomposable structure, a simple asymptotic distribution of the statistic is not available. We suggest a simulation-based method to approximate the null distribution and calculate the corresponding p value. We also study the computational complexity of the proposed methods and give some suggestions on how to improve the performance. In the last section, We give some numerical results including both simulation and an empirical example to study the proposed testing procedure in different scenarios.

CHAPTER I

RESTRICTED EIGENVALUE PROPERTIES FOR VECTOR AUTOREGRESSIVE PROCESSES

1.1 Introduction

High-dimensional data, with the number of features p comparable to or even exceeding the sample size n, brings new opportunities as well as challenges. Many applications in fields such as economics, computational biology and geology involve parameter estimation for high-dimensional time series [43, 63, 80, 48]. Yet the performance of their method is unclear. As shown in [69, 29], a common approach to modeling linear stationary processes is based on an autoregressive (AR) model. In classical setting, where $n \to \infty$ and $p < n$ is fixed, we can easily obtain an ordinary least squares estimation [46]. However, this method fails in a high-dimensional setting. For example, the design matrix in a linear regression model for the AR process is degenerated when $p > n$ and standard least squares estimate is not available. Even if $p < n$, with a large number of predictors, the classical solution would be highly vulnerable to collinearity between the predictors and overfitting.

In recent literature, many regularization approaches based on an assumption of sparsity have been proposed to overcome the curse of dimensionality. To name a few, [75] proposed a least absolute shrinkage and selection operator (Lasso) for linear regression to simultaneously select significant variables and estimate coefficients. [23] proposed a smoothly clipped absolute deviation penalty on the loss function to achieve nearly unbiasedness when estimating large parameters. [14] introduced the Dantzig selector, which selects solutions with the least l_1 norm of weights among candidate solutions that have small losses. [11] applied nonnegative garrote procedure to shrink OLS

estimates and reduce prediction error.

Among these methods, the Lasso type regularization is arguably the most popular one due to its computational efficiency and flexibility [76]. [22] introduced the LARS algorithm to compute the solution path of a LASSO regression. [28] proposed the coordinate descent algorithm which was further studied in [26, 27] for efficient computation with large p. There are many Lasso generalizations and variants developed recently such as group Lasso, elastic net, adaptive Lasso, graphical Lasso, fused Lasso and matrix completion [86, 91, 77, 90, 25, 85, 16].

Some researchers have studied Lasso type regularization for parameter estimation and selection in time series. [82] discussed a linear regression model with autoregressive errors. [57] studied the AR model and obtained selection and estimation consistency but with strong assumptions like $p = O(\log n)$ and an incoherence condition [81] that is difficult to validate. [70] studied large vector autoregressive (VAR) models with both dimensionality and number of lags going to infinity while sample size remains moderate. However, it was also based on the restricted eigenvalue (RE) assumption introduced in [9]. [32] proposed a subset selection method for vector autoregressive processes. But its theoretical performance is unknown. See [74, 49, 59, 36, 4, 84] for more references. Although the Lasso type method is widely used in time series studies, we still need to establish a better theory to justify it. A deep understanding of its connection with time series would be beneficial to future research.

The conditions required for the success of l_1-based relaxation are well-developed. For noiseless models, a restricted nullspace property was proposed in [20] for exact signal recovery of the basis pursuit algorithm in compressed sensing. In the noisy cases, an non-exhaustive list includes the restricted isometry property (RIP) [15], the incoherence condition [21], the RE assumption [9] and the irrepresentable condition [87]. For more references discussing the consistency of Lasso type methods, see [51, 78]. There are also studies on what kind of design matrices satisfy these conditions. [53, 1, 66]

showed that RIP holds with high probability for sub-Gaussian and sub-exponential random matrices with i.i.d. entries and for random matrices from unitary ensembles with moderate sample size. However, for general regression problems we can hardly assume that the design matrices have i.i.d. entries or that they are from unitary ensembles. A weaker assumption such as the RE assumption - instead of the RIP - is more suitable when we do not have control over the design matrix. [65] studied the RE assumption for correlated Gaussian designs and gave inspiring result that RE assumption is satisfied for a broad class of random matrices. [89] relaxed the Gaussian assumption and extended RE assumption to sub-Gaussian random matrices. But they still assumed that the rows are i.i.d. which is generally not applicable to time series design.

In this chapter, we focus on the parameter estimation for an autoregressive process in a high-dimensional setting and study the RE assumption in a α-mixing scenario. The chapter is organized as follows: In Section 1.2, we give some basic introductions on the VAR(p) model. Section 1.3 introduces regularized modeling and RE assumptions. Section 1.4 presents our main theorem on the RE assumption for an α-mixing Gaussian process. In Section 1.5 we study the convergence for the Lasso and Dantzig selector. Section 1.6 discusses the α-mixing property for the VAR(p) model. Section 1.7 focuses on the property of the population covariance of the design matrix. In Section 1.8 we discuss how to select the order of the VAR(p) model. Finally, we give some illustrative numerical results in Section 1.9 to show the performance of the proposed estimators.

1.2 Vector Autoregressive Processes

Instead of a univariate autoregressive model, we consider a more general VAR(p) model

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \qquad t = 0, \pm 1, \pm 2, \ldots, . \tag{1}$$

Here we follow the notation in [46]. $y_t = (y_{1t}, \ldots, y_{Kt})'$ is a random vector. Each A_i is a fixed K-dimensional coefficient matrix, $u_t = (u_{1t}, \ldots, u_{Kt})'$ is a white noise process, that is, u_t satisfies the following conditions:

$$\mathbb{E}(u_i) = 0,\ \mathbb{E}(u_i u_i') = \Sigma_u,\ \mathbb{E}(u_i u_j') = 0, \tag{2}$$

for any $i \neq j$. Without loss of generality, we assume that $\mathbb{E}(y_t) = 0$ for all t and drop the intercept in the model.

We make some more assumptions on the VAR(p) model that are essential to our analysis as follows:

Assumption 1.2.1 *All roots of the reverse characteristic polynomial of the VAR(p) process are outside the complex unit circle, i.e., for* $|z| \leq 1$,

$$\det(I_K - A_1 z - \cdots - A_p z^p) \neq 0. \tag{3}$$

Assumption 1.2.2 u_t *is Gaussian white noise, i.e., in addition to the definition of a white noise process,* u_t *also follows the multivariate normal distribution* $N(0, \Sigma_u)$ *for all* t.

Assumption (1.2.1) is a widely used stability condition that ensures the convergence of the infinite sum of innovations. Assumption (1.2.2) ensures that y_t is a Gaussian process.

Now suppose we have a multiple time series $y_{-p+1}, \ldots, y_0, y_1, \ldots, y_T$ that is generated by this process. Similarly to the notation in [46], we define

$$Y := (y_T, \ldots, y_1),\ A := (A_1, \ldots, A_p),\ U := (u_T, \ldots, u_1), \tag{4}$$

$$Z_t := (y_t', \ldots, y_{t-p+1}')',\ Z := (Z_{T-1}, \ldots, Z_0). \tag{5}$$

Model (1) can be formulated as a linear regression problem:

$$Y = AZ + U. \tag{6}$$

Define

$$\mathbf{y} := \text{vec}(Y),\ \boldsymbol{\beta} := \text{vec}(A),\ \mathbf{u} := \text{vec}(U), \tag{7}$$

where vec is the column stack operator, a least squares estimation for the entries of $A_1, \ldots, A_p$, or $\boldsymbol{\beta}$, can be obtained by

$$\hat{\boldsymbol{\beta}} = ((ZZ')^{-1}Z \otimes I_K)\mathbf{y}. \tag{8}$$

where $\otimes$ is the Kronecker product. Notice that the least squares estimator does not depend on Σ_u.

Based on (8), we can see that it is equivalent to estimate coefficient matrix A row by row. If we denote each rows of the matrices Y, A, U by $\widetilde{\mathbf{y}}_i', \widetilde{\boldsymbol{\beta}}_i', \widetilde{\mathbf{u}}_i'$ where $i = 1, \ldots, K$, the submodels can be written as

$$\widetilde{\mathbf{y}}_i = Z'\widetilde{\boldsymbol{\beta}}_i + \widetilde{\mathbf{u}}_i, \qquad i = 1, \ldots, K. \tag{9}$$

The estimation for A is simply $[\hat{\widetilde{\boldsymbol{\beta}}}_1, \ldots, \hat{\widetilde{\boldsymbol{\beta}}}_K]'$ where

$$\hat{\widetilde{\boldsymbol{\beta}}}_i = (ZZ')^{-1}Z\widetilde{\mathbf{y}}_i, \qquad i = 1, \ldots, K. \tag{10}$$

From equation (10) we can see that the coefficients of submodels (9) can be estimated in the classical low-dimensional setting as if the design matrix Z' is independent of the innovations $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{u}}_i \sim N(0, \widetilde{\sigma}^2 I_T)$ for some $\widetilde{\sigma} > 0$. In the following sections, we will focus on model (9) and discuss its property.

1.3 Regularized autoregressive modeling

Under the previous assumptions (1.2.1) and (1.2.2), if p and K remain fixed while T goes to infinity, it can be shown that we have both asymptotic normality and consistency for the LS estimator (8) of the VAR(p) model (1) and hence the estimators (10) of the submodels (9) [46]. However, in a high-dimensional setting where sample size is limited and $Kp \gg T$, the classical regression approach will fail due to the degeneration of the design matrix Z'.

In order to estimate $\widetilde{\boldsymbol{\beta}}_i$, one of the most widely-adopted assumptions is the sparsity of $\widetilde{\boldsymbol{\beta}}_i$. If $\|\widetilde{\boldsymbol{\beta}}_i\|_0 \leq s$ for $i = 1, \ldots, K$ where $\|\cdot\|_0$ is the l_0 norm operator and s is small compared to Kp and T, it is still possible for us to find a solution that reasonably approximates the true coefficients. A natural formulation of the optimization problem associated with the sparsity assumption for (9) is

$$\underset{\widetilde{\boldsymbol{\beta}}_i \in \mathbb{R}^{Kp}}{\arg\min} \|\widetilde{\mathbf{y}}_i - Z'\widetilde{\boldsymbol{\beta}}_i\|_2^2 \quad \text{s.t.} \quad \|\widetilde{\boldsymbol{\beta}}_i\|_0 \leq s. \tag{11}$$

However, this is a non-convex l_0 optimization problem with combinatorial complexity in computation that is not favorable in a high-dimensional setting. A more computationally friendly relaxation is to impose an l_1 penalty on $\widetilde{\boldsymbol{\beta}}_i$, which leads to the Lasso formulation originally proposed in [75]:

$$\underset{\widetilde{\boldsymbol{\beta}}_i \in \mathbb{R}^{Kp}}{\arg\min} \left\{ \frac{1}{2T} \|\widetilde{\mathbf{y}}_i - Z'\widetilde{\boldsymbol{\beta}}_i\|_2^2 + \lambda_T \|\widetilde{\boldsymbol{\beta}}_i\|_1 \right\}. \tag{12}$$

This is a convex optimization problem that can be solved in a fast and reliable manner. We also consider the Dantzig selector proposed in [14], which shares similar properties with Lasso type methods. The formulation of the optimization problem is:

$$\underset{\widetilde{\boldsymbol{\beta}}_i \in \mathbb{R}^{Kp}}{\arg\min} \|\widetilde{\boldsymbol{\beta}}_i\|_1 \quad \text{s.t.} \quad \left\| \frac{1}{T} Z(\widetilde{\mathbf{y}}_i - Z'\widetilde{\boldsymbol{\beta}}_i) \right\|_\infty \leq \lambda_T. \tag{13}$$

Even though exact recovery of $\widetilde{\boldsymbol{\beta}}_i$ is not possible due to the noise presented, it was shown that if the design matrix Z' satisfies certain conditions and λ_T is properly chosen, we can still control the error bounds of the l_1 regularized methods. Let $u_I \in \mathbb{R}^{|I|}$ denote the subvector of $u \in \mathbb{R}^p$ confined to I. We define the following restricted eigenvalue (RE) assumption on the design matrix introduced in [9]:

Definition 1.3.1 (RE assumption for a $n \times p$ design matrix $RE(s, k_0, X)$) *There exist some integer $1 \leq s \leq p$ and $k_0 > 0$ such that:*

$$\frac{1}{L(s, k_0, X)} := \min_{\substack{I_0 \subset \{1,\ldots,p\}, \\ |I_0| \leq s}} \min_{\substack{u \neq 0, \\ \|u_{I_0^c}\|_1 \leq k_0 \|u_{I_0}\|_1}} \frac{\|Xu\|_2}{\sqrt{n}\|u_{I_0}\|_2} > 0, \tag{14}$$

where $L(s, k_0, X)$ is the restricted eigenvalue constant for the design matrix X.

It is shown in [9] that the RE assumption is less severe than the restricted isometry property (RIP) but still guarantees l_p error bounds. As long as Z' satisfies the restricted eigenvalue assumption $RE(s, k_0, X)$ with constant $k_0 \geq 3$ for the Lasso and $k_0 \geq 1$ for the Dantzig selector, the l_2 error $\|\hat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i\|_2$ can be bounded in order $O(\sqrt{\frac{s\log(Kp)}{T}})$ with high probability. Therefore, given the underlying structure of Z', we are interested in when Z' will satisfy the RE assumption and guarantee such an error bound given the fact that there exists some correlation between the rows of Z'. Here we give another assumption which is useful in characterizing the structure of Z'.

Definition 1.3.2 (RE assumption for a $p\times p$ population covariance matrix $RE(s, k_0, \Sigma)$) *There exist some integer* $1 \leq s \leq p$ *and* $k_0 > 0$ *such that:*

$$\frac{1}{L(s, k_0, \Sigma)} := \min_{\substack{I_0 \subset \{1,\ldots,p\}, \\ |I_0| \leq s}} \min_{\substack{u \neq 0, \\ \|u_{I_0^c}\|_1 \leq k_0 \|u_{I_0}\|_1}} \frac{\|\Sigma^{1/2} u\|_2}{\|u_{I_0}\|_2} > 0, \tag{15}$$

where $L(s, k_0, \Sigma)$ *is the restricted eigenvalue constant for the population covariance matrix* Σ.

1.4 RE assumption for an α-mixing Gaussian process

First, we need to introduce an α-mixing condition following the notations in [57]. Let $\{\Psi_t\}$ be a time series that is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_i^j$ denote the σ-field that is generated by part of the time series $(\Psi_i, \ldots, \Psi_j)$, for $-\infty \leq i \leq j \leq \infty$. Define

$$\alpha(\mathcal{A}_1, \mathcal{A}_2) = \sup_{A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2} |\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)|. \tag{16}$$

for any two σ-fields $\mathcal{A}_1$ and $\mathcal{A}_2$. We say that $\{\Psi_t\}$ is α-mixing if the mixing coefficients $\alpha_\Psi(k) \to 0$, as $k \to \infty$, where

$$\alpha_\Psi(k) = \sup_{s \in \mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^s, \mathcal{F}_{s+k}^\infty). \tag{17}$$

The α-mixing condition plays an important role in the analysis of weakly dependent processes. Many results such as the central limit theorem or concentration of measure

that base on the independent assumption have their counterparts for an α-mixing process with exponential decay rate [88]. This inspires us to study a similar condition that is essential to our results.

The following theorem is the main contribution of this chapter. It is mainly inspired by [89] and extends their results by relaxing the independent rows assumption.

Theorem 1.4.1 *Let* $0 < \theta < 1$, $1 \leq n \leq p$ *and* $0 < s \leq p/2$. *Let* $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)'$ *be an* α*-mixing Gaussian process with mixing coefficients* $\alpha(n) \leq c\rho^n$ *where* $c > 0$ *and* $0 < \rho < 1$ *are constants and* $\Psi_i \sim N(0, I_p)$ *for* $1 \leq i \leq n$. *Let* Σ *be a covariance matrix that satisfies* $RE(s, k_0, \Sigma)$ *with constant* $L(s, k_0, \Sigma)$. *Let*

$$\rho(s, \Sigma) = \sup_{\substack{\|u\|_2=1, \\ |supp(u)|\leq s}} \|\Sigma^{1/2}u\|_2^2, \tag{18}$$

$$C_* = (6 + 3k_0)L(s, k_0, \Sigma)\sqrt{\rho(s, \Sigma)}. \tag{19}$$

Let c'', c''' *be constants defined as in Corollary (A.1.7) which are positive and only depends on* ρ. *Let* e *be the base of natural logarithm. Let* $X = \Psi\Sigma^{1/2}$. *If* n *satisfies*

$$n > \frac{c''}{\theta^2} C_*^2 s \log(5ep/s), \tag{20}$$

then we have $RE(s, k_0, X)$ *holds with constant* $L(s, k_0, X)$ *satisfying*

$$0 < L(s, k_0, X) \leq \frac{1}{1 - \theta} L(s, k_0, \Sigma). \tag{21}$$

with probability at least $1 - 4\exp(-c'''\theta^2 n)$.

See the appendix for the proof of Theorem (1.4.1). In fields such as time series analysis, temporal correlation always exists and is an important factor in modeling. This result ensures that for a design matrix X generated by a Gaussian process, as long as the correlation between the rows of X is moderate, in the sense that the Gaussian process is α-mixing with exponential decay rate, such correlation will not alter the natural structure of X. The restricted eigenvalue assumption is still valid

provided that the sample size is sufficiently large, which still could be relatively small with respect to the number of covariates.

In fact, the conditions in Theorem (1.4.1) do not require that the underlying Gaussian process is actually generated by a VAR(p) model. It can possibly be applied to more general Gaussian process, such as a well defined MA(∞) process

$$Y_t = \sum_{j=0}^{\infty} \phi_j u_{t-j}. \tag{22}$$

with i.i.d. Gaussian innovations $u_t \sim N(0, \Sigma_u)$ and $\sum_{j=0}^{\infty} \phi_j^2 < \infty$. The same argument is valid given $\{Y_t\}$ is α-mixing with exponential decay rate.

1.5 l_p convergence for the Lasso and Dantzig selector

The immediate consequences following Theorem (1.4.1) are the error bounds for the Lasso and Dantzig selector for an α-mixing design. In general, consider the following linear model:

$$Y = X\beta + \epsilon, \tag{23}$$

where $X = \Psi\Sigma^{1/2}$ is an $n \times p$ design matrix defined in Theorem (1.4.1) and $\epsilon \sim N(0, \sigma^2 I_n)$ is independent of X.

The Lasso problem for model (23) is

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}. \tag{24}$$

The Dantzig selector for model (23) is

$$\arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{s.t.} \quad \left\| \frac{1}{n} X'(Y - X\beta) \right\|_\infty \leq \lambda_n. \tag{25}$$

In addition, assume that $\Sigma_{ii} = 1$ for $i = 1, \ldots, p$. This assumption ensures that each entry of β has the same "weight" in the loss functions in (24) and (25). For the submodels (9) of a VAR(p) model, this assumption is not always satisfied. We need to normalize each column of X to reduce the bias introduced by different scales of

the columns.

The following three conditions proposed in [89] are important in the recovery of β. Define

$$\mathcal{A}_1(\theta) := \{X : RE(s, k_0, X) \text{ and (21) holds}\}, \tag{26}$$

to be the event that X satisfies the RE assumption. Let $X_1, \ldots, X_p$ be the column vectors of X and define

$$\mathcal{A}_2(\theta) := \left\{X : 1 - \theta \leq \frac{\|X_j\|_2}{\sqrt{n}} \leq 1 + \theta \text{ holds for all } 1 \leq j \leq p\right\}, \tag{27}$$

to be the event that the l_2 norm of each column of X has order $O(\sqrt{n})$. For $X \in \mathcal{A}_2(\theta)$ and each $a \geq 0$, define

$$\lambda^* := \sqrt{\frac{2(1+a)\sigma^2 \log p}{n}}, \tag{28}$$

and

$$\mathcal{A}_3 := \left\{\epsilon : \text{for } X \in \mathcal{A}_2(\theta) \text{ and } 0 < \theta < 1, \left\|\frac{X'\epsilon}{n}\right\|_\infty \leq (1+\theta)\lambda^*\right\}, \tag{29}$$

to be the event that the covariance between the noise and the columns of X are bounded. Define

$$\mathcal{A}(\theta) := \mathcal{A}_1(\theta) \cap \mathcal{A}_2(\theta) \cap \mathcal{A}_3, \tag{30}$$

to be the event that all the previous three conditions are satisfied. The following two theorems are the extensions of Theorem 3.1 and 3.2 in [89] respectively.

Theorem 1.5.1 (Error bounds for the Lasso) *Let $0 < \theta < 1$, $a > 0$, $1 \leq n \leq p$ and $0 < s \leq p/2$. Let Σ be a covariance matrix that satisfies $RE(s, 3, \Sigma)$ with constant $L(s, 3, \Sigma)$ and additionally let $\Sigma_{ii} = 1, i = 1, \ldots, p$. Let $\Psi, \sqrt{\rho(s, \Sigma)}, C_*, c'', c'''$ be the same as defined in Theorem (1.4.1). Let $\mathcal{A}(\theta)$ be defined as in (30). Suppose that β^* is a solution to (24) with tuning parameter $\lambda_n \geq 2(1+\theta)\lambda^*$. If we have*

$$n > \frac{c''}{\theta^2} \max(C_*^2 s \log(5ep/s), 9 \log p). \tag{31}$$

Then for

$$d \leq \frac{4K^2(s, 3, \Sigma)}{(1-\theta)^2}, \tag{32}$$

we have

$$\|\beta^* - \beta\|_2 \leq 2d\lambda_n\sqrt{s}, \text{ and } \|\beta^* - \beta\|_1 \leq d\lambda_n s. \tag{33}$$

with probability at least

$$\mathbb{P}(\mathcal{A}(\theta)) \geq 1 - 8\exp(-c'''\theta^2 n) - (\pi \log p)^{-1/2} p^{-a}. \tag{34}$$

Theorem 1.5.2 (Error bounds for the Dantzig selector) *Let* $0 < \theta < 1$, $a > 0$, $1 \leq n \leq p$ *and* $0 < s \leq p/2$. *Let* Σ *be a covariance matrix that satisfies* $RE(s, 1, \Sigma)$ *with constant* $L(s, 1, \Sigma)$ *and additionally let* $\Sigma_{ii} = 1, i = 1, \ldots, p$. *Let* $\Psi, \sqrt{\rho(s, \Sigma)}, C_*, c'', c'''$ *be the same as defined in Theorem (1.4.1). Let* $\mathcal{A}(\theta)$ *be defined as in (30). Suppose that* β^* *is a solution to (25) with tuning parameter* $\lambda_n \geq (1+\theta)\lambda^*$. *If we have*

$$n > \frac{c''}{\theta^2} \max(C_*^2 s \log(5ep/s), 9 \log p). \tag{35}$$

Then for

$$d \leq \frac{4K^2(s, 1, \Sigma)}{(1-\theta)^2}, \tag{36}$$

we have

$$\|\beta^* - \beta\|_2 \leq 3d\lambda_n\sqrt{s}, \text{ and } \|\beta^* - \beta\|_1 \leq 2d\lambda_n s. \tag{37}$$

with probability at least

$$\mathbb{P}(\mathcal{A}(\theta)) \geq 1 - 8\exp(-c'''\theta^2 n) - (\pi \log p)^{-1/2} p^{-a}. \tag{38}$$

The proofs of Theorem (1.5.1) and (1.5.2) are essentially the same as in [89].

1.6 α-mixing property for VAR(p) models

We already know that the design matrix Z' in (9) is generated by a Gaussian process corresponding to the VAR(p) model. To be more clear, we explicitly write down Z'

as:

$$Z' = \begin{bmatrix} y_{1,T-1} & \cdots & y_{K,T-1} & \cdots & \cdots & y_{1,T-p} & \cdots & y_{K,T-p} \\ y_{1,T-2} & \cdots & y_{K,T-2} & \cdots & \cdots & y_{1,T-p-1} & \cdots & y_{K,T-p-1} \\ \vdots & & \vdots & & \vdots & & \vdots & \\ y_{1,0} & \cdots & y_{K,0} & \cdots & \cdots & y_{1,-p+1} & \cdots & y_{K,-p+1} \end{bmatrix}. \tag{39}$$

The derived process corresponding to Z' is defined as

$$\{w_t : w_t = (y_{1,t}, \cdots, y_{K,t}, \ldots, \ldots, y_{1,t-p+1}, \cdots, y_{K,t-p+1})'\}, \tag{40}$$

or

$$\{w_t : w_t = (y_t', \ldots, y_{t-p+1}')'\}. \tag{41}$$

It coincides with the VAR(1) representation of the VAR(p) model (1) in [46]:

$$w_t = \widetilde{A} w_{t-1} + r_t, \tag{42}$$

where

$$\widetilde{A} := \begin{bmatrix} A_1 & A_2 & A_3 & \ldots & A_{p-1} & A_p \\ I_K & 0 & 0 & \ldots & 0 & 0 \\ 0 & I_K & 0 & \ldots & 0 & 0 \\ 0 & 0 & I_K & & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & I_K & 0 \end{bmatrix}, r_t := \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{43}$$

The stability assumption (1.2.1) implies that $\{w_t\}$ is also stable as shown in [46] and the Gaussianity of $\{w_t\}$ is obvious.

In order for $\{w_t\}$ to be an α-mixing process, we only need to show that the underlying process $\{y_t\}$ is α-mixing. The following theorem is a direct result of Theorem 1 from [54].

Lemma 1.6.1 *For model (1), if Assumption (1.2.1) and (1.2.2) are satisfied, $\{y_t\}$ is an α-mixing Gaussian process with mixing coefficients $\alpha(n) < c\rho^n$ where $c > 0$ and $0 < \rho < 1$ are absolute constants.*

Proof [54] shows that for a VAR(p) model (1), if the stability assumption (1.2.1) is satisfied and in addition we have the probability law of u_t is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^K$, then the process is geometrically complete regular, which is stronger than α-mixing with exponential decay rate [60]. The absolute continuity condition is obviously true for Gaussian innovations under Assumption (1.2.2). ∎

Theorem 1.6.2 *Let $\{w_t\}$ be a process defined in (41). If the underlying process $\{y_t\}$ is an α-mixing Gaussian process with mixing coefficients $\alpha_y(n) < c\rho^n$ where $c > 0$ and $0 < \rho < 1$ are absolute constants, then $\{w_t\}$ is an α-mixing Gaussian process with mixing coefficients $\alpha_w(n) < c'\rho^n$ where $c' = c\rho^{-p+1} > 0$.*

Proof Let $\{w_t\}$ be defined on a probability space $(\Omega_w, \mathcal{F}_w, \mathbb{P}_w)$ and $\{y_t\}$ be defined on a probability space $(\Omega_y, \mathcal{F}_y, \mathbb{P}_y)$. By definition,

$$\begin{aligned}\alpha_w(k) &= \sup_{s\in\mathbb{Z}} \alpha([\mathcal{F}_w]_{-\infty}^{s}, [\mathcal{F}_w]_{s+k}^{\infty}) \\ &= \sup_{s\in\mathbb{Z}} \sup_{\substack{A_w\in[\mathcal{F}_w]_{-\infty}^{s},\\ B_w\in[\mathcal{F}_w]_{s+k}^{\infty}}} |\mathbb{P}_w(A_w\cap B_w) - \mathbb{P}_w(A_w)\mathbb{P}_w(B_w)|.\end{aligned} \tag{44}$$

For any two events $A_w \in [\mathcal{F}_w]_{-\infty}^{s}$ and $B_w \in [\mathcal{F}_w]_{s+k}^{\infty}$, there is a natural mapping $g : \mathcal{F}_w \mapsto \mathcal{F}_y$ based on the definition of $\{w_t\}$ such that

$$g(A_w) \in [\mathcal{F}_y]_{-\infty}^{s}, \quad g(B_w) \in [\mathcal{F}_y]_{s-p+1+k}^{\infty}. \tag{45}$$

and

$$\begin{aligned}\mathbb{P}_y(g(A_w)) &= \mathbb{P}_w(A_w), \\ \mathbb{P}_y(g(B_w)) &= \mathbb{P}_w(B_w), \\ \mathbb{P}_y(g(A_w)\cap g(B_w)) &= \mathbb{P}_w(A_w\cap B_w).\end{aligned} \tag{46}$$

Therefore

$$\begin{aligned}
\alpha_w(k) &= \sup_{s\in\mathbb{Z}} \sup_{\substack{A_w\in[\mathcal{F}_w]^s_{-\infty},\\ B_w\in[\mathcal{F}_w]^\infty_{s+k}}} |\mathbb{P}_y(g(A_w)\cap g(B_w)) - \mathbb{P}_y(g(A_w))\mathbb{P}_y(g(B_w))| \\
&\leq \sup_{s\in\mathbb{Z}} \sup_{\substack{C_y\in[\mathcal{F}_y]^s_{-\infty},\\ D_y\in[\mathcal{F}_w]^\infty_{s-p+1+k}}} |\mathbb{P}_y(C_y\cap D_y) - \mathbb{P}_y(C_y)\mathbb{P}_y(D_y)| \\
&= \alpha_y(k-p+1) \\
&< c\rho^{k-p+1}.
\end{aligned} \tag{47}$$

which completes the proof. ∎

Remarks From Theorem (1.6.1) we can see that Assumption (1.2.1) and (1.2.2) together ensure the α-mixing property for VAR(p) models. In fact, we can not simply substitute the Gaussian assumption with a sub-Gaussian assumption as in [89]. Consider the following example given in [3]:

Let $\{\epsilon_t\}$ be a doubly infinite sequence of independent Bernoulli(q) random variables where $0 < q < 1$. The AR(1) process $\{x_i\}$ with innovation random variables $\{\epsilon_t\}$ and AR parameter $\rho \in (0, \frac{1}{2}]$ is defined by

$$x_t = \sum_{l=0}^{\infty} \rho^l \epsilon_{t-l}. \tag{48}$$

x_t satisfies Assumption (1.2.1) and is well defined. However, [3] shows that $\alpha_x(m)$ does not converge to 0 as $m \to 0$. Thus, $\{x_t\}$ is not α-mixing.

1.7 Structure of the population covariance matrix of $\{w_t\}$

Theorem 1.4.1 basically states that the sample covariance matrix $\frac{1}{T}ZZ'$ of $\{w_t\}$ inherits the RE property of the population covariance matrix Σ_w of $\{w_t\}$ under mild conditions. If we want to have good finite sample performance out of this design, there are two key conditions in Theorem 1.4.1 we need to pay attention to.

1.7.1 Bounding $\rho(s, \Sigma_w)$

By definition of $\{w_t\}$ (41), Σ_w can be written as

$$\begin{bmatrix} \Gamma_y(0) & \Gamma_y(1) & \dots & \Gamma_y(p-1) \\ \Gamma_y(-1) & \Gamma_y(0) & \dots & \Gamma_y(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_y(-p+1) & \Gamma_y(-p+2) & \dots & \Gamma_y(0) \end{bmatrix}, \tag{49}$$

where $\Gamma_y(h), h = -p+1, \dots, p-1$ are autocovariances of the original VAR(p) model $\{y_t\}$ (1) and we have $\Gamma_y(h) = \Gamma_y(-h)'$. Let

$$\|\Sigma\|_2 := \sup_{x \neq 0} \frac{\|\Sigma x\|_2}{\|x\|_2}, \tag{50}$$

be the l_2/l_2 operator norm of Σ. It is easy to see that

$$\rho(s, \Sigma_w) \leq \|\Sigma_w\|_2. \tag{51}$$

To bound $\rho(s, \Sigma_w)$ such that $\rho(s, \Sigma_w)$ does not grow with K and p, it is sufficient to bound $\|\Sigma_w\|_2$. Under Assumption (1.2.1) for every $\rho > \rho(\widetilde{A})$ where $\rho(\widetilde{A}) < 1$ is the spectral radius of $\widetilde{A}$, there must exist some constant $M \geq 1$ such that for all $k \geq 1$,

$$\|\widetilde{A}^k\|_2 \leq M\rho^k. \tag{52}$$

If $\rho(\widetilde{A})$ is bounded away from 1 and M is bounded regardless of other parameters, it is possible to universally bound $\|\Sigma_w\|_2$ as shown in the following proposition.

Proposition 1.7.1 *Let $\widetilde{A}, r_t$ be defined in (43) and $\Sigma_r = \mathbb{E}(r_t r_t')$ be the covariance matrix of r_t. If there exist some absolute constant $0 < \rho_0 < 1$ and $M_0 \geq 1$ such that*

$$\|\widetilde{A}^k\|_2 \leq M_0\rho_0^k. \tag{53}$$

we have

$$\|\Sigma_w\|_2 \leq (1 + \frac{M_0^2\rho_0^2}{1-\rho_0^2})\|\Sigma_r\|_2. \tag{54}$$

Proof Based on the Yule-Walker equations for the VAR(1) representation (42), we have

$$\Gamma_w(0) = \widetilde{A}\Gamma_w(-1) + \Sigma_r = \widetilde{A}\Gamma_w(1)' + \Sigma_r, \tag{55}$$

and

$$\Gamma_w(1) = \widetilde{A}\Gamma_w(0). \tag{56}$$

where $\Gamma_w(h), h = -1, 0, 1$ are autocovariances of $\{w_t\}$. Combining (55) and (56) we get

$$\Gamma_w(0) = \widetilde{A}\Gamma_w(0)\widetilde{A}' + \Sigma_r. \tag{57}$$

or

$$\Sigma_w = \widetilde{A}\Sigma_w\widetilde{A}' + \Sigma_r. \tag{58}$$

because $\Sigma_w = \Gamma_w(0)$. Recursively apply (58) to the left-hand side of itself,

$$\Sigma_w = \sum_{k=1}^{\infty} \widetilde{A}^k\Sigma_r(\widetilde{A}')^k + \Sigma_r. \tag{59}$$

The summation converges due to (53). Therefore

$$\|\Sigma_w\|_2 \leq \sum_{k=1}^{\infty} \|\widetilde{A}^k\|_2\|\Sigma_r\|_2\|(\widetilde{A}')^k\|_2 + \|\Sigma_r\|_2 \leq (1 + \frac{M_0^2\rho_0^2}{1-\rho_0^2})\|\Sigma_r\|_2. \tag{60}$$

as claimed. ∎

Condition (53) can be viewed as a slightly stronger (universal) version of the stability assumption (1.2.1). As long as $\|\Sigma_r\|_2$ is bounded, we can bound $\|\Sigma_w\|_2$ and thus $\rho(s, \Sigma_w)$. Notice that

$$\Sigma_r = \begin{bmatrix} \Sigma_u & 0 \\ 0 & 0 \end{bmatrix}. \tag{61}$$

Clearly, $\|\Sigma_r\|_2 = \|\Sigma_u\|_2$. There are some typical scenarios:

1. $\Sigma_u = \sigma^2 I_K$.

   Independent innovations is a common assumption. When the correlations across the innovation vectors are small, it is a good approximation to simplify problems. Under this assumption, $\|\Sigma_u\|_2 = \sigma^2$ is a constant thus bounded.

2. Σ_u is a Toeplitz matrix such that $[\Sigma_u]_{ij} = \sigma^2 d^{|i-j|}$ for some constant $0 < d < 1$ as in [65].

   Toeplitz-type covariance matrices naturally appears in spatial-temporal problems if the innovations are arranged in certain order and the correlation decays as the distance increases. $\|\Sigma_u\|_2$ can be bounded as follows:

$$\|\Sigma_u\|_2 \leq \sqrt{\|\Sigma_u\|_1 \cdot \|\Sigma_u\|_\infty} \leq \sqrt{\frac{2\sigma^2}{1-d} \cdot \frac{2\sigma^2}{1-d}} = \frac{2\sigma^2}{1-d}, \tag{62}$$

   where

$$\|\Sigma\|_1 = \max_j \sum_i |\Sigma_{ij}|, \quad \|\Sigma\|_\infty = \max_i \sum_j |\Sigma_{ij}|, \tag{63}$$

   are the l_1/l_1 and l_∞/l_∞ matrix operator norm respectively.

1.7.2 RE property of Σ_w

If Σ_w has no special underlying structure, we simply consider a stronger condition on Σ_w, that is, the minimal eigenvalue of Σ_w is bounded away from 0. We state a fact about Σ_w that requires no additional assumptions.

Proposition 1.7.2 *Let Σ_w be the population covariance matrix of $\{w_t\}$ defined in (41). Under Assumption (1.2.1), if $\Sigma_u \succ 0$, we have $\Sigma_w \succ 0$.*

Proof Obviously, $\Sigma_w \succeq 0$. Assume there exists a $Kp \times 1$ vector $v = (v_1', v_2', \ldots, v_p')'$ such that

$$0 = v'\Sigma_w v = v'\mathbf{A}\Sigma_w\mathbf{A}'v + v'\Sigma_r v \geq v'\Sigma_r v = v_1'\Sigma_u v_1, \tag{64}$$

where $v_i, i = 1, \ldots, p$ are $K \times 1$ subvectors of v. Here we use the equations (58), (61) and the fact that $v'\mathbf{A}\Sigma_w\mathbf{A}'v \geq 0$.

If $v_1 \neq 0$, we have

$$v'\Sigma_w v \geq v_1'\Sigma_u v_1 > 0, \tag{65}$$

based on the fact that $\Sigma_u \succ 0$. If $v_1 = 0$, let k be the smallest integer such that $v_k \neq 0$ and

$$\widetilde{v} = (v_k', v_{k+1}', \ldots, v_p, 0, \ldots, 0)'. \tag{66}$$

It is easy to see that

$$v'\Sigma_w v = \widetilde{v}'\Sigma_w \widetilde{v} \geq v_k'\Sigma_u v_k > 0. \tag{67}$$

Either way it leads to a contradiction, which means that $\Sigma_w \succ 0$. ∎

Proposition (1.7.2) guarantees that under weak assumptions Σ_w is positive definite and thus satisfies RE condition.

It is worth noting that $\Sigma_w \succ 0$ is not a necessary condition for $RE(s, k_0, \Sigma_w)$ to hold. Therefore, even if Σ_w is degenerated, as long as $RE(s, k_0, \Sigma_w)$ holds, Z' will satisfy the RE condition with high probability.

1.8 Order selection

All of the previous results are based on the assumption that the order p of the VAR model is known. However, p is usually not available in real application. In classical VAR model where K, p are fixed and n goes to infinity, even if we know that p is bounded by some constant p_{max} and build a VAR model with order p_{max}, the standard linear regression solution is not sparse and requires additional efforts to avoid overfitting, which could be done by select a model with smaller order to control model complexity. In the high dimensional settings, given n samples, the maximum (estimable) order of the VAR model is naturally bounded by n, that is, $p \leq n$. (Notice that the total number of coefficients, Kp, could still be larger than n.) If we estimate the model based on the Lasso or Dantzig selector, it is possible to estimate the order of the full model by the maximum of the estimated orders of the submodels. Each submodel can be estimated by common techniques such as cross validation. This approach is more likely to over-estimate the order due to the variation of the estimated submodels.

In terms of prediction power, as long as the true non-zero coefficients are correctly included in the models and false-positive variable selections are within tolerance, prediction error can be controlled. Under the assumption that a small number of

variables contain most of the information, it is more important to capture these variables instead of trying to reduce noise. Therefore, a more aggressive model order estimation can even help to get better prediction. In the next section, we also studied the variable selection performance of the Lasso method applied to VAR model.

There are several alternatives to identify the order of the VAR model listed below for reference:

1. Instead of fitting each submodel independently, we can force the tuning parameter, λ_T, to be the same across all submodels and solve the optimization problem (12) and (13) corresponding to all submodels at the same time under this constraint. With only one tuning parameter, it will reduce the variation between different submodels and thus lead to more stable order selection.

2. If we consider a more general non-stationary VAR model, that is, the inverse of Σ_w defined in (49) is a block diagonal matrix, this actually corresponds to a decomposable graphical model introduced in the next chapter. Its (block) bandwidth is the order of the underlying model. We can test its bandwidth based on the proposed procedures. It is more likely to underestimate the bandwidth because of the generality of the model.

1.9 Numerical Results

For illustrative purpose, we designed two sets of simulations to study the performance of the Lasso estimator applied to VAR models. We ran 1000 simulations for each set. For the first set of simulation, we fixed $p = 2$. For $K = n = 8, 16, \ldots, 256$, the VAR models are given as:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + u_t, \tag{68}$$

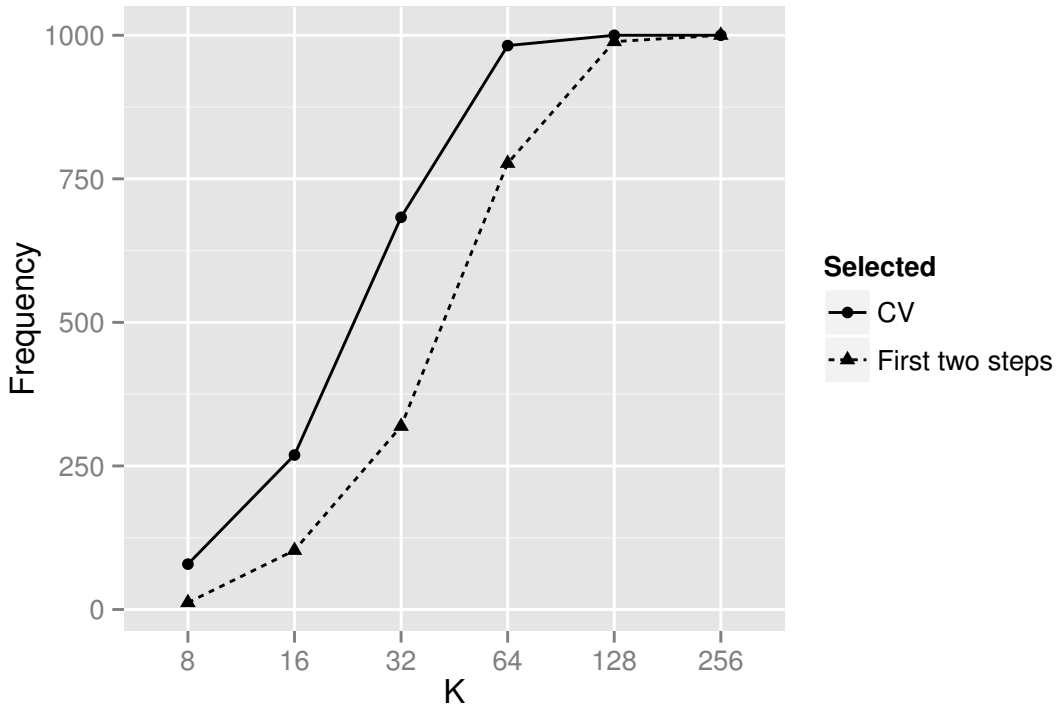

Figure 1: Variable selection results with fixed $p = 2$. The solid line is the frequency (out of 1000 simulations) with which the two non-zero coefficients were selected by cross validation. The dotted line is the frequency with which they were selected in the first two steps of Lasso iterations.

where y_t is a K-dimensional vector, $A_1 = 0.5I_K$, A_2 is a $K \times K$ matrix with

$$[A_2]_{ij} = \begin{cases} 0.4 & j = i + 1 \text{ or } (i, j) = (K, 1), \\ 0 & \text{otherwise,} \end{cases} \tag{69}$$

and u_t's are i.i.d. random vectors distributed as $N(0, I_K)$. We simulated a time series of length 10000 with initial values $y_1 = y_2 = (0, 0, \ldots, 0)'$ and used the last $n + 2$ samples to construct the $n \times 2K$ design matrix Z' and fit the model (9). We only considered the first submodel, that is, estimating the first row of B in (6). The estimation for other submodels are similar. We used the *lars* and *cv.lars* functions in the *lars* package in R to fit the Lasso model and did model selection based on the leave-one-out cross validation with all parameters set to default except for *use.Gram=False*. Figure 1 shows the variable selection performance of Lasso in the fixed-p scenario.

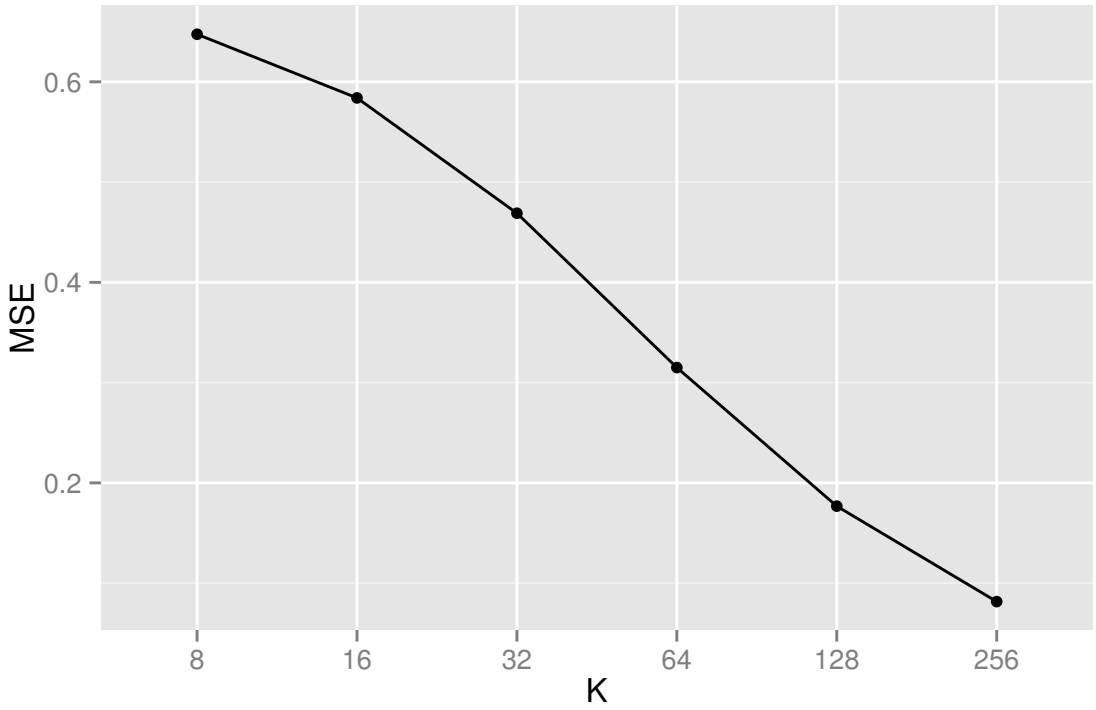

Figure 2: Mean squared error of the Lasso estimator based on cross validation with fixed $p = 2$.

As n and K increase together, the probability that the two non-zero coefficients is selected either in the model that is selected by cross validation or within the first two steps of the Lasso iterations converges to 1 quickly. When $n = K \geq 128$, it is almost certain that the true non-zero coefficients will be included in the model. Figure 2 shows the mean squared error (MSE) of the Lasso estimator. As the number of variables increases, the MSE decreases significantly and approaches zero. It agrees with the variable selection results as estimation bias is reduced if the true non-zero coefficients can be identified correctly.

For the second set of simulation, we fixed $K = 2$. For $p = n = 8, 16, \ldots, 256$, the VAR models are given as:

$$y_t = A_1 y_{t-1} + A_p y_{t-p} + u_t, \tag{70}$$

where y_t is a 2-dimensional vector, $A_1 = 0.5 I_2$,

$$A_p = \begin{bmatrix} 0 & 0.4 \\ 0.4 & 0 \end{bmatrix} \tag{71}$$

and u_t's are i.i.d. random vectors distributed as $N(0, I_2)$. We simulated a time series of length 10000 with initial values $y_1 = y_2 = \ldots = y_p = (0, 0, \ldots, 0)'$ and used the last $n + p$ samples to construct the $n \times 2p$ design matrix Z'. We estimate the first submodel based on the same method described above. The results are plotted in Figure 3 and 4. The patterns of the curves are very similar to the previous results. The variable selection performance and MSE get better as n and p increase and we have a high probability to include critical variables when $n = p \geq 128$.

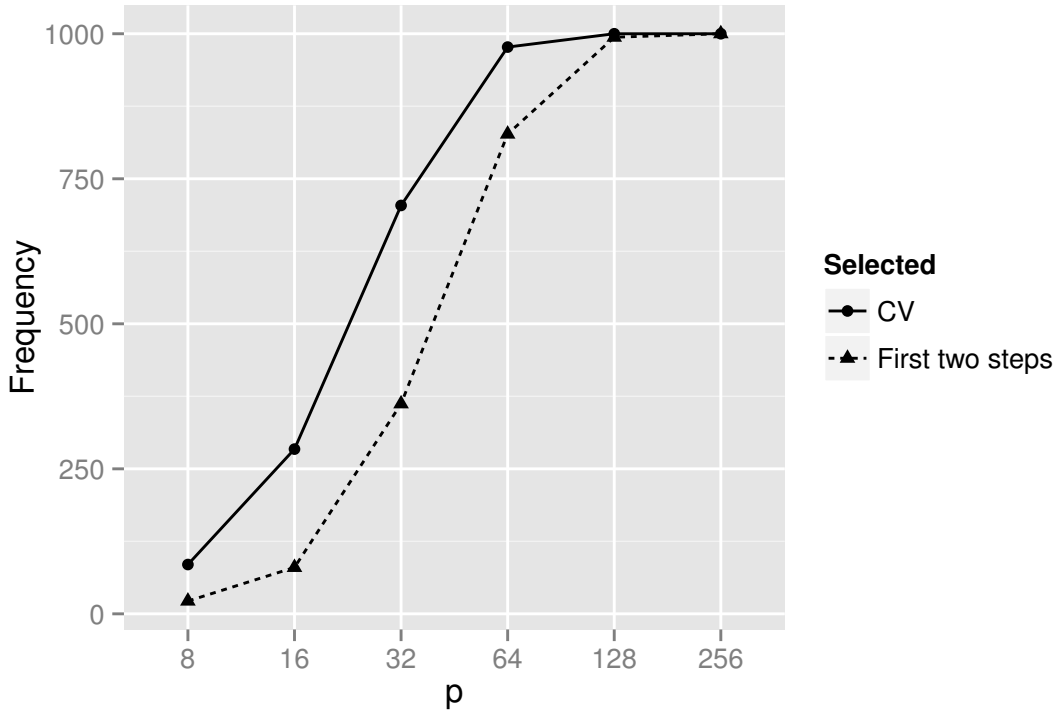

Figure 3: Variable selection results with fixed $K = 2$. The solid line is the frequency (out of 1000 simulations) with which the two non-zero coefficients were selected by cross validation. The dotted line is the frequency with which they were selected in the first two steps of Lasso iterations.

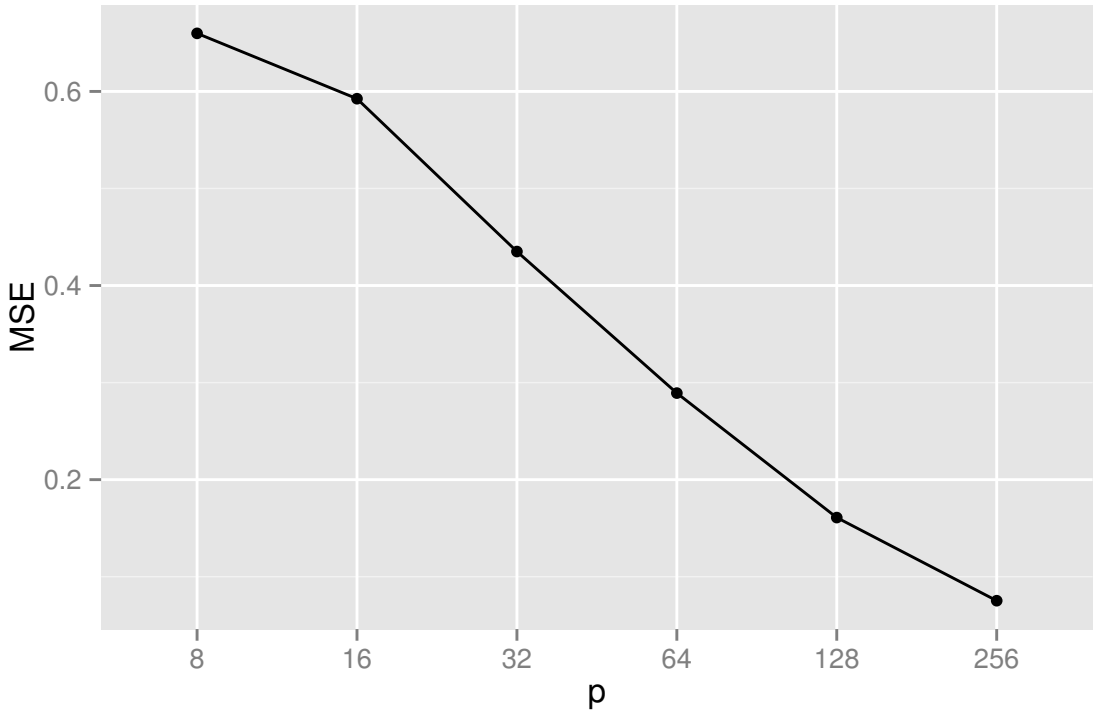

Figure 4: Mean squared error of the Lasso estimator based on cross validation with fixed $K = 2$.

CHAPTER II

DECOMPOSABLE STRUCTURE TEST FOR GAUSSIAN GRAPHICAL MODELS

2.1 Introduction

With the fast growth of technology, high-dimensional statistical inference has drawn a lot of interest recently. One of the most important problem is to estimate the covariance matrix or the inverse covariance matrix (concentration matrix). Usually it is impossible to accurately estimate such matrices when the dimension of data is much larger than the sample size because of the bad behavior of the empirical spectral distribution (See [47] and [6]). However, under certain assumption of the matrices, e.g. sparsity, bandedness or other structures, we are still able to construct good estimators out of the limited information.

[85] and [7] proposed sparse concentration matrix estimation via graphical lasso. [83], [10], [42] and [13] suggested covariance matrix or concentration matrix regularization by banding or tapering, which is demonstrated to have satisfactory performance for certain "bandable" classes of matrices. A strictly banded concentration matrix is a special case of the more general class of concentration matrices corresponding to the decomposable (or chordal) graphical models. Given the decomposable structure of a concentration matrix, [40] derived its maximum likelihood estimator (MLE) when sample size is larger than the maximum size of cliques. This result is quite inspiring because it provides us with an simple explicit estimator in high-dimensional settings as long as the cliques of the underlying graphical structure are not too large. [64] proposed a class of Bayes estimators for decomposable Gaussian graphical models that allows for flexible shrinkage. Even for non-decomposable graphical models, [19]

suggested decomposable graph embedding in general covariance selection problems to reduce the computational complexity. Decomposable graphical models are proved to be widely applicable, easy to use and computation-friendly. To support such structural assumptions, [40] proposed a likelihood ratio test (LRT) based on the MLE estimator. But the failure of the LRT in high-dimensional settings motivates us to develop new testing methods.

Without loss of generality, consider a general framework of covariance matrix testing problems: Let $x_1, \ldots, x_n$ be i.i.d. copies from a common multivariate normal distribution $N_p(\mu, \Sigma)$, where $\Sigma \succ 0$ is a positive definite covariance matrix. We want to test the structure of $\Theta = \Sigma^{-1}$, which leads to the hypothesis testing problem:

$$H_0 : \Theta \in \mathcal{M}_0 \;\leftrightarrow\; H_1 : \Theta \notin \mathcal{M}_0, \tag{72}$$

where $\mathcal{M}_0$ is a specified family of matrices. For example, if we want to test the complete independence of the variables, then $\mathcal{M}_0 = \{M : M = \mathrm{diag}(d_1, d_2, \ldots, d_p), d_i > 0 \text{ for } 1 \leq i \leq p\}$ where $\mathrm{diag}(d_1, d_2, \ldots, d_p)$ is a $p \times p$ diagonal matrix with diagonal entries $d_1, d_2, \ldots, d_p$.

Classical testing procedures for covariance matrices usually assume that the dimension p of the matrix, i.e. the number of variables, is fixed and the sample size n goes to infinity (See [2] and [55] for a general review). However, such procedures may not perform well or even be applicable when p is comparable in magnitude to n. This is mainly because (1) if $p > n$, the covariance matrix degenerates and likelihood ratio tests no longer work; and (2) [47] showed that even if p/n converges to some constant $c \in (0, 1)$, the empirical spectral distribution may not necessarily converge to the spectral distribution of the underlying covariance matrix.

To overcome these difficulties, [41] proposed some statistics based on the eigenvalues of the sample covariance matrix to test sphericity or equality to a given matrix, which is motivated by the work of [34] and [56]. [67] and [71] further extended it to test complete independence of the variables and the latter author also investigated

the non-null distributions of these statistics. [45] introduced a test for group independence and [68] a test for the equality of covariance matrices. [18] proposed tests based on new estimators for $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$ and dropped the normality assumption as well as the explicit relationship between p and n. [62] extended it to test bandedness of high-dimensional covariance matrices. Some tests that are not based on the eigenvalues of the sample covariance matrix are also available. [5] proposed corrections to LRT when $p/n \to c \in (0, 1)$ which utilizes results from random matrix theory (See [6]). [12] extended the work of [33] and suggested statistics based on the largest absolute value of the "off band" correlation coefficients to test bandedness. However, none of the above mentioned works is able to test the graphical structure, i.e. the sparsity pattern of the corresponding concentration matrices.

In this chapter we introduce some test procedures for a given decomposable structure of a Gaussian graphical model, i.e., $\mathcal{M}_0$ is the set of all possible concentration matrices that adopt such structure. First, we give some introduction of Gaussian graphical models and decomposable graphs in Sections 2.2 and 2.3. In Section 2.4, we propose our test statistic decomposable structure test and study its asymptotic behavior in the group independence test case. We also suggest a simulation-based procedure for testing an arbitrary decomposable structure. In Section 2.5, we analyze the computational complexity of the testing procedures and propose several approaches to facilitate the computation. Finally, we illustrate the performance of the test procedures by simulation and give an example based on real data in Section 2.6.

2.2 Gaussian Graphical Models

A Gaussian graphical model describes the conditional independence structure for a Gaussian random vector. Let $X := (X_1, X_2, \ldots, X_p)' \in \mathbb{R}^p$ be a Gaussian random vector distributed as $N_p(\mu, \Sigma)$ where $\Sigma \succ 0$. Let $G := (V, E)$ be an undirected graph

where $V := \{1, 2, \ldots, p\}$ is the vertex set such that vertex i represents the random variable X_i and the edge set $E \subset V \times V$ consists of pairs of distinct vertices, (i, j) where $1 \leq i < j \leq p$, that describes the graphical structure of X. X is said to be Markov with respect to G if for any edge $1 \leq i < j \leq p$ and $(i, j) \notin E$, the (i, j) and (j, i) entries of the concentration matrix $\Theta := \Sigma^{-1}$ are zero, which is also equivalent to the independence of X_i and X_j conditional on all other variables. The graph G is not only a good visualization of the underlying model, but also a good representation of the data structure that helps people understand how different variables interact with each other. For example, from social network data we might be able to infer the diffusion of information on the internet, either between users or between websites.

Many approaches have been proposed to extract structural information and improve the estimation of the covariance or concentration matrix. To name a few, [50] proposed neighborhood selection with the Lasso, that is, for each variable v, they fit a Lasso model against all other variables and select its neighbors based on the variable selection result of the Lasso by solving

$$\underset{\beta:\beta_v=0}{\arg\min} \left\{ \frac{1}{n} \|X_v - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}. \tag{73}$$

where X_v is the data column corresponds to the variable v. With properly chosen regularization parameter λ, they obtained consistency results for their estimator. [85] suggested solving the following optimization problem:

$$\underset{\Theta \succ 0}{\arg\min} \left\{ -\log|\Theta| + \operatorname{tr}(\Theta\hat{\Sigma}) + \lambda \sum_{1 \leq i \neq j \leq p} \sum |\Theta_{ij}| \right\}, \tag{74}$$

where $\hat{\Sigma}$ is the MLE of the covariance matrix Σ. This formulation is derived from minimizing the log-likelihood function subject to l_1 constraints on the magnitude of non-diagonal entries of Θ. They also derived its asymptotic distribution.

Given the estimated graphical model, a natural question to ask is that how much confidence we have in it. There are not many results in the literature regarding how to test a graphical model, especially in the high-dimensional settings. Most of them

focus on tests for the covariance matrix instead of inverse covariance matrix, which inspires us to develop tests for general graphical models and concentration matrices.

2.3 Decomposable Graphs

Decomposable graphs are a family of graphs that have unique properties in terms of interpretation and estimation. In general, given the assumption that Θ is Markov with respect to an arbitrary graph $G := \{V, E\}$, there is no closed-form expression of the MLE of Θ and it needs to be calculated by an iterative proportional scaling algorithm as shown in [40]. However, if G is a decomposable graph, which we will define below, a simple and computationally efficient MLE is available.

Following the definitions in [40], let $G_A = (A, E_A)$ denote the subgraph of G where $E_A = \{(i, j) \in E : i, j \in A\}$. A complete graph is a graph that every pair of vertices are directly connected by an edge. A subset $C \subseteq V$ is called a clique if G_C is complete and for any $A \subseteq V$ such that $C \subset A$, A is not complete. We say that B separates A_1 from A_2 if for every $a_1 \in A_1, a_2 \in A_2$, all paths from a_1 to a_2 intersect B.

Definition 2.3.1 (Decomposition of graph) *A decomposition of an undirected graph G is a triple (A_1, B, A_2) such that $A_1, A_2, B \subseteq V$ are disjoint, B is complete, and B separates A_1 and A_2.*

Definition 2.3.2 (Decomposable graph) *A decomposable graph is an undirected graph that is either complete or if it has a decomposition (A_1, B, A_2) such that both the subgraphs $G_{A_1 \cup B}$ and $G_{A_2 \cup B}$ are decomposable.*

Definition (2.3.2) is recursive and thus not convenient for our study. It has an equivalent form that is easier to use based on a perfect sequence of cliques of G as proposed in [40]. See Definition (2.3.3) and Proposition (2.3.4).

Definition 2.3.3 (Perfect sequence) *For a sequence $A_1, A_2, \ldots, A_m \subseteq V$ of an undirected graph G, define*

$$P_i = A_1 \cup A_2 \cup \cdots \cup A_i, \quad S_i = P_{i-1} \cap A_i. \tag{75}$$

The sequence is said to be perfect if S_j is complete for every $j \geq 1$ and there exists $i < j$ for every $j > 1$ such that $S_j \subseteq A_i$.

Proposition 2.3.4 *The graph G is decomposable if and only if there exists a perfect sequence consisting of all cliques of G.*

For any decomposable graph G, suppose that its cliques form a perfect sequence $C_1, C_2, \ldots, C_m$. Let P_i, S_i be similarly defined as in (75) based on $C_1, C_2, \ldots, C_m$. For every $i = 2, \ldots, m$, S_i separates $P_{i-1} \backslash S_i$ from $C_i \backslash S_i$, and hence $(P_{i-1} \backslash S_i, S_i, C_i \backslash S_i)$ decomposes G_{P_i}, the subgraph of G induced by P_i. The S_i's are not necessarily distinct. For example, if G is a star graph, every S_i is the set that consists of the center vertex and thus identical. See Figure 5 for some examples of decomposable and non-decomposable graphs.

Suppose we have n samples, $x_1, x_2, \ldots, x_n$, of X. The standard MLE of the mean and covariance matrix of X when $n > p$ is

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})', \tag{76}$$

For an arbitrary subset $A \subseteq V$ and $x \in \mathbb{R}^{|V|}$, let x_A denote the subvector of x confined to A. For two arbitrary subsets $A, B \subseteq V$ and a $|V| \times |V|$ matrix M, let $M_{AB} = \{m_{\alpha\beta}\}_{\alpha \in A, \beta \in B}$ denote the $|A| \times |B|$ submatrix of M. For a $|A| \times |B|$ matrix $M = \{m_{\alpha\beta}\}_{\alpha \in A, \beta \in B}$ indexed by $A, B \subseteq V$, let $[M]^V$ denote the $|V| \times |V|$ matrix defined by

$$([M]^V)_{\alpha\beta} = \begin{cases} m_{\alpha\beta} & \text{if } \alpha \in A, \beta \in B \\ 0 & \text{otherwise.} \end{cases} \tag{77}$$

The following proposition from [40] gives the MLE of a concentration matrix Θ that is Markov with respect to a decomposable graph G.

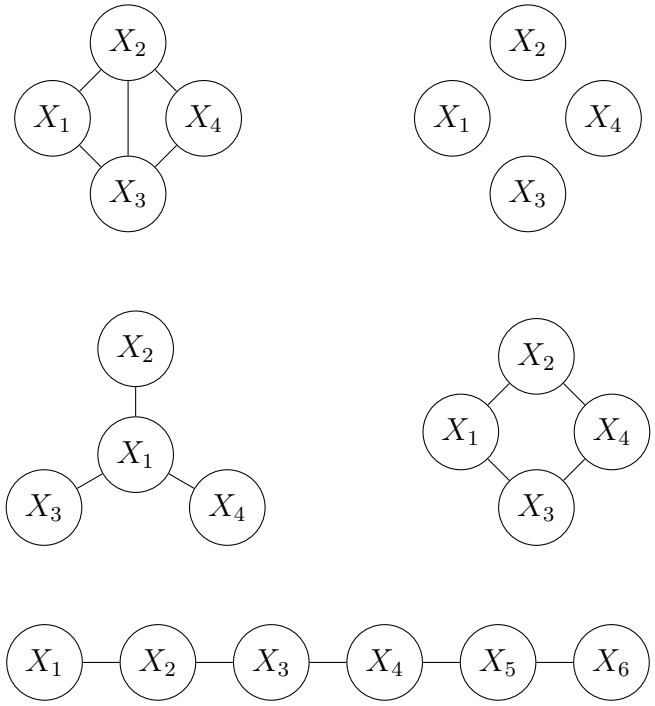

Figure 5: (1) Top left: decomposable graph, corresponding to a 4×4 concentration matrix Θ_1 with $[\Theta_1]_{14} = [\Theta_1]_{41} = 0$, its perfect sequence of cliques is $\{1, 2, 3\}$,$\{2, 3, 4\}$; (2) Top right: decomposable graph, corresponding to a diagonal concentration matrix which implies complete independence of the variables, its perfect sequence of cliques is $\{1\}$,$\{2\}$,$\{3\}$,$\{4\}$; (3) Middle left: decomposable graph, a star graph, its perfect sequence of cliques is $\{1, 2\}$,$\{1, 3\}$,$\{1, 4\}$; (4) Middle right: non-decomposable graph, its cliques are $\{1, 2\}$,$\{1, 3\}$,$\{2, 4\}$ and $\{3, 4\}$; (5) Bottom: decomposable graph, corresponding to a tridiagonal concentration matrix, its perfect sequence of cliques is $\{1, 2\}$,$\{2, 3\}$,$\{3, 4\}$,$\{4, 5\}$,$\{5, 6\}$.

Proposition 2.3.5 *If G is a decomposable graph and its cliques form a perfect sequence $C_1, C_2, \ldots, C_m$ with the corresponding separators $S_2, S_3, \ldots, S_m$ and the concentration matrix Θ is Markov with respect to G, the MLE of the mean vector μ and Θ exists with probability one if and only if $n > \max_{1 \le i \le m} |C_i|$. It is then given as*

$$\hat{\mu}_0 = \bar{x}, \quad \hat{\Theta}_0 = \sum_{i=1}^{m} [(\hat{\Sigma}_{C_i C_i})^{-1}]^V - \sum_{i=2}^{m} [(\hat{\Sigma}_{S_i S_i})^{-1}]^V. \tag{78}$$

where $\hat{\Sigma}$ is the standard MLE of Σ. The determinant of $\hat{\Theta}_0$ can be calculated as

$$\det \hat{\Theta}_0 = \frac{\prod_{i=2}^{m} \det \hat{\Sigma}_{S_i S_i}}{\prod_{i=1}^{m} \det \hat{\Sigma}_{C_i C_i}}. \tag{79}$$

Based on Proposition (2.3.5), we can independently calculate the MLE of the concentration matrix of each clique and separator and then combine them together to get the MLE of the full model. It is worth noting that the sample size n is only required to be greater than the size of the largest clique. It could be much smaller than the number of variables p. For example, when G is a tree, every clique is of size 2, thus $n = 3$ suffices for a non-degenerated MLE of Θ. When G is the graph corresponding to a banded concentration matrix Θ, every clique is of size $2b + 1$ where b is the bandwidth of Θ such that $\Theta_{ij} = 0$ if $|i - j| > b$. In this case, for a fixed bandwidth b, the required sample size is only $2b + 2$.

2.4 The Test Statistic

In the hypothesis testing problem (72), let $\mathcal{M}_0$ represent the set of concentration matrices that are Markov with respect to a given decomposable graph G_0. In a classical setting when n is large compared to p, [40] suggests an exact deviance test that is based on the statistic

$$b = \frac{\det \hat{\Theta}_0}{\det \hat{\Theta}}, \tag{80}$$

where $\hat{\Theta}_0$ is the MLE of Θ under H_0 as defined in (78) and $\hat{\Theta} = \hat{\Sigma}^{-1}$ is the inverse of the standard MLE of the covariance matrix under no structural assumption. They

showed that b is distributed as the product of some independent Beta random variables.

However, in a high-dimensional setting where p is larger than n, $\hat{\Sigma}$ is singular and $\hat{\Theta}$ does not exist. Many new test statistics for high-dimensional covariance or concentration matrices are developed based on the eigenvalues of sample covariance matrix with the general form

$$t = \mathrm{tr}[(\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}} - I_p)^2], \tag{81}$$

where $\hat{\Sigma}$ is the standard MLE of the covariance matrix and $\hat{\Theta}_0$ denotes the estimator of the concentration matrix that corresponds to the specific testing problem. It usually does not involve the inversion of the sample covariance matrix when calculating $\hat{\Theta}_0$ and thus does not degenerate. For example, [41] gave a statistic for testing equality to a given matrix Σ_0 which is essentially equivalent to setting $\hat{\Theta}_0 = \Sigma_0^{-1}$ and a statistic for testing sphericity equivalent to setting

$$\hat{\Theta}_0 = [(\frac{\mathrm{tr}(\hat{\Sigma})}{p})I_p]^{-1} = \frac{p}{\mathrm{tr}(\hat{\Sigma})}I_p. \tag{82}$$

[67] introduced a test for complete independence equivalent to setting

$$\hat{\Theta}_0 = \mathrm{diag}\{\hat{\Sigma}_{11}^{-1}, \hat{\Sigma}_{22}^{-1}, \ldots, \hat{\Sigma}_{pp}^{-1}\}. \tag{83}$$

The intuition behind all of these statistics is clear. If H_0 is true, $\hat{\Theta}_0$ is a good estimator of Θ and thus $\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}}$ should be close to I_p. The statistic t in (81) measures the goodness of fit under H_0 by the Frobenius norm of the difference matrix $\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}} - I_p$. It also has a strong connection with the statistic b in (80) in the sense that

$$b^{-1} = \det(\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}}) = \det[I_p + (\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}} - I_p)] \approx 1 + \mathrm{tr}(\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}} - I_p), \tag{84}$$

if $\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}}$ is sufficiently close to I_p. In fact, [34] proves that the sphericity test is the locally most powerful test and invariant under several families of transformations.

Let $x_{1k}, x_{2k}, \ldots, x_{n_k k}$ be i.i.d. copies of $X_k \sim N_{p_k}(\mu_k, \Sigma_k)$ where $\Sigma_k \succ 0$. Let $\Theta_k =$

Σ_k^{-1}. Consider a series of testing problems

$$H_{0k} : \Theta_k \in \mathcal{M}_{0k} \;\leftrightarrow\; H_{1k} : \Theta_k \notin \mathcal{M}_{0k}, \tag{85}$$

where $\mathcal{M}_{0k}$ is the family of matrices that is Markov with respect to a decomposable graph G_{0k}. Our statistic is given as

$$t_k = \mathrm{tr}[(\hat{\Theta}_{0k}^{\frac{1}{2}} \hat{\Sigma}_k \hat{\Theta}_{0k}^{\frac{1}{2}} - I_{p_k})^2]. \tag{86}$$

where $\hat{\Sigma}_k$ is the standard MLE of Σ_k and $\hat{\Theta}_{0k}$ is defined in (78) based on $\hat{\Sigma}_k$ and G_{0k}. We make the following assumptions to study the asymptotic behavior of our statistic.

Assumption 2.4.1 (Asymptotics of p_k and n_k)

$$\lim_{k\to\infty} p_k = +\infty, \quad \lim_{k\to\infty} n_k = +\infty, \quad \lim_{k\to\infty} p_k/n_k = \gamma, \tag{87}$$

where $\gamma > 0$ is a absolute constant.

We only require that p_k and n_k grow linearly with each other, which means that p_k can grow faster than n_k.

Assumption 2.4.2 (Bounded clique size) *For each k, G_{0k} is a decomposable graph with a perfect sequence of cliques $C_{1k}, C_{2k}, \ldots, C_{m_k k}$ and the corresponding separators $S_{2k}, S_{3k}, \ldots, S_{m_k k}$ with $\max_{1\le i\le m_k} |C_{ik}| < K$ where $K > 0$ is an absolute constant independent of p_k and n_k. Furthermore, $n_k > K$ for all k.*

Some typical graphs that satisfy Assumption (2.4.2) include forests and graphs corresponding to a banded matrix with bounded bandwidth.

2.4.1 Testing the group independence and complete independence

To test the group independence of X, G_0 is a graph consisting of disjointed cliques that represent the groups, that is, each $C_i, i = 1, 2, \ldots, m$ represents one group of

variables with $|C_i| < K$ under Assumption (2.4.2) and $S_i = \emptyset$ for $i = 2, 3, \ldots, m$. In this case, $\hat{\Theta}_0$ is a block diagonal matrix

$$\hat{\Theta}_0 = \text{diag}\{\hat{\Sigma}^{-1}_{C_1C_1}, \hat{\Sigma}^{-1}_{C_2C_2}, \ldots, \hat{\Sigma}^{-1}_{C_mC_m}\}. \tag{88}$$

The following theorem gives the asymptotic null distribution of the test statistic t for a group independence test.

Theorem 2.4.3 *Let t_k be the test statistic in (86) of the testing problem (85). Under Assumption (2.4.1) and (2.4.2), if $C_{1k}, C_{2k}, \ldots, C_{m_kk}$ are disjointed and the null hypothesis is true for every k,*

$$t_k - 2\sum_{l=2}^{m_k}\sum_{i=1}^{l-1}\frac{|C_{ik}||C_{lk}|}{n_k - 1} \to N(0, 4\gamma^2), \tag{89}$$

in distribution as $k \to \infty$.

See appendix for the proof of Theorem (2.4.3). It also reveals the invariance property of the tests as stated in the following corollary.

Corollary 2.4.4 *Under the same setting as in Theorem (2.4.3), the testing problem (85) is invariant under non-singular linear transformations within each group of variables,*

$$G = \{\text{diag}\{O_{|C_{1k}|}, O_{|C_{2k}|}, \ldots, O_{|C_{m_kk}|}\} | O_d \in O(d)\}, \tag{90}$$

where $O(d)$ is the real orthogonal group of dimension d.

When each clique is of size 1, we are actually testing the complete independence of the variables. In this case, our statistic reduces to the statistic introduced in [67], as shown in the following corollary.

Corollary 2.4.5 *Under the same setting as in Theorem (2.4.3), if in addition we have $|C_{ik}| = 1$ for all k and $1 \le i \le m_k$,*

$$t_k - \frac{p_k(p_k - 1)}{n_k - 1} \to N(0, 4\gamma^2), \tag{91}$$

in distribution as $k \to \infty$.

2.4.2 Testing an arbitrary decomposable structure

If G_0 is an arbitrary decomposable graph, a simple asymptotic result for t is not available. Instead, we suggest the following simulation-based testing procedure for the hypothesis testing problem (72) where $\mathcal{M}_0$ is the set of concentration matrices that are Markov with respect to G_0:

1. Calculate $\hat{\Sigma}$, $\hat{\Theta}_0$ and t as in Proposition (2.3.5) and (81).

2. Generate n samples, $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$, from the normal distribution $N(0, \hat{\Theta}_0^{-1})$. Calculate $\tilde{t}$ based on the generated samples similarly to Step 1.

3. Repeat Step 2 N times and obtain $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_N$. Determine the critical value t_0 of the test by computing the $(1-\alpha)$ cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, N$ where α is the significance level. Reject the null hypothesis if $t > t_0$.

The intuition behind this method is that if H_0 is true, $\hat{\Theta}_0$ should be a good proxy of Θ and hence $\hat{\Theta}_0^{-1}$ a good proxy of Σ assuming that Σ is not extremely ill-conditioned. Then the samples generated in Step 2 roughly represents the true distribution and the corresponding test statistics can be used to approximate the null distribution of t.

For the group independence test, the simulation-based procedure is also applicable. It can reduce the bias introduced by the normal approximation and yield a more accurate size and better power with additional computational cost. When computation is not an issue, we suggest always using the simulation-based procedure to test group independence. Some simulation results are given in the last section to compare the size and power of both testing procedures from a more visible perspective.

2.5 Computational Complexity Analysis

For $p \times p$ dense non-singular matrices, the matrix multiplication and inversion operations have complexity up to $O(p^3)$. To calculate the power of a $p \times p$ diagonalizable matrix, it also takes $O(p^3)$ operations. As p and n go to infinity, the computational cost of the testing procedures is not negligible, especially for the simulation-based procedure. Therefore, it is important to understand the complexity of the calculation. Notice that under Assumption (2.4.1), we have $n \sim O(p)$. For notational convenience, all of the following analysis will be based on p.

Recall that our statistic is

$$t = \mathrm{tr}[(\hat{\Theta}_0^{\frac{1}{2}} \hat{\Sigma} \hat{\Theta}_0^{\frac{1}{2}} - I_p)^2]. \tag{92}$$

To calculate t, we need to calculate the MLE, $\hat{\Sigma}$, of the covariance matrix first, which takes $O(p^3)$ operations. Under assumption (2.4.2), where the clique size is bounded, the calculation of $\hat{\Theta}_0$ has complexity up to $O(K^3 p) \sim O(p)$. After taking the square root ($O(p^3)$ operations), $\hat{\Theta}_0^{\frac{1}{2}}$ becomes a dense matrix. The rest of the calculation is basic matrix operation which has complexity up to $O(p^3)$, e.g., matrix multiplication. Hence, the computational complexity is $O(p^3)$ in total. For the simulation-based procedure, it is easy to see that the computational complexity is $O(Np^3)$.

There are several ways that we can accelerate the computation.

- Parallelization

  For simulation-based testing procedures, each run of the simulation is independent. If a computer cluster is available, we can simply divide the simulation into smaller jobs and aggregate the results afterwards. In fact, all of the simulations in Section 2.6 were done in parallel. It is worth noting that the recent development in GPU computing technology also opens new possibilities for high performance matrix computation [58].

- Avoid dense matrix operations

We know that $\hat{\Theta}_0$ is a sparse matrix with only $O(K^2p)$ non-zero entries compared to $O(p^2)$ in the dense matrix case. However, $\hat{\Theta}_0^{\frac{1}{2}}$ is not necessarily sparse. A simple trick to avoid the computation of $\hat{\Theta}_0^{\frac{1}{2}}$ is to notice that t can be reformulated as

$$t = \mathrm{tr}[(\hat{\Theta}_0\hat{\Sigma} - I_p)^2]. \tag{93}$$

Now the multiplication $\hat{\Theta}_0\hat{\Sigma}$ only requires $O(K^2p^2)$ operations which is much better than directly calculating $\hat{\Theta}_0^{\frac{1}{2}}\hat{\Sigma}\hat{\Theta}_0^{\frac{1}{2}}$.

- Normal approximation

  In general, to test an arbitrary decomposable structure, simulation-based procedures are expensive. An alternative is to approximate the null distribution of t by a normal distribution and just estimate its mean and variance, that is, to replace Step 3 with the following procedure.

  3b. Repeat Step 2 by N times and obtain $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_N$. Calculate their sample mean $\hat{\mu}_t$ and (unbiased) sample covariance $\hat{\sigma}_t^2$. Determine the critical value t_0 of the test by computing the $(1-\alpha)$ cutoff point of the normal distribution $N(\hat{\mu}_t, \hat{\sigma}_t^2)$. Reject the null hypothesis if $t > t_0$.

  The accuracy of the normal approximation for an arbitrary decomposable graph is not clear. But it nevertheless serves the purpose when the computational resource is limited and we still want to get a basic idea of the data.

As we can see, even a small algorithm change can lead to large performance gains. It is usually beneficial to develop computation-aware methods in the high-dimensional settings to handle large data sets and enable fast iteration in modeling.

Table 1: Size of the proposed test for group independence under H_0 based on normal approximation. Σ_0 is given in (94). H_0 is rejected when $t - \frac{16m(m-1)}{n-1}$ exceeds the 95% cutoff point of the normal distribution $N(0, 4p^2/n^2)$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.000	0.001	0.007	0.011	0.023	0.028
16	0.000	0.010	0.020	0.028	0.038	0.037
32	0.000	0.015	0.023	0.023	0.055	0.033
64	0.002	0.018	0.031	0.034	0.039	0.048
128	0.001	0.013	0.044	0.035	0.041	0.051
256	0.003	0.013	0.029	0.039	0.043	0.047

2.6 Numerical Results

2.6.1 Simulation Results

The performance of the proposed testing procedures was studied by several sets of simulations for $p, n = 8, 16, \ldots, 256$. For each combination of p and n we run 1000 simulations.

In the first set of simulations, we studied the size of the test for independence of groups of variables based on t_k (or t) defined in (86) by normal approximation. Each group contains 4 variables and there are $m = p/4$ groups in total. The covariance matrix Σ_0 is a block diagonal matrix given as

$$\Sigma_0 = \text{diag}\{\Sigma_{10}, \Sigma_{20}, \ldots, \Sigma_{m0}\}, \tag{94}$$

where $\Sigma_{i0} = I_4 + \frac{1}{2}\vec{1}_4\vec{1}_4'$ and $\vec{1}_4 = (1, 1, 1, 1)'$ for all $1 \leq i \leq m$. $x_1, x_2, \ldots, x_n$ are i.i.d. samples generated from the normal distribution $N_p(0, \Sigma_0)$. H_0 is rejected when $t - \frac{16m(m-1)}{n-1}$ exceeds the 95% cutoff point of the normal distribution $N(0, 4p^2/n^2)$. The results are reported in Table 1. We can see that the size of the test approaches 5% as p and n grow with each other.

In the second set of simulations, we studied the power of the test for independence of groups of variables based on t by normal approximation. The alternative is to set

Table 2: Power of the proposed test for group independence against alternative based on normal approximation. The population covariance matrix Σ_1 is given in (95). H_0 is rejected when $t - \frac{16m(m-1)}{n-1}$ exceeds the 95% cutoff point of the normal distribution $N(0, 4p^2/n^2)$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.00	0.01	0.05	0.24	0.75	0.99
16	0.00	0.06	0.36	0.90	1.00	1.00
32	0.00	0.28	0.86	1.00	1.00	1.00
64	0.03	0.66	1.00	1.00	1.00	1.00
128	0.19	0.94	1.00	1.00	1.00	1.00
256	0.44	0.99	1.00	1.00	1.00	1.00

the population covariance matrix

$$\Sigma_1 = \Sigma_0 + \frac{1}{2}\vec{1}_p\vec{1}'_p, \tag{95}$$

where Σ_0 is defined in (94) and $\vec{1}_p$ is a length p vector of all ones. The results are reported in Table 2. The power of the test approaches 1 as p and n go to infinity. For small p and n, the test does not have much power.

In the third set of simulations, we studied the size of the test for independence of groups of variables based on t by simulation. Under the same setting as in the first set of simulations, H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of the simulated $\tilde{t}_i, i = 1, 2, \ldots, 1000$ and we simulated 1000 tests to compute the size. The results are reported in Table 3. The approximation by simulation is more accurate than normal approximation. For every pair of p and n, we roughly achieved a size of 5%.

In the fourth set of simulations, we studied the power of the test for independence of groups of variables based on t by simulation. Under the same setting as in the second set of simulations, H_0 is rejected by the same criterion as in the third set of simulations. The results are reported in Table 4. The power of the test is consistently better than that of the normal approximation based test at the cost of increased size and computational complexity.

Table 3: Size of the proposed test for group independence under H_0 based on simulation. The population covariance matrix Σ_0 is given in (94). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.054	0.053	0.049	0.045	0.054	0.058
16	0.054	0.046	0.051	0.043	0.048	0.052
32	0.039	0.050	0.049	0.056	0.061	0.055
64	0.058	0.049	0.051	0.058	0.067	0.049
128	0.058	0.045	0.034	0.038	0.044	0.048
256	0.051	0.053	0.047	0.050	0.050	0.057

Table 4: Power of the proposed test for group independence against alternative based on simulation. The population covariance matrix Σ_1 is given in (95). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.06	0.10	0.18	0.46	0.85	1.00
16	0.07	0.17	0.52	0.92	1.00	1.00
32	0.12	0.51	0.90	1.00	1.00	1.00
64	0.27	0.78	0.99	1.00	1.00	1.00
128	0.46	0.96	1.00	1.00	1.00	1.00
256	0.73	0.99	1.00	1.00	1.00	1.00

Table 5: Size of the proposed test for bandedness of the concentration matrix under H_0 based on simulation. The population covariance matrix Σ_0 is given in (96). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.048	0.052	0.050	0.056	0.052	0.049
16	0.044	0.045	0.057	0.051	0.044	0.067
32	0.043	0.051	0.051	0.049	0.044	0.056
64	0.049	0.055	0.049	0.059	0.049	0.060
128	0.045	0.048	0.048	0.044	0.035	0.061
256	0.049	0.038	0.048	0.048	0.052	0.051

Table 6: Power of the proposed test for bandedness of the concentration matrix against alternative based on simulation. The population covariance matrix Σ_1 is given in (96). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.12	0.43	0.83	0.99	1.00	1.00
16	0.24	0.77	0.99	1.00	1.00	1.00
32	0.42	0.97	1.00	1.00	1.00	1.00
64	0.75	1.00	1.00	1.00	1.00	1.00
128	0.94	1.00	1.00	1.00	1.00	1.00
256	0.98	1.00	1.00	1.00	1.00	1.00

In the fifth and sixth sets of simulations, we studied the size and power of the test for bandedness of the concentration matrix. Similarly to the previous simulations, Σ_0 and Σ_1 are given as:

$$\Sigma_0 = \Theta_0^{-1}, \quad \Sigma_1 = \Sigma_0 + \frac{1}{2}\vec{1}_p\vec{1}_p', \tag{96}$$

where

$$[\Theta_0]_{ij} = \begin{cases} 1 & i = j \\ u_k & (i,j) = (k, k+1) \text{ or } (k+1, k),\ 1 \le k \le p-1 \\ 0 & \text{otherwise,} \end{cases} \tag{97}$$

for $1 \le i, j \le p$ and u_k's are i.i.d. random variables with uniform distribution $U(-0.5, 0.5)$. The results are reported in Table 5 and 6. The size of the test is

Table 7: Size of the proposed test for star-shaped graphical models under H_0 based on simulation. The population covariance matrix Σ_0 is given in (98). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.041	0.046	0.056	0.053	0.057	0.048
16	0.062	0.049	0.044	0.038	0.046	0.053
32	0.054	0.063	0.059	0.053	0.054	0.051
64	0.061	0.057	0.056	0.052	0.056	0.058
128	0.046	0.051	0.048	0.057	0.049	0.051
256	0.041	0.061	0.042	0.060	0.050	0.046

close to 5% and its power is slightly better than the case when testing group independence and converges to 1 quickly as p and n go to infinity.

In the seventh and eighth sets of simulations, we studied the size and power of the test for a family of star-shaped graphical model, that is, under H_0, $[\Theta]_{ij}$ is non-zero only if $i = 1$ or $j = 1$ or $i = j$. Similarly, Σ_0 and Σ_1 are given as:

$$\Sigma_0 = \Theta_0^{-1}, \quad \Sigma_1 = \Sigma_0 + \frac{1}{2}\vec{1}_p\vec{1}_p', \tag{98}$$

where

$$[\Theta_0]_{ij} = \begin{cases} 1 & i = j \\ u_k & (i,j) = (k,1) \text{ or } (1,k),\ 2 \le k \le p \\ 0 & \text{otherwise,} \end{cases} \tag{99}$$

for $1 \le i, j \le p$ and u_k's are i.i.d. random variables with uniform distribution $U(-\frac{1}{p}, \frac{1}{p})$. We choose the range of the uniform distribution so that Θ_0 is always positive definite otherwise it is not a valid concentration matrix. The results are reported in Table 7 and 8. Their patterns are expected and similar to previous results. The size of the test is close to 5% for all combination of p and n. The power of the test is close to 1 for large p and n.

Table 8: Power of the proposed test for star-shaped graphical models against alternative based on simulation. The population covariance matrix Σ_1 is given in (98). H_0 is rejected when t exceeds the 95% cutoff point of the empirical distribution function of $\tilde{t}_i, i = 1, 2, \ldots, 1000$. Each computed size is based on 1000 simulations.

	n					
p	8	16	32	64	128	256
8	0.26	0.54	0.90	1.00	1.00	1.00
16	0.45	0.84	0.99	1.00	1.00	1.00
32	0.64	0.96	1.00	1.00	1.00	1.00
64	0.82	1.00	1.00	1.00	1.00	1.00
128	0.94	1.00	1.00	1.00	1.00	1.00
256	0.96	1.00	1.00	1.00	1.00	1.00

2.6.2 An Empirical Example

We used the human activity recognition data set[1] that was previously studied by [17] to illustrate our method. It was collected from 15 participants performing 7 activities. We based our study on the data of working at computer and walking activity of the first participant. There are 33677 and 26860 consecutive samples of the two activities respectively. The x,y,z acceleration were collected at a sampling frequency of 52 Hz from a single chest-mounted accelerometer. We want to study the autocorrelation of the human activity time series data.

For each activity data set and each acceleration direction, we used $400 \times 52 = 20800$ samples starting from the 3001st one which represent 400 seconds of activity. We rearranged the samples into a 52×400 matrix where each column represents one second of data and computed its 400×400 sample covariance matrix. Then for each $b = 0, 1, \ldots, 49$, we tested if the underlying concentration matrix is a banded matrix with a bandwidth b. The p values were estimated using the simulation-based method with 1000 runs. The results are plotted in Figure 6.

[1]This data set is available on the website of the UCI Machine Learning Repository [44]: http://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer.

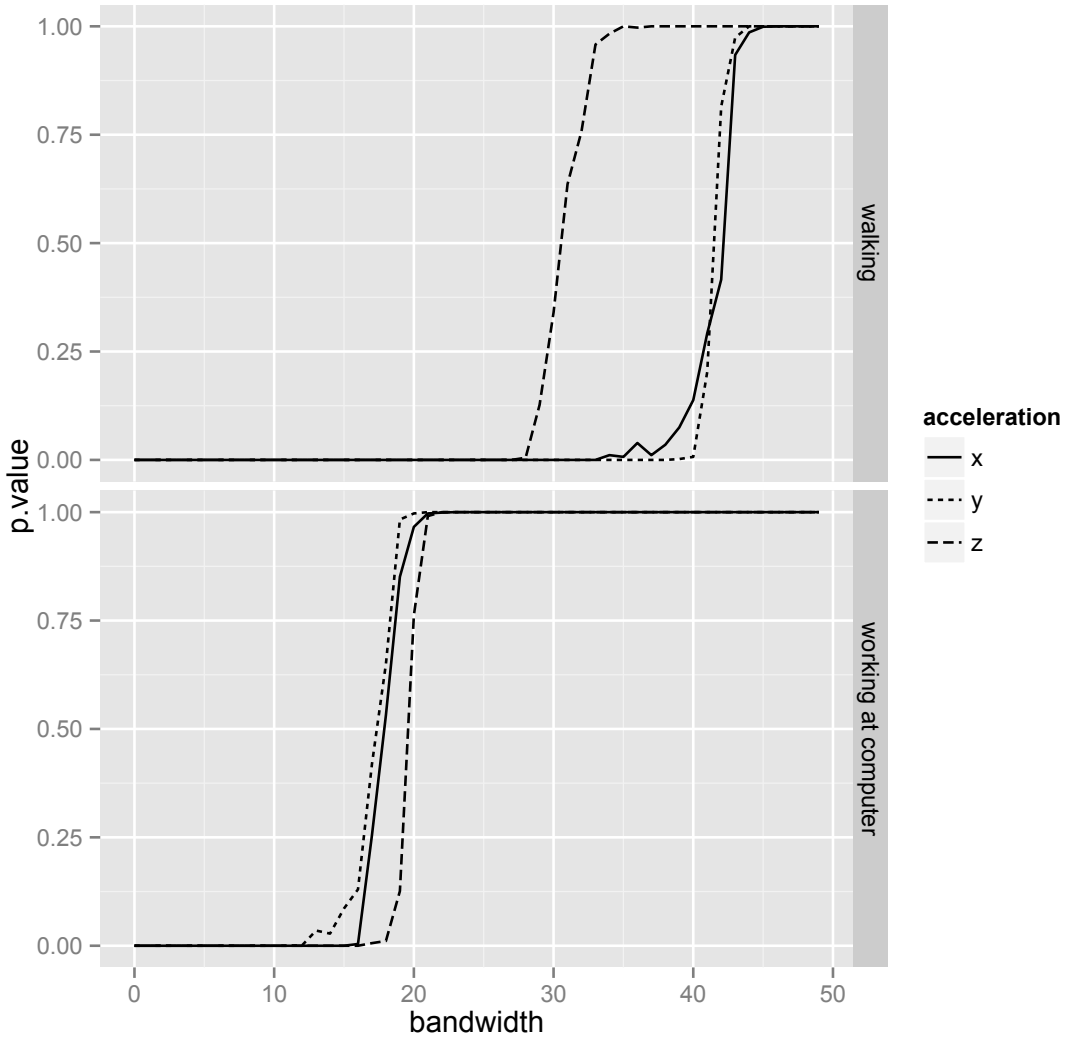

Figure 6: The p values of the bandedness tests of the inverse autocovariance matrix. Each p value was calculated using the simulation-based method and corresponds to one combination of human activity, acceleration direction and bandwidth.

As we can see from the figure, when the participant was walking, there was a significant increase in the autocorrelation for each of the three acceleration directions. Such increase might be the consequence of the periodicity of the walking activity. The actual bandwidth could exceed the calculated $30 \sim 45$ range because the null hypothesis is more likely to be rejected when the test bandwidth is close to the sample size. When the participant was working at computer, the body movement is more random and the acceleration data is less correlated.

APPENDIX A

SUPPLEMENTARY PROOFS

A.1 Proof of Theorem (1.4.1)

This proof is mainly based on the ideas in [89] and [52] that come from the studies on the linear approximate reconstruction problems in $\mathbb{R}^n$.

A.1.1 Some Variations of Bernstein's Inequality

First, we need the following lemma from [88] (Lemma 1.2.4):

Lemma A.1.1 *Let $\{X_n\}$ be an α-mixing time series. Let $\mathcal{F}_i^j$ be the σ-field generated by $(X_i, \ldots, X_j)$ for $i \leq j$ and $i, j \in \mathbb{Z}$. Let $X_1 \in \mathcal{F}_{-\infty}^s$, $X_2 \in \mathcal{F}_{s+k}^{\infty}$ with $\mathbb{E}(|X_1|^p), \mathbb{E}(|X_2|^q) < \infty$ where $\frac{1}{p} + \frac{1}{q} < 1$. Then we have*

$$|\mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)| \leq 10 \|X_1\|_p \|X_2\|_q (\alpha(n))^{1-\frac{1}{p}-\frac{1}{q}}. \tag{100}$$

where $\|\cdot\|_p = [\mathbb{E}(|\cdot|^p)]^{1/p}$ is the L^p norm of a random variable.

Lemma (A.1.1) basically shows that we can bound any single autocovariance with the mixing coefficients for an α-mixing process. Using this result, the norm of the autocovariance matrix of the process can also be bounded as follows.

Lemma A.1.2 *Let $X_1, X_2, \ldots, X_n \sim N(0, \sigma^2 I_n)$ and $X = (X_1, X_2, \ldots, X_n)'$ is an α-mixing Gaussian process with $\alpha(n) \leq c\rho^n$ where $c > 0$ and $0 < \rho < 1$ are constants. Let $\Gamma = \mathbb{E}(XX')$ be the autocovariance matrix of X. We have*

$$\|\Gamma\|_2 \leq \frac{c'\sigma^2}{1 - \rho^{1/3}}, \tag{101}$$

for some absolute constant c' where $\|\cdot\|_2$ is the l_2/l_2 matrix operator norm.

Proof In Lemma A.1.1, take $p = q = 3$, we have

$$|\Gamma_{ij}| \leq c''\sigma^2\rho^{|i-j|/3}, \quad i, j = 1, 2, \ldots, n, \tag{102}$$

for some constant $c'' > 0$. Therefore

$$\|\Gamma\|_2 \leq \sqrt{\|\Gamma\|_1 \cdot \|\Gamma\|_\infty} \leq \sqrt{\frac{2c''\sigma^2}{1-\rho^{1/3}} \cdot \frac{2c''\sigma^2}{1-\rho^{1/3}}} = \frac{2c''\sigma^2}{1-\rho^{1/3}}. \tag{103}$$

finishes the proof. ∎

Remarks The constant $1/3$ in $|\Gamma_{ij}| \leq c''\sigma^2\rho^{|i-j|/3}$ can be easily improved by choosing p and q such that $\frac{1}{p} + \frac{1}{q}$ is arbitrarily close to 1. But it is sufficient to support the following proofs.

The following lemma is a variant of the Bernstein inequality as in [52] (Lemma 1.1):

Lemma A.1.3 *Let $X_1, \ldots, X_n$ be independent centered random variables. If there exist some constant $d > 0$ such that for $\forall i$, $\|X_i\|_{\psi_1} \leq d$, we have*

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i| > t) \leq 2\exp(-cn\min(\frac{t}{d}, \frac{t^2}{d^2})), \tag{104}$$

for $\forall t > 0$, where $c > 0$ is an absolute constant.

We also need the following lemma from [8] (Lemma 0.2):

Lemma A.1.4 *Let $X = z'Az + b'z$ where $z = (z_1, \ldots, z_p)'$ is a Gaussian random vector such that $z_k \sim N(0, 1)$ are i.i.d. random variables, A and b are fixed $p \times p$ matrix and $p \times 1$ vector respectively. Denote the eigenvalues of $\frac{1}{2}(A + A')$ by $\lambda_k, k = 1, \ldots, p$. Let*

$$\lambda^+ = \max\{\max_{k=1,\ldots,p}\{\lambda_k\}, 0\}, \quad \lambda^- = \max\{\max_{k=1,\ldots,p}\{-\lambda_k\}, 0\}. \tag{105}$$

Then for $\forall t > 0$ we have

$$\begin{aligned} &\mathbb{P}(X \geq \mathrm{tr}(A) + \sqrt{\|A + A'\|_2^2 + 2\|b\|_2^2}\sqrt{t} + 2\lambda^+ t) \leq \exp(-t), \\ &\mathbb{P}(X \leq \mathrm{tr}(A) - \sqrt{\|A + A'\|_2^2 + 2\|b\|_2^2}\sqrt{t} - 2\lambda^- t) \leq \exp(-t). \end{aligned} \tag{106}$$

We extend Lemma 1.2 in [52] to Lemma A.1.5 in the α-mixing setting as follows.

Lemma A.1.5 *Let $X = (X_1, X_2, \ldots, X_n)'$ be an α-mixing Gaussian process with mixing coefficients $\alpha(n) \leq c\rho^n$ where $c > 0$ and $0 < \rho < 1$ and $X_i \sim N(0, I_p)$. Let μ denote the probability measure generated by X_1. Define $F = \{\langle \cdot, u\rangle | u \in T \subset S^{p-1}\}$ where $S^{p-1} = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$ and for $f_u = \langle \cdot, u\rangle \in F$ define*

$$Q_{f_u} = \frac{1}{n}\sum_{i=1}^{n} f_u^2(X_i) - \mathbb{E}f_u^2 = \frac{1}{n}\sum_{i=1}^{n}[X_i'u]^2 - \mathbb{E}[(X_1'u)^2], \tag{107}$$

$$R_{f_u} = [\frac{1}{n}\sum_{i=1}^{n} f_u^2(X_i)]^{1/2} = [\frac{1}{n}\sum_{i=1}^{n}(X_i'u)^2]^{1/2}, \tag{108}$$

where the expectations are taken with respect to μ. Let $d = 2\sup_{f\in F}\|f(X_1)\|_{\psi_2} \geq 1$. Then there exists absolute constant $c_1 > 0$ for which the following holds. For every $u, v \in S^{p-1}$ and every $b \geq 2$ we have

$$\mathbb{P}(R_{f_u - f_v} \geq b\|f_u - f_v\|_{\psi_2}) \leq 2\exp(-c_1(1-\rho^{1/3})^2 nb^2), \tag{109}$$

For every $b > 0$ we have

$$\mathbb{P}(|Q_{f_u}| \geq bd^2) \leq 2\exp(-c_1(1-\rho^{1/3})^2 n\min(b, b^2)), \tag{110}$$

$$\mathbb{P}(|Q_{f_u} - Q_{f_v}| \geq bd\|f_u - f_v\|_{\psi_2}) \leq 2\exp(-c_1(1-\rho^{1/3})^2 n\min(b, b^2)). \tag{111}$$

Proof For the first inequality, notice that $\mathbb{E}R^2_{f_u - f_v} = \|f_u - f_v\|_2^2$. Let $Y_i = X_i'(u - v), i = 1, \ldots, n$. $\{Y_i\}$ is an α-mixing Gaussian process with $Y_i \sim N(0, \|u - v\|_2^2)$ and the same mixing coefficients $\alpha(n) \leq c\rho^n$ as $\{X_i\}$. Suppose that the autocovariance matrix for Y is Γ_Y. By Lemma A.1.2, we know that

$$\|\Gamma_Y\|_2 \leq \frac{c'\|u - v\|_2^2}{1 - \rho^{1/3}}, \tag{112}$$

for some constant $c' > 0$. Suppose Γ_Y has a diagonalized representation $\Gamma_Y = U'DU$ for some orthogonal matrix U and diagonal matrix D. By the fact that Γ_Y is positive semidefinite and (112), the diagonal entries D_{ii} of D satisfy

$$0 \leq D_{ii} \leq \frac{c'\|u - v\|_2^2}{1 - \rho^{1/3}}, \tag{113}$$

for all $1 \leq i \leq n$.

Let $\widetilde{Y} = (\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_n)'$ be a random vector with i.i.d. $\widetilde{Y}_i \sim N(0,1)$. Because $Y'Y$ has the same distribution with $\widetilde{Y}'D\widetilde{Y}$, applying Lemma (A.1.3) it follows that for $t > 0$, there exists some constant $c > 0$ such that

$$\begin{aligned}
&\mathbb{P}(|R^2_{f_u - f_v} - \|f_u - f_v\|_2^2| \geq t) \\
=&\mathbb{P}(|\frac{1}{n}Y'Y - \|f_u - f_v\|_2^2| \geq t) \\
=&\mathbb{P}(|\frac{1}{n}\widetilde{Y}'D\widetilde{Y} - \|f_u - f_v\|_2^2| \geq t) \\
\leq&2\exp\left(-cn\min\left(\frac{t}{\frac{c'}{1-\rho^{1/3}}\|(f_u - f_v)^2\|_{\psi_1}}, \left(\frac{t}{\frac{c'}{1-\rho^{1/3}}\|(f_u - f_v)^2\|_{\psi_1}}\right)^2\right)\right) \\
\leq&2\exp\left(-c''(1-\rho^{1/3})^2 n\min\left(\frac{t}{\|(f_u - f_v)^2\|_{\psi_1}}, \left(\frac{t}{\|(f_u - f_v)^2\|_{\psi_1}}\right)^2\right)\right),
\end{aligned} \tag{114}$$

for some constant $c'' > 0$. The first inequality holds because

$$\|D_{ii}\widetilde{Y}_i^2\|_{\psi_1} \leq \frac{c'}{1-\rho^{1/3}}\|Y_i^2\|_{\psi_1} = \frac{c'}{1-\rho^{1/3}}\|(f_u - f_v)^2\|_{\psi_1}, \tag{115}$$

for $i = 1, \ldots, n$. Based on (114), we can prove our first inequality in the same way as the proof in [52].

For the second inequality, let $Y_i = X_i'u$ and similarly define $\Gamma_Y, U, D, \widetilde{Y}_i$ as in the proof of the first inequality. We have

$$0 \leq D_{ii} \leq \frac{c'''\|u\|_2^2}{1-\rho^{1/3}}, \tag{116}$$

for some constant $c''' > 0$ and

$$\begin{aligned}
&\mathbb{P}(|Q_{f_u}| \geq t) \\
=&\mathbb{P}(|\frac{1}{n}Y'Y - \mathbb{E}f_u^2| \geq t) \\
=&\mathbb{P}(|\frac{1}{n}\widetilde{Y}'D\widetilde{Y} - \mathbb{E}f_u^2| \geq t) \\
\leq& 2\exp\left(-cn\min\left(\frac{t}{\frac{c'''}{1-\rho^{1/3}}\|f_u^2\|_{\psi_1}}, \left(\frac{t}{\frac{c'''}{1-\rho^{1/3}}\|f_u^2\|_{\psi_1}}\right)^2\right)\right) \\
\leq& 2\exp\left(-cn\min\left(\frac{t}{\frac{c'''}{1-\rho^{1/3}}\|f_u\|_{\psi_2}}, \left(\frac{t}{\frac{c'''}{1-\rho^{1/3}}\|f_u\|_{\psi_2}}\right)^2\right)\right) \\
\leq& 2\exp\left(-c_2(1-\rho^{1/3})^2 n\min\left(\frac{t}{d}, \left(\frac{t}{d}\right)^2\right)\right),
\end{aligned} \tag{117}$$

where $c_2 > 0$ is a constant. Let $t = bd^2$ and notice that $d \geq 1$, we have

$$\begin{aligned}
&\mathbb{P}(|Q_{f_u}| \geq bd^2) \\
\leq& 2\exp\left(-c_2(1-\rho^{1/3})^2 n\min\left(bd, b^2d^2\right)\right) \\
\leq& 2\exp\left(-c_2(1-\rho^{1/3})^2 n\min\left(b, b^2\right)\right).
\end{aligned} \tag{118}$$

This proves (110).

For the third inequality, notice that

$$Q_{f_u} - Q_{f_v} = \frac{1}{n}\sum_{i=1}^n (f_u - f_v)(X_i)(f_u + f_v)(X_i). \tag{119}$$

Let

$$\begin{aligned}
Y_{i1} &= \frac{(f_u - f_v)(X_i)}{\|u - v\|_2} = \frac{X_i'(u-v)}{\|u-v\|_2} \\
Y_{i2} &= \frac{(f_u + f_v)(X_i)}{\|u + v\|_2} = \frac{X_i'(u+v)}{\|u+v\|_2}.
\end{aligned} \tag{120}$$

We have

$$\begin{aligned}
&\mathbb{P}(|Q_{f_u} - Q_{f_v}| \geq bd\|f_u - f_v\|_{\psi_2}) \\
=&\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n (f_u - f_v)(X_i)(f_u + f_v)(X_i)| \geq bd\|f_u - f_v\|_{\psi_2}) \\
=&\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n \frac{(f_u - f_v)(X_i)}{\|u-v\|_2}\frac{(f_u + f_v)(X_i)}{\|u+v\|_2}| \geq b\frac{d}{\|u+v\|_2}\frac{\|f_u - f_v\|_{\psi_2}}{\|u-v\|_2}) \\
=&\mathbb{P}(|\frac{1}{n}\sum_{i=1}^n Y_{i1}Y_{i2}| \geq c_2 b),
\end{aligned} \tag{121}$$

where $c_2 > 0$ is an absolute constant. Therefore, to prove inequality (111), it is sufficient to show that

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} Y_{i1}Y_{i2}| \geq b) \leq 2\exp(-c_3(1-\rho^{1/3})^2 n \min(b, b^2)). \tag{122}$$

for some absolute constant $c_3 > 0$. Let

$$Y = (Y_{11}, Y_{12}, Y_{21}, Y_{22}, \ldots, Y_{n1}, Y_{n2})'. \tag{123}$$

Obviously, $Y_{im} \sim N(0,1)$ for $1 \leq i \leq n$ and $m \in \{1,2\}$. We have

$$\frac{1}{n}\sum_{i=1}^{n} Y_{i1}Y_{i2} = \frac{1}{n}Y'MY = \widetilde{Y}'\left(\frac{1}{n}\Gamma_Y^{1/2}M\Gamma_Y^{1/2}\right)\widetilde{Y}, \tag{124}$$

where M is a $2n \times 2n$ matrix such that $M_{2i-1,2i} = M_{2i,2i-1} = \frac{1}{2}$ for $i = 1, \ldots, n$ and other entries of M are 0. $\tilde{Y} \sim N_{2n}(0, I_{2n})$ is a standard Gaussian random vector and Γ_Y is the autocovariance matrix of Y. Notice that Y_{im} and $Y_{jm'}$ where $1 \leq i, j \leq n$ and $m, m' \in \{1,2\}$ satisfies α-mixing condition with mixing coefficients $\alpha(|j-i|)$. Similarly to the proof of Lemma (A.1.2), we have

$$\|\Gamma_Y\|_2 \leq \frac{c'}{1-\rho^{1/3}}. \tag{125}$$

Denote $\frac{1}{n}\Gamma_Y^{1/2}M\Gamma_Y^{1/2}$ by A. $\mathrm{tr}(A)$ must be 0 because

$$\mathbb{E}(\widetilde{Y}'A\widetilde{Y}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} Y_{i1}Y_{i2}) = 0. \tag{126}$$

We also have

$$\|A\|_2 \leq \frac{1}{n}\|\Gamma_Y^{1/2}\|_2\|M\|_2\|\Gamma_Y^{1/2}\|_2 \leq \frac{c_3}{n(1-\rho^{1/3})}. \tag{127}$$

for some absolute constant $c_3 > 0$. Applying Lemma (A.1.4),

$$\mathbb{P}(|\widetilde{Y}'A\widetilde{Y}| \geq \frac{2c_3}{n(1-\rho^{1/3})}\sqrt{t} + \frac{2c_3}{n(1-\rho^{1/3})}t) \leq 2\exp(-t), \tag{128}$$

which immediately implies that

$$\mathbb{P}(|\widetilde{Y}'A\widetilde{Y}| \geq b) \leq 2\exp(-c_4(1-\rho^{1/3})^2 n \min(b, b^2)) \tag{129}$$

for some absolute constant $c_4 > 0$ and all $b > 0$. This proves inequality (111). ∎

A.1.2 Bounding Gaussian Process by the Complexity Measure

First, we establish the bounds for a α-mixing Gaussian process based on γ_2-functionals. Define

$$\gamma_2(M, d_M) = \inf \sup_{m \in M} \sum_{s=0}^{\infty} 2^{s/2} d_M(m, M_s), \tag{130}$$

for a metric space (M, d_M) where the infimum is taken with respect to all possible sequences $\{M_s : s \geq 0\}$ such that $\forall s$, $M_s \subseteq M$ and $|M_0| = 1$, $|M_s| = 2^{2^s}$.

A set G is said to be star-shaped if $\forall g \in G, 0 \leq \lambda \leq 1$ we have $\lambda g \in G$.

The following theorem is an extension of Theorem 1.4 in [52] based on Lemma A.1.5.

Theorem A.1.6 *Let X, μ, R_{f_u} be defined as in the previous section. Let $F \subset \{\langle \cdot, v \rangle : v \in \mathbb{R}^p\} \subset L_2(\mu)$ be star-shaped, $d = 2 \sup_{f \in F} \|f(X_1)\|_{\psi_2} \geq 1$ and $n \geq 1$. There exists some absolute constant $C_0 > 0$ and*

$$c_\rho = C_0(1 - \rho^{1/3})^2, \quad c' = \max(64\sqrt{\frac{2}{c_\rho}}, 32\sqrt{2}, \frac{512}{c_\rho}), \quad \bar{c} = \min(\frac{c_\rho}{64}, 1). \tag{131}$$

such that $\forall \theta \in (0, 1)$, with probability more than $1 - 4\exp(-\bar{c}\theta^2 n/d^4)$, for $\forall f \in F$ with $\mathbb{E}f^2 \geq t_n(\theta/c'd)^2$ where

$$t_n(\theta) := \inf \left\{ t > 0 : t \geq \frac{\gamma_2(F \cap tS_{L_2}, \| \ \|_{\psi_2})}{\theta\sqrt{n}} \right\}, \tag{132}$$

and $S_{L_2} = \{f : \|f\|_{L_2} = 1\}$, we have

$$(1 - \theta)\mathbb{E}f^2 \leq R_f^2 \leq (1 + \theta)\mathbb{E}f^2. \tag{133}$$

Proof The proof is similar to the proof of Theorem 1.4 in [52] by modifying some constants.

We use the l_* functional as the measure of complexity of a set $M \subset \mathbb{R}^p$, which is defined by

$$l_*(M) = \mathbb{E} \sup_{m \in M} |\sum_{i=1}^{p} z_i m_i|, \tag{134}$$

where $z_i, i = 1, \ldots, p$ are i.i.d. standard normal random variables and m_i is the ith coordinate of m.

Let $\{Z_m = \sum_{i=1}^{p} z_i m_i : m \in M\}$ be a centered Gaussian process indexed by a symmetric set $M \subset \mathbb{R}^n$. As shown in [24] and [72],

$$c_1 \gamma_2(M, \| \ \|_2) \leq l_*(M) \leq c_2 \gamma_2(M, \| \ \|_2). \tag{135}$$

where $c_1, c_2 > 0$ are absolute constants. We want to bound the α-mixing Gaussian process with the l_*-functional based on Theorem (A.1.6).

Define

$$h_n(\theta, M) := \inf \left\{ t > 0 : t \geq \frac{l_*(M \cap tS^{p-1})}{\theta \sqrt{n}} \right\}. \tag{136}$$

The following corollary is an extension of Corollary 2.7 in [52].

Corollary A.1.7 *Let $X, \mu, \theta, c', \bar{c}$ be the same as in Theorem A.1.6 and $M \subset S^{p-1}$. Let $d = 2\|X_0\|_{\psi_2}$ where X_0 is a standard normal random variable. If n satisfies*

$$n \geq (c''/\theta^2) l_*(M)^2, \tag{137}$$

where $c'' = (c'/c_1)^2 d^4$ and c_1 is the same as in (135), then with probability at least $1 - 4\exp(-c'''\theta^2 n)$ where $c''' = \bar{c}/d^4$, for $\forall m \in M$,

$$1 - \theta \leq \frac{\|Xm\|_2^2}{n} \leq 1 + \theta. \tag{138}$$

Proof Define

$$\widetilde{M} = \{\lambda m : m \in M, \lambda \in [0, 1]\}, \tag{139}$$

$$F_M = \{f_m : f_m = \langle \cdot, m \rangle, m \in \widetilde{M}\}. \tag{140}$$

Because μ is isotropic, $\|f_m\|_{L_2} = \|m\|_2$. On the other hand, notice that $d = 2 \sup_{f \in F} \|f(X_1)\|_{\psi_2}$. Therefore, for all $t > 0$,

$$\gamma_2(F \cap tS_{L_2}, \| \ \|_{\psi_2}) \leq d\gamma_2(F \cap tS_{L_2}, \| \ \|_{L_2}) \leq (d/c_1) l_*(\widetilde{M} \cap tS^{p-1}), \tag{141}$$

Therefore, we hvae

$$t_n(\theta, \widetilde{M}) \leq h_n(c_1\theta/d, \widetilde{M}). \tag{142}$$

Finally, (137) combined with $c'' = (c'/c_1)^2$ is equivalent to $h_n(c_1\theta/\tilde{c}d^2, \widetilde{M}) \leq 1$ thus $t_n(\theta/\bar{c}d) \leq 1$. Then (138) is an immediate consequence of Theorem A.1.6. ∎

A.1.3 Estimating Complexity Measure

Now let $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n)'$ be an α-mixing Gaussian process with $\alpha(n) \leq c\rho^n$ where $c > 0$ and $0 < \rho < 1$ are constants and $\Psi_i \sim N(0, I_p)$ for all $1 \leq i \leq n$. Let Σ be a deterministic positive semidefinite symmetric matrix and $X = \Psi\Sigma^{1/2}$. We are interested in the RE properties of X given the RE properties of Σ.

We say that a non-zero vector $v \in \mathbb{R}^p$ is admissible to (14) (or equivalently, to (15)) if $\|v_{I_0^c}\|_1 \leq k_0\|v_{I_0}\|_1$ holds with some $I_0 \subseteq \{1, 2, \ldots, p\}$ and $|I_0| \leq s$. The following lemma from [89] (Proposition 1.4) gives necessary and sufficient conditions for the RE assumptions (14) and (15).

Lemma A.1.8 *Let $k_0 > 0$, $s \in \mathbb{Z}$ and $1 \leq s \leq p/2$. Let $v \in \mathbb{R}^p$ be a non-zero vector that is admissible to (14). Then*

$$\|v_{T_0^c}\|_1 \leq k_0\|v_{T_0}\|_1. \tag{143}$$

where T_0 is the index set that corresponds to the s largest entries of v. Hence $RE(s, k_0, X)$ is equivalent to the following assumption:

For $\forall v \neq 0$ that is admissible to (14) and for $L(s, k_0, X) > 0$,

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{\|v_{T_0}\|_2}{L(s, k_0, X)} > 0, \tag{144}$$

holds.

And $RE(s, k_0, \Sigma)$ is equivalent to the following assumption:

For $\forall v \neq 0$ that is admissible to (15) and for $L(s, k_0, \Sigma) > 0$,

$$\|\Sigma^{1/2}v\|_2 \geq \frac{\|v_{T_0}\|_2}{L(s, k_0, \Sigma)} > 0, \tag{145}$$

holds.

Finally, if Σ *satisfies* $RE(s, k_0, \Sigma)$*, we have*

$$\|\Sigma^{1/2} v\|_2 \geq \frac{\|v_{I_0}\|_2}{L(s, k_0, \Sigma)} > 0, \tag{146}$$

for $\forall v$ *that is admissible to (15).*

Let

$$B := \left\{ v : \|\Sigma^{1/2} v\|_2 = 1 \text{ and } \|v_{T_0^c}\|_1 \leq k_0 \|v_{T_0}\|_1 \right\}, \tag{147}$$

where T_0 is defined similarly as in Lemma A.1.8. Let

$$\Gamma := \left\{ \delta \in \mathbb{R}^p : \delta = \Sigma^{1/2} v \text{ for some } v \in B \right\}. \tag{148}$$

The following lemma from [89] (Lemma 2.2) gives an estimate for the complexity of the subset Γ that is critical to our result.

Lemma A.1.9 *Let* $1 \leq s \leq p/2$*. Let* Σ *satisfies* $RE(s, k_0, \Sigma)$ *with constant* $L(s, k_0, \Sigma)$*. Let*

$$\rho(s, \Sigma) = \sup_{\substack{\|u\|_2 = 1, \\ |supp(u)| \leq s}} \|\Sigma^{1/2} u\|_2^2, \tag{149}$$

$$C_* = (6 + 3k_0) L(s, k_0, \Sigma) \sqrt{\rho(s, \Sigma)}. \tag{150}$$

Let the l_**-functional be defined by (134). Then*

$$l_*(\Gamma) \leq C_* \sqrt{s \log(5ep/s)}, \tag{151}$$

where e *is the natural logarithm base.*

Now we are ready to prove our main theorem (1.4.1).

Proof Combining Lemma A.1.9 with Corollary A.1.7, if n satisfies

$$n > \frac{c''}{\theta^2} C_*^2 s \log(5ep/s), \tag{152}$$

then with probability at least $1 - 4\exp(-c'''\theta^2 n/d^4)$, we have for all $v \in B$,

$$1 - \theta \leq \frac{\|\Psi\Sigma^{1/2}v\|_2^2}{n} \leq 1 + \theta. \tag{153}$$

Since $0 < \theta < 1$,

$$1 - \theta \leq \frac{\|\Psi\Sigma^{1/2}v\|_2}{\sqrt{n}} \leq 1 + \theta. \tag{154}$$

For all $v \neq 0$ that is admissible to (15), by Lemma A.1.8, we have $\|\Sigma^{1/2}v\|_2 > 0$. Then we can apply (154) to each $\frac{v}{\|\Sigma^{1/2}v\|_2} \neq 0$, which belongs to

$$B' := \left\{v : \|\Sigma^{1/2}v\|_2 = 1 \text{ and } v \text{ is admissible to (15)}\right\}, \tag{155}$$

and hence B (by Lemma (A.1.8)), that

$$\begin{aligned} \frac{\|Xv\|_2}{\sqrt{n}} &= \frac{\left\|\Psi\Sigma^{1/2}\frac{v}{\|\Sigma^{1/2}v\|_2}\right\|_2}{\sqrt{n}}\|\Sigma^{1/2}v\|_2 \\ &\geq (1-\theta)\|\Sigma^{1/2}v\|_2 \\ &\geq (1-\theta)\frac{\|v_{T_0}\|_2}{L(s,k_0,\Sigma)} > 0, \end{aligned} \tag{156}$$

holds with probability at least $1 - 4\exp(-c'''\theta^2 n/d^4)$. By Lemma A.1.8, the result follows immediately. ∎

A.2 Proof of Theorem (2.4.3)

This proof is mainly based on the ideas in [67]. First, we give some general results regarding some random projection matrices.

Lemma A.2.1 *Let n, p_1, p_2, q be positive integers such that $p_1, p_2 \leq q < n$. Let X_1, X_2 and Z be $n \times p_1$, $n \times p_2$ and $n \times q$ matrices with i.i.d. rows distributed as $N(0, I_{p_1})$, $N(0, I_{p_2})$ and $N(0, I_q)$ respectively. Let*

$$P_1 = X_1(X_1'X_1)^{-1}X_1', \quad P_2 = X_2(X_2'X_2)^{-1}X_2', \quad Q = Z(Z'Z)^{-1}Z'. \tag{157}$$

Then for every positive integers a, b, we have

$$\mathbb{E}(Q) = \frac{q}{n}I_n, \tag{158}$$

$$\mathbb{E}[\mathrm{tr}^a(P_1Q)\mathrm{tr}^b(P_2Q)] = \mathbb{E}[\mathrm{tr}^a(P_1Q)]\mathbb{E}[\mathrm{tr}^b(P_2Q)], \tag{159}$$

$$\begin{aligned}&\mathbb{E}[\{\mathrm{tr}(P_1Q) - \mathbb{E}[\mathrm{tr}(P_1Q)]\}^a\{\mathrm{tr}(P_2Q) - \mathbb{E}[\mathrm{tr}(P_2Q)]\}^b] \\ =&\mathbb{E}[\{\mathrm{tr}(P_1Q) - \mathbb{E}[\mathrm{tr}(P_1Q)]\}^a]\mathbb{E}[\{\mathrm{tr}(P_2Q) - \mathbb{E}[\mathrm{tr}(P_2Q)]\}^b],\end{aligned} \tag{160}$$

$$\mathbb{E}[\mathrm{tr}^2(P_1Q)|X_1] = \mathbb{E}[\mathrm{tr}^2(P_1Q)] = \frac{2p_1q(n-p_1)(n-q)}{n^2(n+2)(n-1)} + \frac{p_1^2q^2}{n^2}, \tag{161}$$

$$\mathbb{E}[\mathrm{tr}(P_1Q)\mathrm{tr}(P_2Q)|X_1, X_2] = \frac{2[n\mathrm{tr}(P_1P_2) - p_1p_2]q(n-q)}{n^2(n+2)(n-1)} + \frac{p_1p_2q^2}{n^2}. \tag{162}$$

Proof Notice that X_1, X_2 and Z has rank p_1, p_2 and q with probability 1 respectively. For any matrix $O_n \in O(n)$ where $O(n)$ is the real orthogonal group with dimension n, Z has the same distribution as O_nZ and

$$\begin{aligned}\mathbb{E}(Q) =&\mathbb{E}[Z(Z'Z)^{-1}Z'] \\ =&\mathbb{E}[O_nZ(Z'O_n'O_nZ)^{-1}Z'O_n'] \\ =&O_n\mathbb{E}[Q]O_n',\end{aligned} \tag{163}$$

which implies that $\mathbb{E}(Q) = cI_n$ with

$$c = \frac{\mathbb{E}[\mathrm{tr}(Q)]}{n} = \frac{\mathbb{E}[\mathrm{tr}((Z'Z)^{-1}Z'Z)]}{n} = \frac{q}{n}. \tag{164}$$

This proves (158). (163) also shows that O_nQO_n has the same distribution as Q. We know that P_1 is a projection matrix with rank p_1. There must exist an real orthogonal matrix O_{P_1} such that

$$O_{P_1}P_1O_{P_1}' = D_{p_1,n} := \mathrm{diag}\{1, 1, \ldots, 1, 0, 0, \ldots, 0\}, \tag{165}$$

where $D_{k,n}$ is a diagonal matrix with k 1's and $(n-k)$ 0's in its diagonal entries. Hence,

$$\begin{aligned}\mathbb{E}[\mathrm{tr}^2(P_1Q)|X_1] =&\mathbb{E}[\mathrm{tr}^2(O_{P_1}'D_{p_1,n}O_{P_1}Q)|X_1] \\ =&\mathbb{E}[\mathrm{tr}^2(D_{p_1,n}O_{P_1}QO_{P_1}')|X_1] \\ =&\mathbb{E}[\mathrm{tr}^2(D_{p_1,n}Q)|X_1].\end{aligned} \tag{166}$$

which does not depend on X_1 given that X_1 has rank p_1. Therefore

$$\mathbb{E}[\mathrm{tr}^2(P_1Q)|X_1] = \mathbb{E}[\mathrm{tr}^2(D_{p_1,n}Q)] = \mathbb{E}[\mathrm{tr}^2(P_1Q)] = \frac{2p_1q(n-p_1)(n-q)}{n^2(n+2)(n-1)} + \frac{p_1^2q^2}{n^2}. \tag{167}$$

The latter equality is due to [61]. This proves (161).

Similarly,

$$\begin{aligned}\mathbb{E}[\mathrm{tr}^a(P_1Q)\mathrm{tr}^b(P_2Q)] =& \mathbb{E}\{\mathbb{E}[\mathrm{tr}^a(P_1Q)\mathrm{tr}^b(P_2Q)|Z]\}\\ =& \mathbb{E}\{\mathbb{E}[\mathrm{tr}^a(P_1D_{q,n})\mathrm{tr}^b(P_2D_{q,n})|Z]\}\\ =& \mathbb{E}[\mathrm{tr}^a(P_1D_{q,n})]\mathbb{E}[\mathrm{tr}^b(P_2D_{q,n})]\\ =& \mathbb{E}[\mathrm{tr}^a(P_1Q)]\mathbb{E}[\mathrm{tr}^b(P_2Q)].\end{aligned} \tag{168}$$

This proves (159). (160) is a direct consequence of (159).

To prove (162), notice that

$$\begin{aligned}&\mathbb{E}[\mathrm{tr}(P_1Q)\mathrm{tr}(P_2Q)|X_1,X_2]\\ =&\mathbb{E}\left[\left(\sum_{i=1}^n\sum_{j=1}^n[P_1]_{ij}Q_{ji}\right)\left(\sum_{i=1}^n\sum_{j=1}^n[P_2]_{ij}Q_{ji}\right)|X_1,X_2\right].\end{aligned} \tag{169}$$

For distinct positive integers $1 \le i,j,k,l \le q$,

$$\mathbb{E}(Q_{ii}^2) = \mathbb{E}(Q_{11}^2) = \mathbb{E}[\mathrm{tr}^2(D_{1,n}Q)] = \frac{2q(n-q)}{n^2(n+2)} + \frac{q^2}{n^2}, \tag{170}$$

Obviously, $Q^2 = Q$. Then

$$\begin{aligned}\frac{q}{n} = \mathbb{E}(Q_{11}) =& \mathbb{E}[(1,0,\ldots,0)'Q(1,0,\ldots,0)]\\ =& \mathbb{E}[(1,0,\ldots,0)'Q^2(1,0,\ldots,0)]\\ =& \mathbb{E}[\sum_{i=1}^n Q_{1i}^2]\\ =& \mathbb{E}(Q_{11}^2) + (n-1)\mathbb{E}(Q_{12}^2),\end{aligned} \tag{171}$$

and

$$\mathbb{E}(Q_{ij}^2) = \mathbb{E}(Q_{12}^2) = \frac{\frac{q}{n} - \mathbb{E}(Q_{11}^2)}{n-1} = \frac{q(n-q)}{n(n+2)(n-1)}. \tag{172}$$

Furthermore,

$$\mathbb{E}[\mathrm{tr}^2(D_{2,n}Q)] = 2\mathbb{E}(Q_{11}^2) + 2\mathbb{E}(Q_{11}Q_{22}), \tag{173}$$

$$\begin{aligned}\mathbb{E}(Q_{ii}Q_{jj}) = \mathbb{E}(Q_{11}Q_{22}) =& \frac{4q(n-2)(n-q)}{n^2(n+2)(n-1)} + \frac{4q^2}{n^2} - \left[\frac{4q(n-q)}{n^2(n+2)} + \frac{2q^2}{n^2}\right] \\ =& -\frac{2q(n-q)}{n^2(n+2)(n-1)} + \frac{q^2}{n^2}.\end{aligned} \tag{174}$$

It is easy to verify that

$$\mathbb{E}(Q_{12}^2) = \frac{1}{2}[\mathbb{E}(Q_{11}^2) - \mathbb{E}(Q_{11}Q_{22})]. \tag{175}$$

Consider three $n \times n$ projection matrices

$$\begin{aligned}M_1 =& (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \ldots, 0)'(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, \ldots, 0), \\ M_2 =& (\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, 0, \ldots, 0)'(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, 0, \ldots, 0), \\ M_3 =& (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)'(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0).\end{aligned} \tag{176}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}[\mathrm{tr}^2(M_1Q)] =& \mathbb{E}[\mathrm{tr}^2(D_{1,n}Q)] = \mathbb{E}(Q_{11}^2), \\ \mathbb{E}[\mathrm{tr}^2(M_1Q)] =& \mathbb{E}[(\frac{1}{2}Q_{11} + Q_{12} + \frac{1}{2}Q_{22})^2], \\ =& \frac{1}{2}\mathbb{E}(Q_{11}^2) + \mathbb{E}(Q_{12}^2) + 2\mathbb{E}(Q_{11}Q_{12}) + \frac{1}{2}\mathbb{E}(Q_{11}Q_{22}), \\ \mathbb{E}(Q_{ii}Q_{ij}) =& \mathbb{E}(Q_{11}Q_{12}) = \frac{1}{2}\mathbb{E}(Q_{11}^2) - \frac{1}{2}\mathbb{E}(Q_{11}Q_{22}) - \mathbb{E}(Q_{12}^2) = 0.\end{aligned} \tag{177}$$

and

$$\begin{aligned}\mathbb{E}[\mathrm{tr}^2(M_2Q)] =& \mathbb{E}[\mathrm{tr}^2(D_{1,n}Q)] = \mathbb{E}(Q_{11}^2), \\ \mathbb{E}[\mathrm{tr}^2(M_2Q)] =& \frac{1}{9}\mathbb{E}[(Q_{11} + Q_{22} + Q_{33} + 2Q_{12} + 2Q_{13} + 2Q_{23})^2], \\ =& \frac{1}{9}[3\mathbb{E}(Q_{11}^2) + 6\mathbb{E}(Q_{11}Q_{22}) + 24\mathbb{E}(Q_{11}Q_{12}) + \\ & 12\mathbb{E}(Q_{11}Q_{23}) + 12\mathbb{E}(Q_{12}^2) + 24\mathbb{E}(Q_{12}Q_{13})], \\ =& \frac{1}{3}[3\mathbb{E}(Q_{11}^2) + 4\mathbb{E}(Q_{11}Q_{23}) + 8\mathbb{E}(Q_{12}Q_{13})], \\ \mathbb{E}(Q_{11}Q_{23}) =& -2\mathbb{E}(Q_{12}Q_{13}).\end{aligned} \tag{178}$$

Combining with

$$\begin{aligned}0=\mathbb{E}(Q_{12}) =&\mathbb{E}[(1,0,\ldots,0)'Q(0,1,0,\ldots,0)]\\ =&\mathbb{E}[(1,0,\ldots,0)'Q^2(0,1,0,\ldots,0)]\\ =&\mathbb{E}[\sum_{i=1}^{n}Q_{1i}Q_{i2}]\\ =&2\mathbb{E}(Q_{11}Q_{12})+(n-2)\mathbb{E}(Q_{12}Q_{13})\\ =&(n-2)\mathbb{E}(Q_{12}Q_{13}),\end{aligned} \tag{179}$$

We get

$$\mathbb{E}(Q_{ii}Q_{jk})=\mathbb{E}(Q_{11}Q_{23})=\mathbb{E}(Q_{ij}Q_{ik})=\mathbb{E}(Q_{12}Q_{13})=0. \tag{180}$$

Applying the same technique to M_3,

$$\mathbb{E}(Q_{ij}Q_{kl})=\mathbb{E}(Q_{12}Q_{34})=0. \tag{181}$$

Hence (169) can be calculated by

$$\begin{aligned}&\mathbb{E}\left[(\sum_{i=1}^{n}\sum_{j=1}^{n}[P_1]_{ij}Q_{ji})(\sum_{i=1}^{n}\sum_{j=1}^{n}[P_2]_{ij}Q_{ji})|X_1,X_2\right]\\ =&2\sum_{i=1}^{n}\sum_{j=1}^{n}[P_1]_{ij}[P_2]_{ij}\mathbb{E}(Q_{12}^2)+\sum_{i=1}^{n}[P_1]_{ii}\sum_{i=1}^{n}[P_2]_{ii}\mathbb{E}(Q_{11}Q_{22})\\ &+\sum_{i=1}^{n}[P_1]_{ii}[P_2]_{ii}[\mathbb{E}(Q_{11}^2)-2\mathbb{E}(Q_{12}^2)-\mathbb{E}(Q_{11}Q_{22})]\\ =&2\mathrm{tr}(P_1P_2)\mathbb{E}(Q_{12}^2)+\mathrm{tr}(P_1)\mathrm{tr}(P_2)\mathbb{E}(Q_{11}Q_{22})\\ =&\frac{2[n\mathrm{tr}(P_1P_2)-p_1p_2]q(n-q)}{n^2(n+2)(n-1)}+\frac{p_1p_2q^2}{n^2}.\end{aligned} \tag{182}$$

This proves (162). ∎

Now we are ready to prove Theorem (2.4.3).

Proof As $n_k\to\infty$, we assume that $n_k>K$ where K is defined in Assumption (2.4.1). By the definition of t_k,

$$t_k=\mathrm{tr}[(\hat{\Theta}_{0k}^{\frac{1}{2}}\hat{\Sigma}_k\hat{\Theta}_{0k}^{\frac{1}{2}}-I_{p_k})^2]=\mathrm{tr}[(\hat{\Sigma}_k\hat{\Theta}_{0k}-I_{p_k})^2]=\mathrm{tr}(\hat{\Sigma}_k\hat{\Theta}_{0k}\hat{\Sigma}_k\hat{\Theta}_{0k})-2\mathrm{tr}(\hat{\Sigma}_k\hat{\Theta}_{0k})+p_k. \tag{183}$$

If we partition $\hat{\Sigma}_k$ according to the cliques or groups,

$$\operatorname{tr}(\hat{\Sigma}_k\hat{\Theta}_{0k}) = \sum_{i=1}^{m_k} \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}) = \sum_{i=1}^{m_k} |C_{ik}| = p_k, \tag{184}$$

$$\begin{aligned}\operatorname{tr}(\hat{\Sigma}_k\hat{\Theta}_{0k}\hat{\Sigma}_k\hat{\Theta}_{0k}) =& \sum_{i=1}^{m_k}\sum_{j=1}^{m_k} \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{jk}}[\hat{\Sigma}_k]^{-1}_{C_{jk}C_{jk}}[\hat{\Sigma}_k]_{C_{jk}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}) \\ =& 2\sum_{1\le i<j\le m_k}\sum \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{jk}}[\hat{\Sigma}_k]^{-1}_{C_{jk}C_{jk}}[\hat{\Sigma}_k]_{C_{jk}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}) \\ &+ \sum_{i=1}^{m_k} \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}[\hat{\Sigma}_k]_{C_{ik}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}) \\ =& 2\sum_{1\le i<j\le m_k}\sum \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{jk}}[\hat{\Sigma}_k]^{-1}_{C_{jk}C_{jk}}[\hat{\Sigma}_k]_{C_{jk}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}) + p_k.\end{aligned} \tag{185}$$

All of the slicing of the matrices are done before the inversion. Therefore we have

$$t_k = 2\sum_{1\le i<j\le m_k}\sum \operatorname{tr}([\hat{\Sigma}_k]_{C_{ik}C_{jk}}[\hat{\Sigma}_k]^{-1}_{C_{jk}C_{jk}}[\hat{\Sigma}_k]_{C_{jk}C_{ik}}[\hat{\Sigma}_k]^{-1}_{C_{ik}C_{ik}}). \tag{186}$$

Notice that $\hat{\Sigma}_k$ has the same distribution as $\Sigma_k^{\frac{1}{2}} Z_k' Z_k \Sigma_k^{\frac{1}{2}}$, where the rows of the $(n_k - 1) \times p_k$ matrix Z_k are independent $N_{p_k}(0, I_{p_k})$ random vectors. Under the null hypothesis,

$$\Sigma_k = \operatorname{diag}\{[\Sigma_k]_{C_{1k}C_{1k}}, [\Sigma_k]_{C_{2k}C_{2k}}, \ldots, [\Sigma_k]_{C_{m_kk}C_{m_kk}}\}, \tag{187}$$

is a block diagonal matrix and

$$\Sigma_k^{\frac{1}{2}} = \operatorname{diag}\{[\Sigma_k]^{\frac{1}{2}}_{C_{1k}C_{1k}}, [\Sigma_k]^{\frac{1}{2}}_{C_{2k}C_{2k}}, \ldots, [\Sigma_k]^{\frac{1}{2}}_{C_{m_kk}C_{m_kk}}\}. \tag{188}$$

t_k has the same distribution as

$$\begin{aligned}2\sum_{1\le i<j\le m_k}\sum \operatorname{tr}([\Sigma_k]^{\frac{1}{2}}_{C_{ik}C_{ik}}[Z_k'Z_k]_{C_{ik}C_{jk}}[\Sigma_k]^{\frac{1}{2}}_{C_{jk}C_{jk}}[\Sigma_k]^{-\frac{1}{2}}_{C_{jk}C_{jk}}[Z_k'Z_k]^{-1}_{C_{jk}C_{jk}}[\Sigma_k]^{-\frac{1}{2}}_{C_{jk}C_{jk}} \\ [\Sigma_k]^{\frac{1}{2}}_{C_{jk}C_{jk}}[Z_k'Z_k]_{C_{jk}C_{ik}}[\Sigma_k]^{\frac{1}{2}}_{C_{ik}C_{ik}}[\Sigma_k]^{-\frac{1}{2}}_{C_{ik}C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}[\Sigma_k]^{-\frac{1}{2}}_{C_{ik}C_{ik}}).\end{aligned} \tag{189}$$

or

$$2\sum_{1\le i<j\le m_k}\sum \operatorname{tr}([Z_k'Z_k]_{C_{ik}C_{jk}}[Z_k'Z_k]^{-1}_{C_{jk}C_{jk}}[Z_k'Z_k]_{C_{jk}C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}). \tag{190}$$

Without loss of generality, we assume that $|C_i| \leq |C_j|$ for $1 \leq i < j \leq m_k$, otherwise we can simply rearrange the order of the cliques. For notational convenience, for $1 \leq i < j \leq m_k$, define

$$r_{ijk} := \mathrm{tr}([Z_k'Z_k]_{C_{ik}C_{jk}}[Z_k'Z_k]^{-1}_{C_{jk}C_{jk}}[Z_k'Z_k]_{C_{jk}C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}). \tag{191}$$

Define $W_{1k} = 0$. For $l = 2, 3, \ldots, m_k$, define

$$D_{lk} := \sum_{i=1}^{l-1} r_{ilk} - \frac{|C_{lk}|}{n_k - 1}\sum_{i=1}^{l-1}|C_{ik}|, \qquad W_{lk} = \sum_{i=2}^{l} D_{ik}, \tag{192}$$

$$\mathcal{F}_{l-1,k} = \{\text{The } i\text{th column vector of } Z_k : i \in \cup_{j=1}^{l-1} C_{jk}\}. \tag{193}$$

Obviously,

$$t_k - 2\sum_{l=2}^{m_k}\sum_{i=1}^{l-1}\frac{|C_{ik}||C_{lk}|}{n_k - 1} = 2W_{m_k k}. \tag{194}$$

In addition, for each $i = 1, 2, \ldots, l-1$,

$$\begin{aligned}\mathbb{E}\left\{r_{ilk}|\mathcal{F}_{l-1,k}\right\} =& \mathrm{tr}([Z_k]_{\cdot C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}[Z_k]'_{\cdot C_{ik}}\mathbb{E}\{[Z_k]_{\cdot C_{lk}}[Z_k'Z_k]^{-1}_{C_{lk}C_{lk}}[Z_k]'_{\cdot C_{lk}}\}) \\ =& \mathrm{tr}([Z_k]_{\cdot C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}[Z_k]'_{\cdot C_{ik}}\frac{|C_{lk}|}{n_k-1}I_{n_k-1}) \\ =& \frac{|C_{ik}||C_{lk}|}{n_k-1}.\end{aligned} \tag{195}$$

where $[Z_k]_{\cdot C_{ik}}$ is the submatrix of Z_k with columns corresponding to the index set C_{ik}. The second equality is due to Lemma (A.2.1). From (195) we know that $\mathbb{E}(D_{lk}|\mathcal{F}_{l-1,k}) = 0$. For each k, $\{W_{lk}, l = 2, 3, \ldots, m_k\}$ is a martingale with corresponding martingale differences $D_{2k}, D_{3k}, \ldots, D_{m_k k}$.

Based on the results in [61] and Lemma (A.2.1), for distinct integers h, i, j, l, we have

$$\mathbb{E}(r_{ilk}) = \frac{|C_{ik}||C_{lk}|}{n_k - 1}, \tag{196}$$

$$\mathbb{E}(r_{ilk}^2) = \frac{2|C_{ik}||C_{lk}|(n_k - 1 - |C_{ik}|)(n_k - 1 - |C_{lk}|)}{(n_k-1)^2(n_k+1)(n_k-2)} + \frac{|C_{ik}|^2|C_{lk}|^2}{(n_k-1)^2}, \tag{197}$$

$$\mathbb{E}(r_{ilk}r_{hlk}) = \mathbb{E}(r_{ilk})\mathbb{E}(r_{hlk}) = \mathbb{E}(r_{ijk})\mathbb{E}(r_{hlk}) = \frac{|C_{ik}||C_{hk}||C_{lk}|^2}{(n_k-1)^2}. \tag{198}$$

Hence,

$$\begin{aligned}\mathbb{E}[\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k})] = \mathbb{E}(D_{lk}^2) =& \mathbb{E}\left[\left(\sum_{i=1}^{l-1} r_{ilk}\right)^2\right] - \left[\mathbb{E}\left(\sum_{i=1}^{l-1} r_{ilk}\right)\right]^2 \\ =& \sum_{i=1}^{l-1}\{\mathbb{E}(r_{ilk}^2) - [\mathbb{E}(r_{ilk})]^2\} \\ =& \sum_{i=1}^{l-1} \frac{2|C_{ik}||C_{lk}|(n_k - 1 - |C_{ik}|)(n_k - 1 - |C_{lk}|)}{(n_k-1)^2(n_k+1)(n_k-2)}.\end{aligned} \tag{199}$$

Under assumption (2.4.2), $\max_{1\le i\le m_k}|C_{ik}| < K$. It is easy to see that

$$\begin{aligned}\mathbb{E}\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k})\right] \ge& \frac{(n_k-1-K)(n_k-1-K)}{(n_k-1)^2(n_k+1)(n_k-2)}\sum_{l=2}^{m_k}\sum_{i=1}^{l-1} 2|C_{ik}||C_{lk}| \\ \ge& \frac{(n_k-1-K)(n_k-1-K)}{(n_k-1)^2(n_k+1)(n_k-2)}(p_k^2 - Kp_k),\end{aligned} \tag{200}$$

$$\begin{aligned}\mathbb{E}\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k})\right] \le& \frac{(n_k-1-K)(n_k-1-K)}{(n_k-1)^2(n_k+1)(n_k-2)}\sum_{l=2}^{m_k}\sum_{i=1}^{l-1} 2|C_{ik}||C_{lk}| \\ \le& \frac{(n_k-1)(n_k-1)}{(n_k-1)^2(n_k+1)(n_k-2)}p_k^2.\end{aligned} \tag{201}$$

Under assumption (2.4.1), both bounds converges to γ^2 as $k\to\infty$, which means that

$$\lim_{k\to\infty}\mathbb{E}\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k})\right] = \gamma^2. \tag{202}$$

Now by Lemma (A.2.1), for $1 \le i \ne j \le l-1$,

$$\mathbb{E}(r_{ilk}^2|\mathcal{F}_{l-1,k}) = \frac{2|C_{ik}||C_{lk}|(n_k-1-|C_{ik}|)(n_k-1-|C_{lk}|)}{(n_k-1)^2(n_k+1)(n_k-2)} + \frac{|C_{ik}|^2|C_{lk}|^2}{(n_k-1)^2}, \tag{203}$$

$$\mathbb{E}(r_{ilk}r_{jlk}|\mathcal{F}_{l-1,k}) = \frac{2[(n_k-1)r_{ijk} - |C_{ik}||C_{jk}|]|C_{lk}|(n_k-1-|C_{lk}|)}{(n_k-1)^2(n_k+1)(n_k-2)} + \frac{|C_{ik}||C_{jk}||C_{lk}|^2}{(n_k-1)^2}, \tag{204}$$

and

$$\begin{aligned}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k}) =& \mathbb{E}\left[Var\left(\sum_{i=1}^{l-1} r_{ilk}\right)|\mathcal{F}_{l-1,k}\right] \\ =& \sum_{i=1}^{l-1}\sum_{j=1}^{l-1}\frac{2[(n_k-1)r_{ijk} - |C_{ik}||C_{jk}|]|C_{lk}|(n_k-1-|C_{lk}|)}{(n_k-1)^2(n_k+1)(n_k-2)},\end{aligned} \tag{205}$$

where

$$r_{iik} := \mathrm{tr}([Z_k'Z_k]_{C_{ik}C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}[Z_k'Z_k]_{C_{ik}C_{ik}}[Z_k'Z_k]^{-1}_{C_{ik}C_{ik}}) = |C_{ik}|. \tag{206}$$

Combining (205) with (197) and (198),

$$Var\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k})\right] = Var\left[\sum_{l=2}^{m_k}\sum_{i=1}^{l-1}\sum_{j=1}^{l-1}\frac{2r_{ijk}|C_{lk}|(n_k-1-|C_{lk}|)}{(n_k-1)(n_k+1)(n_k-2)}\right] \to 0. \tag{207}$$

This guarantees that

$$\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^2|\mathcal{F}_{l-1,k}) \to \gamma^2, \tag{208}$$

in probability. In addition, the Liapounov condition,

$$\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^4|\mathcal{F}_{l-1,k}) \to 0. \tag{209}$$

in probability, holds since

$$\begin{aligned}
&\mathbb{P}\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^4|\mathcal{F}_{l-1,k}) \geq \epsilon\right] \\
\leq&\frac{1}{\epsilon}\mathbb{E}\left[\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^4|\mathcal{F}_{l-1,k})\right] \\
=&\frac{1}{\epsilon}\sum_{l=2}^{m_k}\mathbb{E}(D_{lk}^4) \\
=&\frac{1}{\epsilon}\sum_{l=2}^{m_k}\mathbb{E}\left[\left\{\sum_{i=1}^{l-1}[r_{ilk}-\mathbb{E}(r_{ilk})]\right\}^4\right] \\
=&\frac{1}{\epsilon}\sum_{l=2}^{m_k}\left[\sum_{i=1}^{l-1}\mathbb{E}\{[r_{ilk}-\mathbb{E}(r_{ilk})]^4\} + 2\sum\sum_{1\leq i<j\leq l-1}\mathbb{E}\{[r_{ilk}-\mathbb{E}(r_{ilk})]^2[r_{jlk}-\mathbb{E}(r_{jlk})]^2\}\right] \\
=&O(\frac{1}{n}) \to 0,
\end{aligned} \tag{210}$$

for every $\epsilon > 0$. The third equality is due to (160) extended to the case with up to four matrices P_1, P_2, P_3, P_4 with similar technique and the last equality is based on the results in [61] that

$$\mathbb{E}\{[r_{jlk}-\mathbb{E}(r_{jlk})]^s\} = O(\frac{1}{n^s}). \tag{211}$$

The Liapounov condition implies the Lindeberg condition,

$$\sum_{l=2}^{m_k} \mathbb{E}[D_{lk}^2 I(|D_{lk}| > \epsilon)|\mathcal{F}_{l-1,k}] \to 0. \tag{212}$$

in probability for all $\epsilon > 0$. Applying Corollary 3.1 in [30] with (208) and (212),

$$t_k - 2\sum_{l=2}^{m_k}\sum_{i=1}^{l-1} \frac{|C_{ik}||C_{lk}|}{n_k - 1} = 2W_{m_k k} = 2\sum_{l=1}^{m_k} D_{lk} \to N(0, 4\gamma^2) \tag{213}$$

in distribution. This completes the proof. ∎

REFERENCES

[1] Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N., "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approximation*, vol. 34, no. 1, pp. 61–88, 2011.

[2] Anderson, T., Anderson, T., Anderson, T., and Anderson, T., *An introduction to multivariate statistical analysis*, vol. 2. Wiley New York, 1958.

[3] Andrews, D. W., "Non-strong mixing autoregressive processes," *Journal of Applied Probability*, pp. 930–934, 1984.

[4] Audrino, F. and Camponovo, L., "Oracle properties and finite sample inference of the adaptive lasso for time series regression models," *arXiv preprint arXiv:1312.1473*, 2013.

[5] Bai, Z., Jiang, D., Yao, J., and Zheng, S., "Corrections to lrt on large-dimensional covariance matrix by rmt," *The Annals of Statistics*, vol. 37, no. 6B, pp. 3822–3840, 2009.

[6] Bai, Z. and Silverstein, J., *Spectral analysis of large dimensional random matrices*. Springer, 2009.

[7] Banerjee, O., El Ghaoui, L., and d'Aspremont, A., "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[8] Bechar, I., "A bernstein-type inequality for stochastic processes of quadratic forms of gaussian variables," *arXiv preprint arXiv:0909.3595*, 2009.

[9] Bickel, P. J., Ritov, Y., and Tsybakov, A. B., "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[10] Bickel, P. and Levina, E., "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[11] Breiman, L., "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[12] Cai, T. and Jiang, T., "Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices," *The Annals of Statistics*, vol. 39, no. 3, pp. 1496–1525, 2011.

[13] Cai, T., Zhang, C., and Zhou, H., "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, vol. 38, no. 4, pp. 2118–2144, 2010.

[14] Candes, E. and Tao, T., "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, pp. 2313–2351, 2007.

[15] Candes, E. J. and Tao, T., "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[16] Candès, E. J. and Tao, T., "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[17] Casale, P., Pujol, O., and Radeva, P., "Personalization and user verification in wearable systems using biometric walking patterns," *Personal and Ubiquitous Computing*, vol. 16, no. 5, pp. 563–580, 2012.

[18] Chen, S., Zhang, L., and Zhong, P., "Tests for high-dimensional covariance matrices," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 810–819, 2010.

[19] Dahl, J., Vandenberghe, L., and Roychowdhury, V., "Covariance selection for nonchordal graphs via chordal embedding," *Optimization Methods & Software*, vol. 23, no. 4, pp. 501–520, 2008.

[20] Donoho, D. L., "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[21] Donoho, D. L., Elad, M., and Temlyakov, V. N., "Stable recovery of sparse overcomplete representations in the presence of noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, 2006.

[22] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., and others, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[23] Fan, J. and Li, R., "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[24] Fernique, X., "Regularité des trajectoires des fonctions aléatoires gaussiennes," in *Ecole dEté de Probabilités de Saint-Flour IV1974*, pp. 1–96, Springer, 1975.

[25] Friedman, J., Hastie, T., and Tibshirani, R., "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[26] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., and others, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[27] Friedman, J., Hastie, T., and Tibshirani, R., "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[28] Fu, W. J., "Penalized regressions: the bridge versus the lasso," *Journal of computational and graphical statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[29] Goldenshluger, A. and Zeevi, A., "Nonasymptotic bounds for autoregressive time series modeling," *Annals of statistics*, pp. 417–444, 2001.

[30] Hall, P. and Heyde, C. C., *Martingale limit theory and its application*. Academic press, 1980.

[31] Hamilton, J. D., *Time series analysis*, vol. 2. Princeton university press Princeton, 1994.

[32] Hsu, N.-J., Hung, H.-L., and Chang, Y.-M., "Subset selection for vector autoregressive processes using lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 7, pp. 3645–3657, 2008.

[33] Jiang, T., "The asymptotic distributions of the largest entries of sample correlation matrices," *The Annals of Applied Probability*, vol. 14, no. 2, pp. 865–880, 2004.

[34] John, S., "Some optimal multivariate tests," *Biometrika*, vol. 58, no. 1, pp. 123–127, 1971.

[35] Klartag, B. and Mendelson, S., "Empirical processes and random projections," *Journal of Functional Analysis*, vol. 225, no. 1, pp. 229–245, 2005.

[36] Kock, A. B. and Callot, L., "Oracle inequalities for high dimensional vector autoregressions," *Aarhus University, CREATES Research Paper*, vol. 16, 2012.

[37] Koltchinskii, V. and others, "The dantzig selector and sparsity oracle inequalities," *Bernoulli*, vol. 15, no. 3, pp. 799–828, 2009.

[38] Kosorok, M. R., *Introduction to empirical processes and semiparametric inference*. Springer, 2007.

[39] Lam, C. and Souza, P. C., "Regularization for spatial panel time series using the adaptive lasso," tech. rep., Mimeo, 2013.

[40] Lauritzen, S., *Graphical models*, vol. 17. Oxford University Press, USA, 1996.

[41] Ledoit, O. and Wolf, M., "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Annals of Statistics*, pp. 1081–1102, 2002.

[42] Levina, E., Rothman, A., and Zhu, J., "Sparse estimation of large covariance matrices via a nested lasso penalty," *The Annals of Applied Statistics*, pp. 245–263, 2008.

[43] Li, J. and Chen, W., "Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models," *Available at SSRN 2410214*, 2014.

[44] Lichman, M., "UCI machine learning repository," 2013.

[45] Lin, Z. and Xiang, Y., "A hypothesis test for independence of sets of variates in high dimensions," *Statistics & Probability Letters*, vol. 78, no. 17, pp. 2939–2946, 2008.

[46] Lütkepohl, H., "Introduction to multiple time series analysis, 1993," 1993.

[47] Marchenko, V. and Pastur, L., "Distribution of eigenvalues for some sets of random matrices," *Matematicheskii Sbornik*, vol. 114, no. 4, pp. 507–536, 1967.

[48] McShane, B. B., Wyner, A. J., and others, "A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable?," *The Annals of Applied Statistics*, vol. 5, no. 1, pp. 5–44, 2011.

[49] Medeiros, M. C. and Mendes, E., "Estimating high-dimensional time series models," *CREATES Research Paper*, vol. 37, 2012.

[50] Meinshausen, N. and Bühlmann, P., "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, pp. 1436–1462, 2006.

[51] Meinshausen, N. and Yu, B., "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, pp. 246–270, 2009.

[52] Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N., "Reconstruction and subgaussian operators in asymptotic geometric analysis," *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.

[53] Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N., "Uniform uncertainty principle for bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008.

[54] Mokkadem, A., "Mixing properties of arma processes," *Stochastic processes and their applications*, vol. 29, no. 2, pp. 309–315, 1988.

[55] Muirhead, R., *Aspects of multivariate statistical theory*, vol. 42. Wiley Online Library, 1982.

[56] Nagao, H., "On some test criteria for covariance matrix," *The Annals of Statistics*, pp. 700–709, 1973.

[57] Nardi, Y. and Rinaldo, A., "Autoregressive process modeling via the lasso procedure," *Journal of Multivariate Analysis*, vol. 102, no. 3, pp. 528–549, 2011.

[58] Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., and Phillips, J. C., "Gpu computing," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 879–899, 2008.

[59] Park, H. and Sakaori, F., "Lag weighted lasso for time series model," *Computational Statistics*, vol. 28, no. 2, pp. 493–504, 2013.

[60] Pham, T. D. and Tran, L. T., "Some mixing properties of time series models," *Stochastic processes and their applications*, vol. 19, no. 2, pp. 297–303, 1985.

[61] Pillai, K. S., *On some distribution problems in multivariate analysis*. University of North Carolina, 1954.

[62] Qiu, Y. and Chen, S., "Test for bandedness of high-dimensional covariance matrices and bandwidth estimation," *The Annals of Statistics*, vol. 40, no. 3, pp. 1285–1314, 2012.

[63] Rajapakse, J. C. and Mundra, P. A., "Stability of building gene regulatory networks with sparse autoregressive models," *BMC bioinformatics*, vol. 12, no. Suppl 13, p. S17, 2011.

[64] Rajaratnam, B., Massam, H., and Carvalho, C., "Flexible covariance estimation in graphical gaussian models," *The Annals of Statistics*, vol. 36, no. 6, pp. 2818–2849, 2008.

[65] Raskutti, G., Wainwright, M. J., and Yu, B., "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[66] Rudelson, M. and Vershynin, R., "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[67] Schott, J., "Testing for complete independence in high dimensions," *Biometrika*, vol. 92, no. 4, pp. 951–956, 2005.

[68] Schott, J., "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6535–6542, 2007.

[69] Shibata, R., "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *The Annals of Statistics*, pp. 147–164, 1980.

[70] Song, S. and Bickel, P. J., "Large vector auto regressions," *arXiv preprint arXiv:1106.3915*, 2011.

[71] Srivastava, M., "Some tests concerning the covariance matrix in high dimensional data," *J. Japan Statist. Soc*, vol. 35, no. 2, pp. 251–272, 2005.

[72] Talagrand, M., "Regularity of gaussian processes," *Acta mathematica*, vol. 159, no. 1, pp. 99–149, 1987.

[73] Talagrand, M., *The generic chaining*. Springer, 2005.

[74] Tang, L., Zhou, Z., and Wu, C., "Efficient estimation and variable selection for infinite variance autoregressive models," *Journal of Applied Mathematics and Computing*, vol. 40, no. 1-2, pp. 399–413, 2012.

[75] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[76] Tibshirani, R., "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[77] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K., "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[78] Van De Geer, S. A., Bühlmann, P., and others, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[79] Van der Vaart, A. and Wellner, J., "Weak convergence and empirical processes. 1996."

[80] Vidaurre, D., Bielza, C., and Larrañaga, P., "Classification of neural signals from sparse autoregressive features," *Neurocomputing*, vol. 111, pp. 21–26, 2013.

[81] Wainwright, M. J., "Sharp thresholds for high-dimensional and noisy recovery of sparsity using l1-constrained quadratic programming," 2006.

[82] Wang, H., Li, G., and Tsai, C.-L., "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 1, pp. 63–78, 2007.

[83] Wu, W. and Pourahmadi, M., "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, vol. 90, no. 4, pp. 831–844, 2003.

[84] Xu, G., Xiang, Y., Wang, S., and Lin, Z., "Regularization and variable selection for infinite variance autoregressive models," *Journal of Statistical Planning and Inference*, vol. 142, no. 9, pp. 2545–2553, 2012.

[85] Yuan, M. and Lin, Y., "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[86] Yuan, M. and Lin, Y., "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[87] Zhao, P. and Yu, B., "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[88] Zhengyan, L. and Chuanrong, L., *Limit theory for mixing dependent random variables*, vol. 378. Springer, 1996.

[89] Zhou, S., "Restricted eigenvalue conditions on subgaussian random matrices," *arXiv preprint arXiv:0912.4045*, 2009.

[90] Zou, H., "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[91] Zou, H. and Hastie, T., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.