STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DATA

Yin Xia

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

Tony Cai, Professor, Department of Statistics

Graduate Group Chairperson

Charles L. Epstein, Professor, Department of Mathematics

Dissertation Committee
Mark Low, Professor, Department of Statistics
Zongming Ma, Assistant Professor, Department of Statistics
Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Professor Tony Cai for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge.

I am deeply grateful to Professor Mark Low for his guidance and encouragement during my graduate studies and for his inspiration as a teacher and scholar. I wish to thank my collaborator, Professor Weidong Liu, for valuable discussions and for many stimulating ideas. I want to thank my committee member, Professor Zongming Ma. He is always supportive and willing to help. I also want to thank Professors Abba Krieger and Nancy Zhang for their continuous support and very helpful discussions.

My sincere thanks also goes to Professor Charles Epstein, the graduate chair of Applied Mathematics and Computational Science, for offering me this wonderful opportunity to study in this program.

Last but not the least, I would like to thank my parents and my friends, for always supporting me throughout my life.
ABSTRACT

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DATA

Yin Xia

Tony Cai

This thesis considers in the high dimensional setting two canonical testing problems in multivariate analysis, namely testing the equality of two mean vectors and testing the equality of two covariance matrices. The construction of adaptive confidence intervals for regression functions under shape constraints of monotonicity and convexity is also studied.

For testing two mean vectors, we introduce a new test statistic based on a linear transformation of the data by the precision matrix which incorporates the correlations among the variables. Limiting null distribution of the test statistic and the power of the test, both for the case the precision matrix is known and the case it is unknown, are analyzed. The test is particularly powerful against sparse alternatives and enjoys certain optimality. Numerical results show that the proposed test significantly outperforms other tests against sparse alternatives.

For testing two covariance matrices, the limiting null distribution of a new test statistic is derived. The test enjoys certain optimality and is especially powerful against sparse alternatives. Simulation results show that the test significantly outperforms existing methods both in terms of size and power. Analysis of prostate
cancer datasets is carried out to demonstrate the application of the testing procedures. Motivated by applications in genomics, we also consider two related problems, recovering the support of the difference of two covariance matrices and testing the equality of two covariance matrices row by row. New testing procedures are introduced and their properties are studied.

For construction of adaptive confidence intervals, a natural benchmark is established for the minimum expected length of confidence intervals at a given function in terms of an analytic quantity, the local modulus of continuity. This bound depends not only on the function but also the assumed function class. These benchmarks show that the constructed confidence intervals have near minimum expected length for each individual function, while maintaining a given coverage probability for functions within the class. Such adaptivity is much stronger than adaptive minimaxity over a collection of large parameter spaces.
Contents

1 Introduction 1

1.1 High-Dimensional Mean Testing ........................................... 2
1.2 High-Dimensional Covariance Matrix Testing ............................ 7
1.3 Adaptive Confidence Intervals For Regression Functions Under Shape Constraints .......................................................... 11

2 Testing High Dimensional Means 18

2.1 Methodology ........................................................................... 18
2.1.1 Oracle Procedures ................................................................. 19
2.1.2 Data-Driven Procedure .......................................................... 22
2.2 Theoretical Analysis ................................................................. 23
2.2.1 Asymptotic Distributions of the Oracle Test Statistics .......... 24
2.2.2 The Asymptotic Properties of $\Phi_\alpha(\Omega)$ And $\Phi_\alpha(\hat{\Omega})$ ....... 30
2.3 Extension to Non-Gaussian Distributions .................................... 39
2.4 Simulation Study .................................................................... 43
3 Testing High Dimensional Covariance Matrices

3.1 The testing procedure

3.2 Theoretical analysis of size and power

3.2.1 Definitions and assumptions

3.2.2 Limiting null distribution and optimality

3.2.3 Comparison with other tests

3.3 Support recovery of $\Sigma_1 - \Sigma_2$ and application to gene selection

3.3.1 Support recovery of $\Sigma_1 - \Sigma_2$

3.3.2 Testing rows of two covariance matrices

3.4 Numerical results

3.4.1 Support recovery

3.4.2 Real data analysis

3.5 Technical Lemmas

4 Adaptive Confidence Intervals For Regression Functions Under Shape Constraints

4.1 Benchmark and Lower Bound on Expected Length

4.1.1 Benchmark and Lower Bound

4.1.2 Examples of Bounds For Monotone Functions and Convex Functions
4.2 Confidence Procedures ........................................... 108
  4.2.1 Construction ...................................................... 108
  4.2.2 Analysis of the Confidence Intervals ....................... 114
4.3 Nonparametric Regression ....................................... 122
  4.3.1 Monotone Regression ........................................... 123
  4.3.2 Convex Regression .............................................. 124
4.4 Technical Lemmas ................................................... 126
  4.4.1 Proof of Lemma 12 ............................................... 126
  4.4.2 Proof of Lemma 13 ............................................... 127

5 Conclusion .................................................................. 129

6 Future Research .......................................................... 137
  6.1 High Dimensional Sparse Discriminant Analysis ............ 138
  6.2 High Dimensional Testing of Precision Matrices ............ 139
  6.3 Robust and Adaptive Confidence Bands Under Shape Constraints 140

Appendix ....................................................................... 142

A Testing High Dimensional Means .................................. 142
  A.1 Power Comparison of the Oracle Tests ....................... 142
  A.2 Additional Simulation Results ................................. 147
  A.3 Proof of Technical Lemmas ..................................... 148
  A.4 Proof of Propositions ............................................. 152
B Testing High Dimensional Covariance Matrices

B.1 Proof of Lemmas 10-11 .................................................. 166

B.2 Additional simulation results ............................................. 170

B.2.1 Simulation results for non-Gaussian distributions .......... 170

B.2.2 Simulation results for small samples ......................... 174

Reference ................................................................. 178
List of Tables

2.1 Empirical sizes with $\alpha = 0.05$ and $n = 100$. Based on 1000 replications. 47

2.2 Powers of tests with $\alpha = 0.05$ and $n = 100$. Signal strength keeps the same. Based on 1000 replications. 48

2.3 Powers of tests with $\alpha = 0.05$ and $n = 100$. Signal Strength varies. Based on 1000 replications. 49

3.1 $N(0,1)$ random variables. Model 1-4 Empirical sizes and powers in percents. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications. 88

3.2 Average (standard deviation) of $s(\bar{\Psi}, \Psi)$ in percents. 89

A.1 Sizes and Powers of tests with $\alpha = 0.05$ and $n = 100$ for Model 6. $\mu_1 - \mu_2$ has $m = \sqrt{p}$ nonzero elements. Signal strength keeps the same. Based on 1000 replications. 148

B.1 $\Gamma(10,1)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications. 171
B.2 Mixture random variables. Model 1-4 Empirical sizes and powers.

\[ \alpha = 0.05. \ n = 60 \text{ and } 100. \ 5000 \text{ replications.} \]

B.3 $t(12)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = $ 0.05. $n = 60 \text{ and } 100. \ 5000 \text{ replications.}$

B.4 $N(0,1)$ random variables. Model 1-4 Empirical sizes and powers.

\[ \alpha = 0.05. \ n = 30. \ 5000 \text{ replications.} \]

B.5 $\Gamma(10,1)$ random variables. Model 1-4 Empirical sizes and powers.

\[ \alpha = 0.05. \ n = 30. \ 5000 \text{ replications.} \]
List of Figures

3.1 Heat maps of the frequency of the 0s identified for each entry of $\Sigma_1 - \Sigma_2$ ($n = 100$, $p = 50$ for the top row and $p = 100$ for the second row) out of 100 replications. White indicates 1000s identified out of 100 runs; black, 0/100. . . . . . . . . . . . . . . . . . . . . . . . . . 90

3.2 Heat map of the the selected genes by exactly recovery. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.1 Plots of $f_k$ and a least favorable function $g$ in Example 1 with the constraints $|g(0) - f_k(0)| = a$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 103

4.2 Plots of $f$ and a least favorable function $g$ in Example 2 with the constraints $|g(0) - f(0)| = a$. . . . . . . . . . . . . . . . . . . . . . . . . . . 106

4.3 Plots of $f$ and a least favorable function $g$ in Examples 3 and 4 with the constraints $|g(0) - f(0)| = a$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
Chapter 1

Introduction

High-dimensional testing problems are very important in multivariate analysis. Many statistical procedures rely on the fundamental assumption of the equality of two means or the equality of two covariance matrices. In the conventional low dimensional settings, these problems have already been well studied, see, for example, Anderson (2003). But in the high-dimensional settings, those traditional tests do not perform well, and they are even not well defined when the dimensional $p$ is larger than the sample size $n$. Although recently considerable attentions have been received in the high-dimensional setting, most of the new proposed tests have been focusing on the dense cases, where two means or two covariance matrices differ in a lot of entries under the alternative. In this thesis, we are interested in the high-dimensional sparse testing problems, where the data is a high dimensional noisy vector (matrix) with a potentially low dimensional signal. In those cases,
the current existing tests would always suffer from low power. Thus, we develop optimal procedures to deal with each of these two testing problems under the high-dimensional sparse alternatives, which are commonly assumed in many recent large dataset.

The construction of useful confidence sets is one of the more challenging problems in nonparametric function estimation. My focus has been on the construction of confidence sets for the underlying regression or density function in settings where the unknown function obeys particular shape constraints. Such constraints are often natural in many applications. In this thesis, a data driven procedure is constructed and shown to be adaptive to each and every function in the parameter space. This confidence interval has the smallest expected length, up to a universal constant factor, for individual convex (or monotone) functions within the class of all confidence intervals which guarantee a $1 - \alpha$ coverage probability over all convex (or monotone) functions. Such adaptivity is much stronger than the classical adaptive minimaxity over a collection of large parameter spaces.

In the remainder of this chapter, we define each of our problems and give a quick overview of Chapters 2 - 4.

### 1.1 High-Dimensional Mean Testing

One canonical testing problem is that of testing the equality of two mean vectors $\mu_1$ and $\mu_2$ based on independent random samples, one from a distribution with mean
and covariance matrix $\Sigma$ and another from a distribution with mean $\mu_2$ and the same covariance matrix $\Sigma$. This testing problem arises in many scientific applications, including genetics, econometrics, and signal processing. In the Gaussian setting where one observes $X_k \overset{iid}{\sim} N_p(\mu_1, \Sigma), k = 1, \ldots, n_1$, and $Y_k \overset{iid}{\sim} N_p(\mu_2, \Sigma), k = 1, \ldots, n_2$, the classical test for testing the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

is Hotelling’s $T^2$ test with the test statistic given by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X} - \bar{Y})' \tilde{\Sigma}^{-1} (\bar{X} - \bar{Y}),$$

where $\bar{X} = n_1^{-1} \sum_{k=1}^{n_1} X_k$ and $\bar{Y} = n_2^{-1} \sum_{k=1}^{n_2} Y_k$ are the sample means and $\tilde{\Sigma}$ is the sample covariance matrix. The properties of Hotelling’s $T^2$ test has been well studied in the conventional low-dimensional setting. It enjoys desirable properties when the dimension $p$ is fixed. See, e.g., Anderson (2003).

In many contemporary applications, high dimensional data, whose dimension is often comparable to or even much larger than the sample size, are commonly available. Examples include genomics, medical imaging, risk management, and web search problems. In such high dimensional settings, classical methods designed for the low-dimensional case either perform poorly or are no longer applicable. For example, the performance of Hotelling’s $T^2$ test is unsatisfactory when the dimension is high relative to the sample sizes.

Several proposals for correcting Hotelling’s $T^2$ statistic have been introduced in the high dimensional settings. For example, Bai and Saranadasa (1996) proposed
to remove $\hat{\Sigma}^{-1}$ in $T^2$ and introduced a new statistic based on the squared Euclidean norm $\|\bar{X} - \bar{Y}\|_2^2$. Srivastava and Du (2008) and Srivastava (2009) constructed test statistics by replacing $\hat{\Sigma}^{-1}$ with the inverse of the diagonal of $\hat{\Sigma}$. Chen and Qin (2010) introduced a test statistic by removing the terms $\sum_{i=1}^{n_1} X'_i X_i$ and $\sum_{i=1}^{n_2} Y'_i Y_i$ in $\|\bar{X} - \bar{Y}\|_2^2$. All of the above test statistics are based on an estimator of $(\mu_1 - \mu_2)^T A (\mu_1 - \mu_2)$ for some given positive definite matrix $A$. We shall call these test statistics sum of squares type statistics as they all aim to estimate the squared Euclidean norm $\|A^{1/2}(\mu_1 - \mu_2)\|_2^2$.

It is known that tests based on the sum of squares type statistics can have good power against the “dense” alternatives. That is, under the alternative hypothesis $H_1$, the signals in $\mu_1 - \mu_2$ spread out over a large number of coordinates. In anomaly detection, medical imaging, genomics and many other applications, however, the means of the two populations are typically either identical or are quite similar in the sense that they only possibly differ in a small number of coordinates. In other words, under the alternative $H_1$, the difference of the two means $\mu_1 - \mu_2$ is sparse. For example, for ultrasonic flaw detection in highly-scattering materials, many scattering centers such as grain boundaries produce echoes and the ensemble of these echoes is usually defined as background noise, while small cracks, flaws, or other metallurgical defects would be defined as signals. See, for example, Zhang, Zhang and Wang (2000). In this case, it is natural to take $\mu_1 - \mu_2$ to be the sparse signals when the metallurgical defects exist. Similarly, for detection of hydrocarbons
in materials, instantaneous spectral analysis is often used to detect hydrocarbons through low-frequency shadows, which is usually considered as sparse signals. See Castagna, Sun and Siegfried (2003). In medical imaging, MRI is commonly used for breast cancer detection. It is used to visualize microcalcifications, which can be an indication of breast cancer. The signals are rare in such applications, see James, Clymer and Schmalbrock (2001). Another application is the shape analysis of brain structures, in which the shape differences, if any, are commonly assumed to be confined to a small number of isolated regions inside the whole brain. This is equivalent to the sparse alternative. See Cao and Worsley (1999) and Taylor and Worsley (2008). In these sparse settings, tests based on the sum of squares type statistics are not powerful. For example, the three tests mentioned earlier all require \((n_1 + n_2)\|\mu_1 - \mu_2\|^2/\sigma^2 \to \infty\) in order for any of the tests to be able to distinguish between the null and the alternative with probability tending to 1.

The goal of Chapter 2 is to develop a test that is powerful against sparse alternatives in the high dimensional setting under dependency. To explore the advantages of the dependence between the variables, we introduce in this thesis a new test statistic that is based on a linear transformation of the observations by the precision matrix \(\Omega\). Suppose for the moment the precision matrix \(\Omega = \Sigma^{-1}\) is known. For testing the null hypothesis \(H_0 : \mu_1 = \mu_2\), we first transform the samples \(\{X_k; 1 \leq k \leq n_1\}\) and \(\{Y_k; 1 \leq k \leq n_2\}\) by multiplying with \(\Omega\) to obtain the transformed samples \(\{\Omega X_k; 1 \leq k \leq n_1\}\) and \(\{\Omega Y_k; 1 \leq k \leq n_2\}\). The new test
statistic is then defined to be the maximum of the squared two sample $t$-statistics of the transformed observations $\{\Omega X_k; 1 \leq k \leq n_1\}$ and $\{\Omega Y_k; 1 \leq k \leq n_2\}$. We shall first show that the limiting null distribution of this test statistic is the extreme value distribution of type I, and then construct an asymptotically $\alpha$ level test based on the limiting distribution. It will be shown that this test enjoys certain optimality and uniformly outperforms two other natural tests against sparse alternatives. The asymptotic properties including the power of the tests are investigated in Section 2.2.

The precision matrix $\Omega$ is typically unknown in practice and thus needs to be estimated. In order to get a good estimator of $\Omega$, in this thesis we focus on the case where $\Omega$ is sparse. This is a natural assumption in many applications including the Gaussian graphical models. See, e.g., Ravikumar et al. (2008), Yuan (2010), and Cai, Liu, and Luo (2011). In this case, we estimate $\Omega$ by the constrained $\ell_1$ minimization approach proposed by Cai, Liu, and Luo (2011) and then plug into the test statistic mentioned above. In principle, other “good” estimators of $\Omega$ can also be used. We show that, under regularity conditions, the test based on this data-driven test statistic performs asymptotically as well as the test based on the oracle statistic and thus shares the same optimality. A simulation study is carried out to examine the numerical performance of these tests and compare them with other tests proposed in the literature. The numerical results show that the proposed test significantly outperforms those based on the sum of squares type statistics when
$\mathbf{\mu}_1 - \mathbf{\mu}_2$ is sparse under the alternative.

### 1.2 High-Dimensional Covariance Matrix Testing

Another important testing problem is that of testing the equality of two covariance matrices, $\Sigma_1$ and $\Sigma_2$. Many statistical procedures including the classical Fisher's linear discriminant analysis rely on the fundamental assumption of equal covariance matrices. This testing problem has been well studied in the conventional low-dimensional setting. See, for example, Sugiura and Nagao (1968), Gupta and Giri (1973), Perlman (1980), Gupta and Tang (1984), O'Brien (1992), and Anderson (2003). In particular, the likelihood ratio test (LRT) is commonly used and enjoys certain optimality under regularity conditions.

Driven by a wide range of contemporary scientific applications, analysis of high dimensional data is of significant current interest. In the high dimensional setting, where the dimension can be much larger than the sample size, the conventional testing procedures such as the LRT either perform poorly or are not even well defined. Several tests for the equality of two large covariance matrices have been proposed. For example, Schott (2007) introduced a test based on the Frobenius norm of the difference of the two covariance matrices. Srivastava and Yanagihara (2010) constructed a test that relied on a measure of distance by $\frac{\text{tr}(\Sigma^2_1)/\text{tr}(\Sigma_1)^2}{\text{tr}(\Sigma^2_2)/\text{tr}(\Sigma_2)^2}$ - Both of these two tests are designed for the multivariate normal populations. Li and Chen (2012) proposed a test using a linear combination of three
one-sample $U$-statistics which was also motivated by an unbiased estimator of the Frobenius norm of $\Sigma_1 - \Sigma_2$.

In many applications such as gene selection in genomics, the covariance matrices of the two populations can be either equal or quite similar in the sense that they only possibly differ in a small number of entries. In such a setting, under the alternative the difference of the two covariance matrices $\Sigma_1 - \Sigma_2$ is sparse. The above mentioned tests, which are all based on the Frobenius norm, are not powerful against such sparse alternatives. See Section 3.2.3 for further details.

The first goal of Chapter 3 is to develop a test that is powerful against sparse alternatives and robust with respect to the population distributions. Let $X$ and $Y$ be two $p$ variate random vectors with covariance matrices $\Sigma_1$ and $\Sigma_2$ respectively. Let $\{X_1, \ldots, X_{n_1}\}$ be i.i.d. random samples from $X$ and let $\{Y_1, \ldots, Y_{n_2}\}$ be i.i.d. random samples from $Y$ that are independent of $\{X_1, \ldots, X_{n_1}\}$. We wish to test the hypotheses

$$H_0 : \Sigma_1 = \Sigma_2 \quad \text{versus} \quad H_1 : \Sigma_1 \neq \Sigma_2 \quad (1.2.1)$$

based on the two samples. We are particularly interested in the high dimensional setting where $p$ can be much larger than $n = \max(n_1, n_2)$. In many applications, if the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$ is rejected, it is often of significant interest to further investigate in which way the two covariance matrices differ from each other. Motivated by applications to gene selection, in this thesis we also consider two related problems, recovering the support of $\Sigma_1 - \Sigma_2$ and testing the equality
of the two covariance matrices row by row.

We propose a test for the hypotheses in (1.2.1) based on the maximum of the standardized differences between the entries of the two sample covariance matrices and investigate its theoretical and numerical properties. The limiting null distribution of the test statistic is derived. It is shown that the distribution of the test statistic converges to a type I extreme value distribution under the null $H_0$. This fact implies that the proposed test has the pre-specified significance level asymptotically. The power of the test is investigated. The theoretical analysis shows that the proposed test enjoys certain optimality against a large class of sparse alternatives in terms of the power. We show that it only requires one of the entries of $\Sigma_1 - \Sigma_2$ having a magnitude more than $C \sqrt{\log p/n}$ in order for the test to correctly reject the null hypothesis $H_0$. It is also shown that this lower bound $C \sqrt{\log p/n}$ is rate-optimal.

In addition to the theoretical properties, we also consider the numerical performance of the proposed testing procedure using both simulated and real datasets. The numerical results show that the new test significantly outperforms the existing methods both in terms of size and power. A prostate cancer dataset is used to illustrate our testing procedures for gene selection.

In addition to the global test of equal covariance matrices, we also consider recovery of the support of $\Sigma_1 - \Sigma_2$ as well as testing $\Sigma_1$ and $\Sigma_2$ row by row. Support recovery can also be viewed as simultaneous testing of the equality of individual
entries between the two covariance matrices. We introduce a procedure for support
recovery based on the thresholding of the standardized differences between the
entries of the two covariance matrices. It is shown that under certain conditions,
the procedure recovers the true support of $\Sigma_1 - \Sigma_2$ exactly with probability tending
to 1. The procedure is also shown to be minimax rate optimal.

The problem of testing the equality of two covariance matrices row by row is
motivated by applications in genomics. A commonly used approach in microarray
analysis is to select “interesting” genes by applying multiple testing procedures
on the two-sample $t$-statistics. This approach has been successful in finding genes
with significant changes in the mean expression levels between diseased and non-
diseased populations. It has been noted recently that these mean-based methods
lack the ability to discover genes that change their relationships with other genes
and new methods that are based on the change in the gene’s dependence structure
are thus needed to identify these genes. See, for example, Ho, et al. (2008), Hu,
et al. (2009) and Hu, et al. (2010). In this thesis, we propose a procedure which
simultaneously tests the $p$ null hypotheses that the corresponding rows of the two
covariance matrices are equal to each other. Asymptotic null distribution of the
test statistics is derived and properties of the test are studied. It is shown that the
procedure controls the family-wise error rate at a pre-specified level. Applications
to gene selection is also considered.
1.3 Adaptive Confidence Intervals For Regression Functions Under Shape Constraints

The construction of useful confidence sets is one of the more challenging problems in nonparametric function estimation. There are two main interrelated issues which need to be considered together, coverage probability and the expected size of the confidence set. For a fixed parameter space it is often possible to construct confidence sets which have guaranteed coverage probability over the parameter space while controlling the maximum expected size. However such minimax statements are often thought to be too conservative and a more natural goal is to have the expected size of the confidence set reflect in some sense the difficulty of estimating the particular underlying function.

These issues are well illustrated by considering confidence intervals for the value of a function at a fixed point. Let $Y$ be an observation from the white noise model:

$$dY(t) = f(t)dt + n^{-1/2}dW(t), \quad -\frac{1}{2} \leq t \leq \frac{1}{2} \quad (1.3.1)$$

where $W(t)$ is standard Brownian motion and $f$ belongs to some parameter space $\mathcal{F}$. Suppose that we wish to construct a confidence interval for $f$ at some point $t_0 \in (-\frac{1}{2}, \frac{1}{2})$. Let $CI$ be a confidence interval for $f(t_0)$ based on observing the process $Y$ and let $L(CI)$ denote the length of the confidence interval. The minimax point of view can then be expressed by: subject to the constraint on the coverage probability $\inf_{f \in \mathcal{F}} P(f(t_0) \in CI) \geq 1 - \alpha$, minimize the maximum expected length.
sup_{f \in \mathcal{F}} E_f(L(CI)).

As an example it is common to consider the Lipschitz classes

\[ \Lambda(\beta, M) = \{ f : |f(y) - f(x)| \leq M|y - x|^{\beta} \text{ for } x, y \in [-\frac{1}{2}, \frac{1}{2}] \}, \quad \text{if } 0 < \beta \leq 1 \]

and for \( \beta > 1 \)

\[ \Lambda(\beta, M) = \{ f : |f^{([\beta])}(x) - f^{([\beta])}(y)| \leq M|x - y|^{\beta'} \text{ for } x, y \in [-\frac{1}{2}, \frac{1}{2}] \}, \]

where \([\beta]\) is the largest integer less than \( \beta \) and \( \beta' = \beta - [\beta] \). For these classes it easily follows from results of Donoho (1994), Low (1997), and Evans, Hansen and Stark (2005) that the minimax expected length of confidence intervals, which have guaranteed coverage of \( 1 - \alpha \) over \( \Lambda(\beta, M) \), is of order \( M^{\frac{1}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}} \).

It should however be stressed that confidence intervals which achieve such an expected length rely on the knowledge of the particular smoothness parameters \( \beta \) and \( M \), which are not known in most applications. Unfortunately, Low (1997) and Cai and Low (2004) have shown that the natural goal of constructing an adaptive confidence interval which has a given coverage probability and has expected length that is simultaneously close to these minimax expected lengths for a range of smoothness parameters is not in general attainable. More specifically suppose that a confidence interval has guaranteed coverage probability of \( 1 - \alpha \) over \( \Lambda(\beta, M) \). Then for any \( f \) in the interior of \( \Lambda(\beta, M) \) the expected length for this \( f \) must also be of order \( n^{-\frac{\beta}{1+2\beta}} \). In other words the minimax rate describes the actual rate for all functions in the class other than those on the boundary of the set. For example
in the case that a confidence interval has guaranteed coverage probability of \(1 - \alpha\) over the Lipschitz class \(\Lambda(1, M)\) then even if the underlying function has two derivatives and first derivative smaller than \(M\), the confidence interval for \(f(x)\) must still have expected length of order \(n^{-1/3}\) even though one would hope that an adaptive confidence interval would have a much shorter length of order \(n^{-2/5}\).

Despite these very negative results there are some settings where some degree of adaptation has been shown to be possible. In particular under certain shape constraints Hengartner and Stark (1995) constructed confidence bands which have a guaranteed coverage probability of at least \(1 - \alpha\) over the collection of all monotone densities and which have maximum expected length of order \(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta + 1}}\) for those monotone densities which are in \(\Lambda(\beta, M)\) for a particular choice of \(\beta\) where \(0 < \beta \leq 1\). This construction relies on the selection of a tuning parameter and is thus not adaptive. Dümbgen (2003) however does provide adaptive confidence bands with optimal rates for both isotonic and convex functions under supremum norm loss on arbitrary compact subintervals. These results are however still framed in terms of the maximum length over particular large parameter spaces and the existence of such intervals raises the question of exactly how much adaption is possible. It is this question that is the focus of Chapter 4 of this thesis.

Rather than considering the maximum expected length over large collections of functions we study the problem of adaptation to each and every function in the parameter space. We examine this problem in detail for two commonly used
collections of functions that have shape constraints, namely the collection of convex functions and the collection of monotone functions. We focus on these parameter spaces as it is for such shape constrained problems for which there is some hope for adaptation. Within this context we consider the problem of constructing a confidence interval for the value of a function at a fixed point under both the white noise with drift model given in (1.3.1) as well as a nonparametric regression model. We show that within the class of convex functions and the class of monotone functions it is indeed possible to adapt to each individual function, and not just to the minimax expected length over different parameter spaces in a collection. The notion of adaptivity to a single function is also discussed in Lepski, Mammen and Spokoiny (1997) and Lepski and Spokoiny (1997) for the related point estimation problem but in that context a logarithmic penalty of the noise level must be paid and so in that context the notion of adaptivity is somewhat different.

This result is achieved in two steps. First we study the problem of minimizing the expected length of a confidence interval assuming that the data is generated from a particular function \( f \) in the parameter space, subject to the constraint that the confidence interval has guaranteed coverage probability over the entire parameter space. The solution to this problem gives a benchmark for the expected length which depends on the function \( f \) considered. It gives a bound on the expected length of any adaptive interval because if the expected length is smaller than this bound for any particular function, the confidence interval cannot have the desired coverage
probability. In applications it is more useful to express the benchmark in terms of a local modulus of continuity, an analytic quantity that can be easily calculated for individual functions. In situations where adaptation is not possible this local modulus of continuity does not vary significantly from function to function. Such is the case in the settings considered in Low (1997). However in the context of convex or monotone functions the resulting benchmark does vary significantly and this opens up the possibility for adaptation in those settings.

Our second step is to actually construct adaptive confidence intervals. This is done separately for monotone functions and convex functions, with similar results. For example, an adaptive confidence interval is constructed which is shown to have expected length uniformly within an absolute constant factor of the benchmark for every convex function, while maintaining coverage probability over the collection of all convex functions. In other words, this confidence interval has smallest expected length, up to a universal constant factor, for each and every convex function within the class of all confidence intervals which guarantee a $1 - \alpha$ coverage probability over all convex functions. A similar result is established for a confidence interval designed for monotone functions.

The rest of this thesis is organized as follows.

In Chapter 2, we discuss the problem of testing two high-dimensional means. After reviewing basic notations and definitions, Section 2.1 introduces the new test statistics. Theoretical properties of the proposed tests are investigated in Section
2.2. Limiting null distributions of the test statistics and the power of the tests, both for the case the precision matrix \( \Omega \) is known and the case \( \Omega \) is unknown, are analyzed. Extensions to the non-Gaussian distributions are given in Section 2.3. A simulation study is carried out in Section 2.4 to investigate the numerical performance of the tests. The technical lemmas are collected in Section 2.5.

Two-sample covariance matrix testing problem is discussed in Chapter 3. In this chapter, Section 3.1 introduces the procedure for testing the equality of two covariance matrices. Theoretical properties of the test are investigated in Section 3.2. After Section 3.2.1 in which basic definitions and assumptions are given, Section 3.2.2 develops the asymptotic null distribution of the test statistic and presents the optimality results for testing against sparse alternatives. Comparisons with other tests are given in Section 3.2.3. Section 3.3 considers support recovery of \( \Sigma_1 - \Sigma_2 \) and testing the two covariance matrices row by row. Section 3.4 investigates the numerical performance of the proposed test by simulations and by an analysis of a prostate cancer dataset. The technical lemmas are collected in Section 3.5.

Chapter 4 considers the construction of adaptive confidence intervals under shape constraints. In section 4.1 the benchmark for the expected length at each monotone function or each convex function is established under the constraint that the interval has a given level of coverage probability over the collection of monotone functions or the collection of convex functions. Section 4.2 constructs data driven confidence intervals for both monotone functions and convex functions and shows
that these confidence intervals maintain coverage probability and have expected
length within an absolute constant factor of the benchmark given in Section 4.1 for
each monotone function and convex function. Section 4.3 considers the nonpara-
metric regression model and technical lemmas are collected in Section 4.4.

Chapter 5 discusses our results and other related work in the literature and
Chapter 6 summarizes the future research. More technical details and extensive
simulation results are collected in Appendices A and B.
Chapter 2

Testing High Dimensional Means

2.1 Methodology

This section considers the testing problem in the setting of Gaussian distributions. Extensions to the non-Gaussian case will be discussed in Section 2.3. We shall first present our testing procedure in the oracle setting in Section 2.1.1 where the precision matrix $\Omega$ is assumed to be known. In addition, two other natural testing procedures are introduced in this setting. A data-driven procedure is given in Section 2.1.2 for the general case of the unknown precision matrix $\Omega$ by using an estimator of the precision matrix $\Omega$ through the constrained $\ell_1$ minimization method.

We begin with basic notations and definitions. For a vector $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$, define the $\ell_q$ norm by $|\beta|_q = (\sum_{i=1}^p |\beta_i|^q)^{1/q}$ for $1 \leq q \leq \infty$ with the usual
modification for $q = \infty$. A vector $\beta$ is called $k$-sparse if it has at most $k$ nonzero entries. For a matrix $\Omega = (\omega_{ij})_{p \times p}$, the matrix 1-norm is the maximum absolute column sum, $\|\Omega\|_{L_1} = \max_{1 \leq j \leq p} \sum_{i=1}^{p} |\omega_{ij}|$, the matrix elementwise infinity norm is defined to be $|\Omega|_{\infty} = \max_{1 \leq i, j \leq p} |\omega_{ij}|$ and the elementwise $\ell_1$ norm is $\|\Omega\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |\omega_{ij}|$. For a matrix $\Omega$, we say $\Omega$ is $k$-sparse if each row/column has at most $k$ nonzero entries. We shall denote the difference $\mu_1 - \mu_2$ by $\delta$ so the null hypothesis can be equivalently written as $H_0 : \delta = 0$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant $C$ such that $|a_n| \leq C|b_n|$ holds for all sufficiently large $n$, write $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n / b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants $c$ and $C$ such that $c \leq a_n / b_n \leq C$ for all $n \geq 1$.

2.1.1 Oracle Procedures

Suppose we observe independent $p$-dimensional random samples

$$X_1, \ldots, X_{n_1} \sim iid N(\mu_1, \Sigma) \text{ and } Y_1, \ldots, Y_{n_2} \sim iid N(\mu_2, \Sigma)$$

where the precision matrix $\Omega = \Sigma^{-1}$ is known. In this case, the null hypothesis $H_0 : \delta = 0$ is equivalent to $H_0 : \Omega \delta = 0$. An unbiased estimator of $\Omega \delta$ is the sample mean vector $\Omega(\bar{X} - \bar{Y}) =: \bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_p)^T$. We propose to test the null hypothesis $H_0 : \delta = 0$ based on the test statistic

$$M_\Omega = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{\bar{Z}_i^2}{\omega_{ii}},$$

(2.1.1)
At first sight, the test statistic $M_{\Omega}$ is not the most intuitive choice for testing $H_0 : \delta = 0$. We first briefly illustrate the motivation on the linear transformation of the data by the precision matrix. Under a sparse alternative, the power of a test mainly depends on the magnitudes of the signals (nonzero coordinates of $\delta$) and the number of the signals. It will be shown in Section 2.5 that $(\Omega \delta)_i$ is approximately equal to $\delta_i \omega_{ii}$ for all $i$ in the support of $\delta$. The magnitudes of the nonzero signals $\delta_i$ are then transformed to $|\delta_i| \omega_{ii}^{1/2}$ after normalized by the standard deviation of the transformed variable $(\Omega X)_i$. In comparison, the magnitudes of the signals in the original data are $|\delta_i| / \sigma_{ii}^{1/2}$. It can be seen from the elementary inequality $\omega_{ii} \sigma_{ii} \geq 1$ for $1 \leq i \leq p$ that $|\delta_i| \omega_{ii}^{1/2} \geq |\delta_i| / \sigma_{ii}^{1/2}$. That is, such a linear transformation magnifies the signals and the number of the signals due to the dependence in the data. The transformation thus helps to distinguish the null and alternative hypotheses. The advantage of this linear transformation will be proved rigorously in Section 2.5.

The asymptotic null distribution of $M_{\Omega}$ will be studied in Section 2.2. Note that $M_{\Omega}$ is the maximum of $p$ dependent normal random variables. It is well known that the limiting distribution of the maximum of $p$ independent $\chi_1^2$ random variables after normalization is the extreme value distribution of type I. This result was generalized by Berman (1964) to the dependent case, where the limiting distribution for the maximum of a stationary sequence was considered. In the setting of this thesis, the precision matrix $\Omega$ does not have any natural order and the result in Berman (1964) thus does not apply. We shall prove by using different techniques that $M_{\Omega}$
still converges to the extreme value distribution of type I under the null $H_0$.

More generally, for a given invertible $p \times p$ matrix $A$, the null hypothesis $H_0 : \delta = 0$ is equivalent to $H_0 : A\delta = 0$. Set $\delta^A = (\delta^A_1, ..., \delta^A_p)' := A(X - Y)$. Denote the covariance matrix of $AX$ by $B = (b_{ij})$ and define the test statistic

$$M_A = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{(\delta^A_i)^2}{b_{ii}}. \quad (2.1.2)$$

The most natural choices of $A$ are arguably $A = \Omega^\sharp_1$ and $A = I$. In the case of $A = \Omega^\sharp_1$, the components of $\Omega^\sharp_1 X$ and $\Omega^\sharp_1 Y$ are independent. Set $\bar{W} = (\bar{W}_1, ..., \bar{W}_p)^T := \Omega^\sharp_1 (\bar{X} - \bar{Y})$. It is natural to consider the test statistic

$$M_{\Omega^\sharp_1} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \bar{W}_i^2. \quad (2.1.3)$$

As we will show later that the test based on $M_\Omega$ uniformly outperforms the test based on $M_{\Omega^\sharp_1}$ for testing against sparse alternatives.

Another natural choice is $A = I$. That is, the test is directly based on the difference of the sample means $\bar{X} - \bar{Y}$. Set $\bar{\delta} = (\bar{\delta}_1, ..., \bar{\delta}_p)' := \bar{X} - \bar{Y}$ and define the test statistic

$$M_I = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{\bar{\delta}_i^2}{\sigma_{ii}} \quad (2.1.4)$$

where $\sigma_{ii}$ are the diagonal elements of $\Sigma$. Here $M_I$ is the maximum of the squared two sample $t$ statistics based on the samples $\{X_k\}$ and $\{Y_k\}$ directly. It will be shown that the test based on $M_I$ is uniformly outperformed by the test based on $M_\Omega$ for testing against sparse alternatives.
2.1.2 Data-Driven Procedure

We have so far focused on the oracle case in which the precision matrix $\Omega$ is known. However, in most applications $\Omega$ is unknown and thus needs to be estimated. Estimation of the precision matrix $\Omega$ under the sparsity assumption has attracted much recent attention, see, for example, Yuan and Lin (2007); d’Aspremont, Banerjee and El Ghaoui (2008); Friedman, Hastie and Tibshirani (2008); Rothman et al. (2008); Yuan (2010); and Cai, Liu, and Luo (2011).

In this thesis, we use the constrained $\ell_1$ minimization method given in Cai, Liu, and Luo (2011) to estimate $\Omega$. Let $\hat{\Omega}_1 = (\hat{\omega}^1_{ij})$ be a solution of the following optimization problem:

$$\min \|\Omega\|_1 \text{ subject to } |\Sigma_n \Omega - I|_{\infty} \leq \lambda_n,$$

where $\| \cdot \|_1$ is the elementwise $\ell_1$ norm defined at the beginning of Section 2.1, and $\lambda_n = C \sqrt{\log p/n}$ for some sufficiently large constant $C$. In practice, $\lambda_n$ can be chosen through cross validation. See Cai, Liu, and Luo (2011) for further details.

The estimator of the precision matrix $\Omega$ is defined to be $\hat{\Omega} = (\hat{\omega}_{ij})_{p \times p}$, where

$$\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}^1_{ij} I\{|\hat{\omega}^1_{ij}| \leq |\hat{\omega}^1_{ji}|\} + \hat{\omega}^1_{ji} I\{|\hat{\omega}^1_{ij}| > |\hat{\omega}^1_{ji}|\}.$$

The estimator $\hat{\Omega}$ is called the CLIME estimator and can be implemented by linear programming. It enjoys desirable theoretical and numerical properties. See Cai, Liu, and Luo (2011) for more details on the properties and implementation of this estimator.
For testing the hypothesis $H_0 : \mu_1 = \mu_2$ in the case of unknown precision matrix $\Omega$, motivated by the oracle procedure $M_{\Omega}$ given in Section 2.1.1, our final test statistic is $M_{\bar{\Omega}}$ defined by

$$M_{\bar{\Omega}} = \frac{n_1n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{\hat{Z}_i^2}{\hat{\omega}_{ii}^{(0)}},$$

(2.1.5)

where $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_p)^T := \hat{\Omega}(\bar{X} - \bar{Y})$ and $\hat{\omega}_{ii}^{(0)} = \frac{n_1}{n_1 + n_2} \hat{\omega}_{ii}^{(1)} + \frac{n_2}{n_1 + n_2} \hat{\omega}_{ii}^{(2)}$ with

$$\hat{\omega}_{ij}^{(1)} := \frac{1}{n_1} \sum_{k=1}^{n_1} (\hat{\Omega}X_k - \bar{X}_{\hat{\Omega}})(\hat{\Omega}X_k - \bar{X}_{\hat{\Omega}})^T,$$

$$\hat{\omega}_{ij}^{(2)} := \frac{1}{n_2} \sum_{k=1}^{n_2} (\hat{\Omega}Y_k - \bar{Y}_{\hat{\Omega}})(\hat{\Omega}Y_k - \bar{Y}_{\hat{\Omega}})^T,$$

$$X_{\bar{\Omega}} = n_1^{-1} \sum_{k=1}^{n_1} \hat{\Omega}X_k, \quad Y_{\bar{\Omega}} = n_2^{-1} \sum_{k=1}^{n_2} \hat{\Omega}Y_k.$$  (2.1.6)

We will show in Section 2.2 that $M_{\bar{\Omega}}$ and $M_{\Omega}$ have the same asymptotic null distribution and power under certain sparsity conditions on $\Omega$. The simulation results in Section 2.4 show that the numerical performance of the test based on $M_{\bar{\Omega}}$ is similar to that of the test based on $M_{\Omega}$. Note that other estimators of the precision matrix $\Omega$ can also be used to construct a good test. See more discussions in Section 2.2.2.

### 2.2 Theoretical Analysis

We now turn to the analysis of the properties of $M_{\Omega}$ and $M_{\bar{\Omega}}$ including the limiting null distribution and the power of the corresponding tests. We will show that the test based on $M_{\Omega}$ enjoys certain optimality when testing against sparse alternatives. Moreover, under sparsity conditions on the precision matrix $\Omega$ the test based on
$M_\Omega$ performs as well as that based on $M_\Omega$ and thus shares the same optimality. The asymptotic null distributions of $M_{\Omega^2}$ and $M_I$ are also derived. We are particular interested in the comparison of power of the test based on $M_\Omega$ and that of the tests based on $M_{\Omega^2}$ and $M_I$. We will show that the test based on $M_\Omega$ is asymptotically more powerful than those based on $M_{\Omega^2}$ and $M_I$ when $\mu_1 - \mu_2$ is sparse.

2.2.1 Asymptotic Distributions of the Oracle Test Statistics

We first establish the asymptotic null distributions for the oracle test statistics $M_\Omega$, $M_{\Omega^2}$ and $M_I$. Let $D_1 = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ and $D_2 = \text{diag}(\omega_{11}, \ldots, \omega_{pp})$, where $\sigma_{kk}$ and $\omega_{kk}$ are the diagonal entries of $\Sigma$ and $\Omega$ respectively. The correlation matrix of $X$ and $Y$ is then $\Gamma = (\gamma_{ij}) = D_1^{-1/2} \Sigma D_1^{-1/2}$ and the correlation matrix of $\Omega X$ and $\Omega Y$ is $R = (r_{ij}) = D_2^{-1/2} \Omega D_2^{-1/2}$. To obtain the limiting null distributions, we assume that the eigenvalues of the covariance matrix $\Sigma$ are bounded from above and below, and the correlations in $\Gamma$ and $R$ are bounded away from $-1$ and $1$. More specifically we assume the following:

(C1): $C_0^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$ for some constant $C_0 > 0$;

(C2): $\max_{1 \leq i < j \leq p} |\gamma_{ij}| \leq r_1 < 1$ for some constant $0 < r_1 < 1$;

(C3): $\max_{1 \leq i < j \leq p} |r_{ij}| \leq r_2 < 1$ for some constant $0 < r_2 < 1$.

Condition (C1) on the eigenvalues is a common assumption in high-dimensional setting. Conditions (C2) and (C3) are also mild. For example, if $\max_{1 \leq i < j \leq p} |r_{ij}| = 1$, then $\Sigma$ is singular. The following theorem states the asymptotic null distributions
for the three oracle statistics $M_\Omega$, $M_{\Omega^\perp}$ and $M_I$.

**Theorem 1.** Let the test statistics $M_\Omega$, $M_{\Omega^\perp}$ and $M_I$ be defined as in (2.1.1), (2.1.3) and (2.1.4), respectively.

(i). Suppose that (C1) and (C3) hold. Then for any $x \in \mathbb{R}$, we have, as $p \to \infty$,

$$
P_{H_0}(M_\Omega - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right).$$

(ii). For any $x \in \mathbb{R}$, we have, as $p \to \infty$,

$$
P_{H_0}(M_{\Omega^\perp} - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right).$$

(iii). Suppose that (C1) and (C2) hold. Then for any $x \in \mathbb{R}$, we have, as $p \to \infty$,

$$
P_{H_0}(M_I - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right).$$

Proof of Theorem 1: Let $(Z_1, \ldots, Z_p)'$ be a zero mean multivariate normal random vector with covariance matrix $\Omega = (\omega_{ij})_{1 \leq i, j \leq p}$ and the diagonal $\omega_{ii} = 1$ for $1 \leq i \leq p$. To prove Theorem 1, it suffices to prove the following lemma.

**Lemma 1.** Suppose that $\max_{1 \leq i \neq j \leq p} |\omega_{ij}| \leq r < 1$ and $\lambda_{\max}(\Omega) \leq C_0$. Then for any $x \in \mathbb{R}$ as $p \to \infty$

$$
P\left(\max_{1 \leq i \leq p} Z_i^2 - 2 \log p + \log \log p \leq x\right) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp(-x/2)\right), \quad (2.2.1)$$

$$
P\left(\max_{1 \leq i \leq p} Z_i \leq \sqrt{2 \log p - \log \log p + x}\right) \to \exp\left(-\frac{1}{2 \sqrt{\pi}} \exp(-x/2)\right) \quad (2.2.2)$$
We only need to prove (2.2.1) because the proof of (2.2.2) is similar. Set \( x_p = 2 \log p - \log \log p + x \). By Lemma 2, we have for any fixed \( k \leq \lfloor p/2 \rfloor \),

\[
\sum_{t=1}^{2k} (-1)^{t-1} E_t \leq p \left( \max_{1 \leq i \leq p} |Z_i| \geq \sqrt{x_p} \right) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} E_t, \tag{2.2.3}
\]

where

\[
E_t = \sum_{1 \leq i_1 < \cdots < i_t \leq p} P \left( |Z_{i_1}| \geq \sqrt{x_p}, \cdots, |Z_{i_t}| \geq \sqrt{x_p} \right) =: \sum_{1 \leq i_1 < \cdots < i_t \leq p} P_{i_1, \cdots, i_t}.
\]

Define \( \mathcal{I} = \left\{ 1 \leq i_1 < \cdots < i_t \leq p : \max_{1 \leq k < l \leq t} |\text{Cov}(Z_{i_k}, Z_{i_l})| \geq p^{-\gamma} \right\} \), where \( \gamma > 0 \) is a sufficiently small number to be specified later. For \( 2 \leq d \leq t - 1 \), define

\[
\mathcal{I}_d = \left\{ 1 \leq i_1 < \cdots < i_t \leq p : \text{Card}(S) = d, \text{ where } S \text{ is the largest subset of } \{i_1, \ldots, i_t\} \text{ such that } \forall i_k \neq i_t \in S, |\text{Cov}(Z_{i_k}, Z_{i_t})| < p^{-\gamma} \right\}.
\]

For \( d = 1 \), define

\[
\mathcal{I}_1 = \left\{ 1 \leq i_1 < \cdots < i_t \leq p : |\text{Cov}(Z_{i_k}, Z_{i_l})| \geq p^{-\gamma} \text{ for every } 1 \leq k < l \leq t \right\}.
\]

So we have \( \mathcal{I} = \bigcup_{d=1}^{t-1} \mathcal{I}_d \). Let \( \text{Card}(\mathcal{I}_d) \) denote the total number of the vectors \((i_1, \ldots, i_t)\) in \( \mathcal{I}_d \). We can show that \( \text{Card}(\mathcal{I}_d) \leq C p^{d+2\gamma t} \). In fact, the total number of the subsets of \( \{i_1, \ldots, i_t\} \) with cardinality \( d \) is \( C^t_p d \). For a fixed subset \( S \) with cardinality \( d \), the number of \( i \) such that \( |\text{Cov}(Z_i, Z_j)| \geq p^{-\gamma} \) for some \( j \in S \) is no more than \( Cdp^{2\gamma} \). This implies that \( \text{Card}(\mathcal{I}_d) \leq C p^{d+2\gamma t} \). Define \( \mathcal{I}^c = \{ 1 \leq i_1 < \cdots < i_t \leq p \} \setminus \mathcal{I} \). Then the number of elements in the sum \( \sum_{(i_1, \ldots, i_t) \in \mathcal{I}^c} P_{i_1, \cdots, i_t} \) is

\[
C_p^t - O(\sum_{d=1}^{t-1} p^{d+2\gamma t}) = C_p^t - O(p^{t-1+2\gamma t}) = (1 + o(1))C_p^t.
\]
To prove Lemma 1, it suffices to show that

$$P_{i_1, \ldots, i_t} = (1 + o(1)) \pi^{-\frac{t}{2}} p^{-t} \exp(-\frac{tx}{2}) \quad (2.2.4)$$

uniformly in \((i_1, \ldots, i_t) \in \mathcal{I}^c\), and for \(1 \leq d \leq t - 1\),

$$\sum_{(i_1, \ldots, i_t) \in \mathcal{I}_d} P_{i_1, \ldots, i_t} \to 0. \quad (2.2.5)$$

Putting (2.2.3) - (2.2.5) together, we obtain that

$$(1 + o(1)) S_{2k} \leq P\left( \max_{1 \leq i \leq p} |Z_i| \geq \sqrt{xp} \right) \leq (1 + o(1)) S_{2k-1}, \quad (2.2.6)$$

where \(S_k = \sum_{t=1}^{k} (-1)^{t-1/2} \pi^{-\frac{t}{2}} \exp(-\frac{tx}{2})\). Note that

$$\lim_{k \to \infty} S_k = 1 - \exp(-\frac{1}{\sqrt{\pi}} e^{-x/2}).$$

By letting \(p \to \infty\) first and then \(k \to \infty\) in (2.2.6), we prove Lemma 1.

We now prove (2.2.4). Let \(z = (z_{i_1}, \ldots, z_{i_t})'\) and \(|z|_{\min} = \min_{1 \leq j \leq t} |z_{i_j}|\). Write

$$P_{i_1, \ldots, i_t} = \frac{1}{(2\pi)^{t/2} \det(\Omega_t)^{1/2}} \int_{|z|_{\min} \geq \sqrt{xp}} \exp\left(-\frac{1}{2} z' \Omega_t^{-1} z\right) dz,$$

where \(\Omega_t\) is the covariance matrix of \(Z = (Z_{i_1}, \ldots, Z_{i_t})'\), and \(\Omega_t = (a_{kl})_{t \times t}\), where \(a_{kl} = \text{Cov}(Z_{i_k}, Z_{i_l})\). Since \(i_1, \ldots, i_t \in \mathcal{I}^c\), \(a_{kk} = 1\) and \(|a_{kl}| < p^{-\gamma}\) for \(k \neq l\). Write

$$\int_{|z|_{\min} \geq \sqrt{xp}} \exp\left(-\frac{1}{2} z' \Omega_t^{-1} z\right) dz$$

$$= \int_{|z|_{\min} \geq \sqrt{xp}, \|z\|^2 > (\log p)^2} \exp\left(-\frac{1}{2} z' \Omega_t^{-1} z\right) dz$$

$$+ \int_{|z|_{\min} \geq \sqrt{xp}, \|z\|^2 \leq (\log p)^2} \exp\left(-\frac{1}{2} z' \Omega_t^{-1} z\right) dz. \quad (2.2.7)$$

Then

$$\int_{|z|_{\min} \geq \sqrt{xp}, \|z\|^2 > (\log p)^2} \exp\left(-\frac{1}{2} z' \Omega_t^{-1} z\right) dz \leq C \exp(-\log p)^2/2t) \leq C p^{-2t}, \quad (2.2.8)$$
uniformly in \((i_1, \ldots, i_t) \in \mathcal{I}^c\). For the second part of the sum in (2.2.7), note that

\[ \| \Omega_t^{-1} - I \|_2 \leq \| \Omega_t^{-1} \|_2 \| \Omega_t - I \|_2 \leq C p^{-\gamma}. \]  

(2.2.9)

Let \( A = \{ |z|_{\text{min}} \geq \sqrt{x_p}, \| z \|_2 \leq (\log p)^2 \} \). It follows that

\[ \int_A e^{-\frac{1}{2}z' \Omega_t^{-1} z} dz = \int_A e^{-\frac{1}{2}z' (\Omega_t^{-1} - I) z - \frac{1}{2} \| z \|^2} dz \]

\[ = \left( 1 + O(p^{-\gamma}(\log p)^2) \right) \int_A e^{-\frac{1}{2} \| z \|^2} dz \]

\[ = \left( 1 + O(p^{-\gamma}(\log p)^2) \right) \int_{|z|_{\text{min}} \geq \sqrt{x_p}} e^{-\frac{1}{2} \| z \|^2} dz + C p^{-\gamma}. \]

(2.2.10)

uniformly in \((i_1, \ldots, i_t) \in \mathcal{I}^c\). This, together with (2.2.7) and (2.2.8), implies (2.2.4).

It remains to prove (2.2.5). For \( S \subset \mathcal{I}_d \) with \( d \geq 1 \), without loss of generality, we can assume \( S = \{ i_{t-d+1}, \ldots, i_t \} \). By the definition of \( S \) and \( \mathcal{I}_d \), for any \( k \in \{ i_1, \ldots, i_{t-d} \} \), there exists at least one \( l \in S \) such that \( |\text{Cov}(Z_k, Z_l)| \geq p^{-\gamma} \). We divide \( \mathcal{I}_d \) into two parts:

\[ \mathcal{I}_{d,1} = \left\{ 1 \leq i_1 < \cdots < i_t \leq p : \text{there exists an } k \in \{ i_1, \ldots, i_{t-d} \} \text{ such that} \right. \]

\[ \text{for some } l_1, l_2 \in S \text{ with } l_1 \neq l_2, |\text{Cov}(Z_k, Z_{l_1})| \geq p^{-\gamma} \]

\[ \text{and } |\text{Cov}(Z_k, Z_{l_2})| \geq p^{-\gamma} \right\} \]

and \( \mathcal{I}_{d,2} = \mathcal{I}_d \setminus \mathcal{I}_{d,1} \). Clearly, \( \mathcal{I}_{1,1} = \emptyset \) and \( \mathcal{I}_{1,2} = \mathcal{I}_1 \). Moreover, we can show that \( \text{Card}(\mathcal{I}_{d,1}) \leq C p^{d-1+2\gamma}. \) For any \((i_1, \ldots, i_t) \in \mathcal{I}_{d,1}, \)

\[ P\left( |Z_{i_1}| \geq \sqrt{x_p}, \ldots, |Z_{i_t}| \geq \sqrt{x_p} \right) \leq P\left( |Z_{i_{t-d+1}}| \geq \sqrt{x_p}, \ldots, |Z_{i_t}| \geq \sqrt{x_p} \right) = O(p^{-d}). \]

Hence by letting \( \gamma \) be sufficiently small,

\[ \sum_{\mathcal{I}_{d,1}} P_{i_1, \ldots, i_t} \leq C p^{-1+2\gamma} = o(1). \]  

(2.2.11)
For any $(i_1, \ldots, i_t) \in \mathcal{I}_{d,2}$, without loss of generality, assume that $|\text{Cov}(Z_{i_1}, Z_{i_{t-d+1}})| \geq p^{-\gamma}$. Note that

$$P \left( |Z_{i_1}| \geq \sqrt{xp}, \ldots, |Z_{i_t}| \geq \sqrt{xp} \right) \leq P \left( |Z_{i_1}| \geq \sqrt{xp}, |Z_{i_{t-d+1}}| \geq \sqrt{xp}, \ldots, |Z_{i_t}| \geq \sqrt{xp} \right).$$

Let $U_l$ be the covariance matrix of $(Z_{i_1}, Z_{i_{t-d+1}}, \ldots, Z_{i_t})$. We can show that $\|U_l - \bar{U}_l\|_2 = O(p^{-\gamma})$, where $\bar{U}_l = \text{diag}(D, I_{d-1})$ and $D$ is the covariance matrix of $Z_{i_1}$ and $Z_{i_{t-d+1}}$. Using the similar arguments as in (2.2.7)-(2.2.10), we can get

$$P \left( |Z_{i_1}| \geq \sqrt{xp}, |Z_{i_{t-d+1}}| \geq \sqrt{xp}, \ldots, |Z_{i_t}| \geq \sqrt{xp} \right) \leq \frac{1}{1+o(1)}P(|Z_{i_1}| \geq \sqrt{xp}, |Z_{i_{t-d+1}}| \geq \sqrt{xp}) \times O(p^{-d+1}) \leq Cp^{-\frac{2}{1+\gamma}} \times O(p^{-d+1}),$$

where the last inequality follows from Lemma 7 and the assumption $\max_{1 \leq i \neq j \leq p} |\omega_{ij}| \leq r < 1$. Thus by letting $\gamma$ be sufficiently small,

$$\sum_{\mathcal{I}_{d,2}} P_{i_1, \ldots, i_t} \leq Cp^{-\left(\frac{d+2}{1+\gamma} - d + 1 + \frac{1}{1+\gamma}\right)} = o(1).$$

Combining (2.2.11) and (2.2.12), we prove (2.2.5). The proof of Lemma 1 is then complete.

Theorem 1 holds for any fixed sample sizes $n_1$ and $n_2$ and it shows that $M_{\Omega}$, $M_{\Omega^2}$ and $M_I$ have the same asymptotic null distribution. Based on the limiting null distribution, three asymptotically $\alpha$-level tests can be defined as follows:

$$\Phi_{\alpha}(\Omega) = I\{M_{\Omega} \geq 2\log p - \log \log p + q_{\alpha}\},$$
Φ_α(Ω^1) = I\{M_Ω^+ ≥ 2 log p - log log p + q_α\},
Φ_α(I) = I\{M_I ≥ 2 log p - log log p + q_α\},

where q_α is the 1 − α quantile of the type I extreme value distribution with the cumulative distribution function exp \left(-\frac{1}{\sqrt{\pi}} \exp(-x/2)\right), i.e.,

q_α = -\log(\pi) − 2 \log \log(1 − α)^{-1}.

The null hypothesis H_0 is rejected if and only if Φ_α(·) = 1. Although the asymptotic null distribution of the test statistics M_Ω, M_I, and M_Ω^\perp are the same, the power of the tests Φ_α(Ω), Φ_α(Ω^\perp), and Φ_α(I) are quite different. We shall show in Section 1 in Appendix A that the power of Φ_α(Ω) uniformly dominates those of Φ_α(Ω^\perp) and Φ_α(I) when testing against sparse alternatives, and the results are briefly summarized in Section 2.2.2.

2.2.2 The Asymptotic Properties of Φ_α(Ω) And Φ_α(\hat{Ω})

In this section, the asymptotic power of M_Ω is analyzed and the test Φ_α(Ω) is shown to be minimax rate optimal. In practice, Ω is unknown and the test statistic M_Ω should be used instead of M_Ω. Define the set of k_p-sparse vector by

S(k_p) = \{δ : Σ_{j=1}^p I\{δ_j \neq 0\} = k_p\}.

Throughout the section, we analyze the power of M_Ω and M_\hat{Ω} under the alternative

H_1 : δ ∈ S(k_p) with k_p = p^r, 0 ≤ r < 1, and the nonzero
locations are randomly uniformly drawn from \{1, \ldots, p\},

As discussed in the introduction, the condition on the nonzero coordinates in \(H_1\) is mild. Similar conditions have been imposed in Hall and Jin (2008), Hall and Jin (2010) and Arias-Castro, Candes and Plan (2011). We show that, under certain sparsity assumptions on \(\Omega\), \(M_{\Omega}\) performs as well as \(M_{\Omega}\) asymptotically. For the following sections, we assume \(n_1 \asymp n_2\) and write \(n = \frac{n_1 n_2}{n_1 + n_2}\).

**The Asymptotic Power of \(\Phi_\alpha(\Omega)\) And Its Optimality**

The asymptotic power of \(\Phi_\alpha(\Omega)\) is analyzed under certain conditions on the separation between \(\mu_1\) and \(\mu_2\). Furthermore, a lower bound is derived to show that this condition is minimax rate optimal in order to distinguish \(H_1\) and \(H_0\) with probability tending to 1.

**Theorem 2.** Suppose \((C1)\) holds for some \(r < 1/4\). Under the alternative \(H_1\), if

\[
\max_i |\delta_i/\sigma_i^2| \geq \sqrt{2\beta \log p/n}
\]

with \(\beta \geq 1/(\min_i \sigma_i \omega_{ii}) + \varepsilon\) for some constant \(\varepsilon > 0\), then as \(p \to \infty\)

\[
P_{H_1}(\Phi_\alpha(\Omega) = 1) \to 1.
\]

Proof of Theorem 2: It suffices to prove

\[
P\left( \max_{1 \leq i \leq p} \left| (\Omega \delta)_i/\sqrt{\omega_{ii}} \right| \geq \sqrt{(2 + \varepsilon/2) \log p/n} \right) \to 1.
\]

By Lemma 4 and the condition \(\max_i |\delta_i/\sigma_i^2| \geq \sqrt{2\beta \log p/n}\) with \(\beta \geq 1/(\min_i \sigma_i \omega_{ii}) + \varepsilon\) for some constant \(\varepsilon > 0\), we can get \(\max_{1 \leq i \leq p} \left| (\Omega \delta)_i/\sqrt{\omega_{ii}} \right| \geq \sqrt{(2 + \varepsilon/2) \log p/n}\) with probability tending to one. So Theorem 2 follows. \(\square\)
We shall show that the condition $\max_i |\delta_i / \sigma_{ii}^{1/2}| \geq \sqrt{2\beta \log p/n}$ is minimax rate optimal for testing against sparse alternatives. First we introduce some conditions.

(C4) $k_p = p^r$ for some $r < 1/2$ and $\Omega = \Sigma^{-1}$ is $s_p$-sparse with $s_p = O((p/k_p^2)^\gamma)$ for some $0 < \gamma < 1$.

(C4') $k_p = p^r$ for some $r < 1/4$.

(C5) $\|\Omega\|_{L_1} \leq M$ for some constant $M > 0$.

Define the class of $\alpha$-level tests by

$$\mathcal{T}_\alpha = \{ \Phi_\alpha : P_{H_0}(\Phi_\alpha = 1) \leq \alpha \}.$$

The following theorem shows that the condition $\max_i |\delta_i / \sigma_{ii}^{1/2}| \geq \sqrt{2\beta \log p/n}$ is minimax rate optimal.

**Theorem 3.** Assume that (C4) (or (C4')) and (C5) hold. Let $\alpha, \nu > 0$ and $\alpha + \nu < 1$. Then there exists a positive constant $c$ such that for all sufficiently large $n$ and $p$,

$$\inf_{\delta \in \mathcal{S}(k_p) \cap \{ |\delta|_\infty \geq c\sqrt{\log p/n} \}} \sup_{\Phi_\alpha \in \mathcal{T}_\alpha} P(\Phi_\alpha = 1) \leq 1 - \nu.$$

Proof of Theorem 3: First we assume $k_p = o(p^r)$ for some $r < 1/4$, and we can get similar argument if $k_p = O(p^r)$ for some $r < 1/2$ and $\Omega = \Sigma^{-1}$ is $s_p$ sparse with $s_p = O((p/k_p^2)^\gamma)$ for some $0 < \gamma < 1$. Let $\mathcal{M}_{s,p}$ denote the set of all subsets of $\{1, ..., p\}$ with cardinality $k_p$. Let $\hat{m}$ be a random set of $\{1, ..., p\}$, which is uniformly distributed on $\mathcal{M}$. Let $\omega_j$, $1 \leq j \leq p$ be i.i.d. variables with $P(\omega_j = 1) = P(\omega_j = -1) = 1/2$. We construct a class of $\delta = \mu_1 - \mu_2$ by letting $\mu_1 = 0$ and $\delta = -\mu_2$ satisfy $\delta = (\delta_1, \ldots, \delta_p)'$ with $\delta_j = \frac{\rho}{\sqrt{k_p}} \omega_j 1_{j \in \hat{m}}$, where $\rho = c \sqrt{\frac{k_p \log p}{n}}$ and $c > 0$.
is sufficiently small that will be specified later. Clearly, $|\delta|_2 = \rho$. Let $\mu_\rho$ be the distribution of $\delta$. Note that $\mu_\rho$ is a probability measure on $\{\delta \in S_{k_p} : |\delta| = \rho\}$.

We now calculate the likelihood ratio $L_{\mu_\rho} = \frac{d\mu_\rho}{d\rho_0}(\{X_n, Y_n\})$. It is easy to see that $L_{\mu_\rho} = E_{\hat{m}, \omega}(\exp(-\sqrt{n}Z'\delta - \frac{n}{2}\delta'\Omega\delta))$, where $Z$ is a multivariate normal vector with mean 0 and $\text{Cov}(Z) = \Omega$, and is independent with $\hat{m}$ and $\Omega$. For any fixed $\hat{m} = m$, let $\delta_{m}^i$, $1 \leq i \leq 2^{k_p}$ be all the possible values of $\delta$. That is, $P(\delta = \delta_{m}^i|\hat{m} = m) = 2^{-k_p}$. Thus

$$E_{\hat{m}, \omega}(\exp(-\sqrt{n}Z'\delta - \frac{n}{2}\delta'\Omega\delta))$$

$$= \frac{1}{(\frac{p}{k_p})^2 \frac{1}{2^{k_p}}} \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} \exp \left( -\sqrt{n}Z'\delta_{(m)}^i - \frac{n}{2}\delta_{(m)}^i\Omega\delta_{(m)}^i \right).$$

It follows that

$$EL_{\mu_\rho}^2 = E \left\{ \frac{1}{(\frac{p}{k_p})^2 \frac{1}{2^{k_p}}} \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} \exp \left( -\sqrt{n}Z'\delta_{(m)}^i - \frac{n}{2}\delta_{(m)}^i\Omega\delta_{(m)}^i \right) \right\}^2$$

$$= \frac{1}{(\frac{p}{k_p})^2 \frac{1}{2^{k_p}}} \sum_{m,m' \in \mathcal{M}} \sum_{i,j=1}^{2^{k_p}} \exp \left( -\sqrt{n}Z'\delta_{(m)}^i + \delta_{(m')}^j \right)$$

$$\times \exp \left( -\frac{n}{2}(\delta_{(m)}^i\Omega\delta_{(m)}^i + \delta_{(m')}^j\Omega\delta_{(m')}^j) \right)$$

$$= \frac{1}{(\frac{p}{k_p})^2 \frac{1}{2^{k_p}}} \sum_{m,m' \in \mathcal{M}} \sum_{i,j=1}^{2^{k_p}} \exp \left( \frac{n}{2}(\delta_{(m)}^i\Omega\delta_{(m)}^i + \delta_{(m')}^j\Omega\delta_{(m')}^j) \right)$$

$$= \frac{1}{(\frac{p}{k_p})^2 \frac{1}{2^{k_p}}} \sum_{m,m' \in \mathcal{M}} \sum_{i,j=1}^{2^{k_p}} \exp \left( \frac{n\rho^2}{k_p} \sum_{k \in m,l \in m'} a_{k,l}\omega_{k}^i \omega_{l}^j \right).$$
where \( \frac{\alpha}{\sqrt{k_p}}(\omega^{(i)}_k I_{k \in m}) := \delta^{(i)}_{(m)} \) and \( \Omega = (a_{kl})_{p \times p} \). Thus

\[
\text{EL}^2_{\mu_p} = \frac{1}{(p_{k_p})^2} \sum_{m, m' \in M} 2^{k_p} \prod_{k,l=1}^{k_p} \left( \exp\left(\frac{n \rho^2}{k_p} a_{kl}\right) + \exp\left(-\frac{n \rho^2}{k_p} a_{kl}\right) \right) \]

\[
= \frac{1}{(p_{k_p})^2} \sum_{m, m' \in M} \prod_{k \in m, l \in m'} \cosh\left(\frac{n \rho^2}{k_p} a_{kl}\right) \]

\[
\leq \frac{1}{(p_{k_p})^2} \sum_{m, m' \in M} \prod_{k \in m, l \in m'} \exp\left(\frac{n \rho^2}{k_p} |a_{kl}|\right).
\]

For every \( m \), let

\[
B := B_m = \{ l : |a_{k,l}| \geq \frac{M}{d}, k \in m \},
\]

where \( d = \left(\frac{p}{k_p}\right)^{1-\gamma} \) and \( \gamma \) is sufficiently small. For every \( k \), the number of \( l \) such that \( |a_{kl}| \geq \frac{M}{d} \) is at most \( d \).

Hence

\[
\text{EL}^2_{\mu_p} \leq \frac{1}{(p_{k_p})^2} \sum_{m \in M} \sum_{j=0}^{k_p} I\{|m' \cap B| = j\} \exp\left(\sum_{k \in m, l \in m'} \frac{n \rho^2}{k_p} |a_{kl}|\right)
\]

\[
= \frac{1}{(p_{k_p})^2} \sum_{m \in M} \sum_{j=0}^{k_p} I\{|m' \cap B| = j\} \cdot \exp\left(\sum_{k \in m, l \in m'} \frac{n \rho^2}{k_p} |a_{kl}|\right) + \sum_{k \in m, l \in m'} \frac{n \rho^2}{k_p} |a_{kl}|
\]

\[
\leq \frac{1}{(p_{k_p})^2} \sum_{m \in M} \sum_{j=0}^{k_p} \left(\frac{k_p d}{j}\right) \left(\frac{p - k_p}{k_p - j}\right) \exp\left(\frac{M n \rho^2}{k_p} j + \frac{M k_p^2 \log p}{d}\right)
\]

\[
\leq \left(1 + o(1)\right) \left(\frac{dk_p t}{p}\right)^{k_p} \exp\left(\frac{M k_p^2 \log p}{d}\right),
\]

where \( t = \exp\left(\frac{M n \rho^2}{k_p}\right) = p^{M c^2} \). It follows that

\[
\text{EL}^2_{\mu_p} \leq (1 + o(1)) \exp\left(k_p \log\left(1 + \frac{dk_p t}{p}\right) + \frac{M k_p^2 \log p}{d}\right)
\]

\[
\leq (1 + o(1)) \exp\left(\frac{dk_p t}{p} + \frac{M k_p^2 \log p}{d}\right) \leq 1 + 4(1 - \alpha - \nu)^2
\]
by letting \( c \) be sufficiently small. If \( k_p = O(p^r) \) for some \( r < 1/2 \) and \( \Omega = \Sigma^{-1} \) is \( s_p \) sparse with \( s_p = O((p/k^2_p)^\gamma) \) for some \( 0 < \gamma < 1 \), we let \( B := B_m = \{ l : a_{k,l} \neq 0, k \in m \} \). Then we can similarly get

\[
E \left[ \frac{\mu^2}{\rho} \right] \leq \frac{1}{(k_p p)^2} \sum_{m \in M} \sum_{j=0}^{k_p} I\{|m' \cap B| = j\} \exp \left( \sum_{k \in m, l \in m'} \frac{n \rho^2}{k_p} |a_{k,l}| \right)
\]

\[
\leq \frac{1}{(k_p p)^2} \sum_{m \in M} \sum_{j=0}^{k_p} I\{|m' \cap B| = j\} \exp \left( \frac{Mn \rho^2}{k_p} j \right)
\]

\[
\leq (1 + o(1)) \left( 1 + \frac{s_p k_p t}{p} \right)^{k_p},
\]

So we can still get \( E \left[ \frac{\mu^2}{\rho} \right] \leq 1 + 4(1 - \alpha - \nu)^2 \) by letting \( c \) be sufficiently small.

Theorem 3 now follows from Lemma 6. \( \square \)

Theorem 3 shows that, if \( c \) is sufficiently small, then any \( \alpha \) level test is unable to reject the null hypothesis correctly uniformly over \( \delta \in S(k_p) \cap \{|\delta|_{\infty} \geq c \sqrt{\log p/n}\} \) with probability tending to one. So the order of the lower bound \( \max_i |\delta_i / \sigma_{ii}^{1/2}| \geq \sqrt{2\beta \log p/n} \) can not be improved.

**The Asymptotic Properties of \( \Phi_\alpha(\hat{\Omega}) \) And Its Optimality**

We now analyze the properties of \( M_{\hat{\Omega}} \) and the corresponding test including the limiting null distribution and the asymptotic power. We shall show that \( M_{\hat{\Omega}} \) has the same limiting null distribution as \( M_{\Omega} \) and define the corresponding test \( \Phi_\alpha(\hat{\Omega}) \) by

\[
\Phi_\alpha(\hat{\Omega}) = I\{M_{\hat{\Omega}} \geq 2 \log p - \log \log p + q_\alpha\}.
\]
Under certain sparsity assumptions, the asymptotic properties of $\Phi_\alpha(\hat{\Omega})$ are similar to those of $\Phi_\alpha(\Omega)$. Define the following class of matrices that belong to an $\ell_q$ ball with $0 \leq q < 1$:

$$
\mathcal{U}_q(s_p, M_p) = \left\{ \Omega \succ 0 : \|\Omega\|_{L_1} \leq M_p, \max_{1 \leq j \leq p} \sum_{i=1}^p |\omega_{ij}|^q \leq s_p \right\}.
$$

We assume that $\Omega \in \mathcal{U}_q(s_p, M_p)$ so $\Omega$ can be well estimated by the CLIME estimator $\hat{\Omega}$ under some conditions on $s_p$ and $M_p$; see Cai, Liu and Luo (2011).

**Theorem 4.** Suppose that (C1) and (C3) hold and $\Omega \in \mathcal{U}_q(s_p, M_p)$ with

$$
s_p = o\left(\frac{n(1-q)^2}{M_p(1-q)(\log p)^{(3-q)/2}}\right),
$$

(2.2.13)

(i). Then under the null hypothesis $H_0$, for any $x \in \mathbb{R}$,

$$
P_{H_0}(\hat{\Omega} - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right), \quad \text{as } n, p \to \infty.
$$

(ii). Under the alternative hypothesis $H_1$ with $r < 1/6$, we have, as $n, p \to \infty$,

$$
\frac{P_{H_1}(\Phi_\alpha(\hat{\Omega}) = 1)}{P_{H_1}(\Phi_\alpha(\Omega) = 1)} \to 1.
$$

Furthermore, if $\max_i |\delta_i/\sigma_i^2| \geq \sqrt{2/\beta \log p/n}$ with $\beta \geq 1/(\min_i \sigma_i^2 \omega_{ii}) + \varepsilon$ for some constant $\varepsilon > 0$, then

$$
P_{H_1}(\Phi_\alpha(\hat{\Omega}) = 1) \to 1, \quad \text{as } n, p \to \infty.
$$

Proof of Theorem 4: We only prove part (ii) of Theorem 4 in this section, part (i) follows from the proof of part (ii) directly. Without loss of generality, we assume
that $\sigma_{ii} = 1$ for $1 \leq i \leq p$. Define the event $A = \{\max_{1 \leq i \leq p} |\delta_i| \leq 8\sqrt{\log p/n}\}$. We have $P\left(\|\hat{\Sigma}_X - \Sigma\|_\infty \leq C\sqrt{\log p/n}\right) \to 1$ as $n, p \to \infty$; see Cai and Liu (2011). On the event $\{\|\hat{\Sigma}_X - \Sigma\|_\infty \leq C\sqrt{\log p/n}\}$,

$$\max_{1 \leq i \leq p} |\hat{\omega}_{ii}^{(1)} - \omega_{ii}| \leq |\hat{\Omega}\hat{\Sigma}_X \hat{\Omega} - \Omega \Sigma \Omega|_\infty \leq Cs_p M_p^{2-q} \left(\frac{\log p}{n}\right)^{(1-q)/2} = o(1/ \log p).$$

Similarly, $\max_{1 \leq i \leq p} |\hat{\omega}_{ii}^{(2)} - \omega_{ii}| = o(1/ \log p)$. Hence, as in the proof of Proposition 1 (i) in Appendix A, it is easy to show that $P\left(M_{\Omega} \in R_\alpha, A^c\right) = P\left(A^c\right) + o(1)$ and $P\left(M_{\Omega} \in R_\alpha, A^c\right) = P\left(A^c\right) + o(1)$. Note that $\hat{\Omega}(\bar{X} - \bar{Y}) = (\hat{\Omega} - \Omega)(\bar{X} - \bar{Y} - \delta) + (\hat{\Omega} - \Omega)\delta + \Omega(\bar{X} - \bar{Y})$. On $A$, we have

$$\left|\hat{\Omega} - \Omega\right|(\bar{X} - \bar{Y} - \delta) + (\hat{\Omega} - \Omega)\delta\right|_\infty = O_p\left(s_p M_p^{1-q} \left(\frac{\log p}{n}\right)^{(1-q)/2}\right)$$

$$= o_p\left(1/ \sqrt{n \log p}\right).$$

To prove Theorem 4, it suffices to show that

$$P\left(\max_{1 \leq i \leq p} |Z_i^0| \geq \sqrt{x_p} + a_n, A\right) = P\left(\max_{1 \leq i \leq p} |Z_i^0| \geq \sqrt{x_p}, A\right) + o(1), \quad (2.2.14)$$

for any $a_n = o((\log p)^{-1/2})$, where $Z_i^0 = (\Omega Z)_i/\sqrt{\omega_{ii}}$ defined in the proof of Proposition 1 (i) in Appendix A. From the proof of Proposition 1 (i), let $H = \text{supp}(\delta) = \{l_1, ..., l_p\}$, then we can get

$$P\left(\max_{1 \leq i \leq p} |Z_i^0| \geq \sqrt{x_p} + a_n, A\right)$$

$$= \alpha P(A) + (1 - \alpha)P\left(\max_{i \in H} |Y_i| \geq \sqrt{x_p} + a_n, A\right) + o(1),$$

$$P\left(\max_{1 \leq i \leq p} |Z_i^0| \geq \sqrt{x_p}, A\right) = \alpha P(A) + (1 - \alpha)P\left(\max_{i \in H} |Y_i| \geq \sqrt{x_p}, A\right) + o(1),$$

where given $\delta, Y_i, i \in H$ are independent normal random variables with unit variance. This, together with Lemma 5, implies (2.2.14).
By Theorem 4, we see that $M_{\hat{\Omega}}$ and $M_{\Omega}$ have the same asymptotic distribution and power, and so the test $\Phi_\alpha(\hat{\Omega})$ is also minimax rate optimal.

Remark 1. The CLIME estimator in Cai, Liu and Luo (2011) is considered in this section. Other “good” estimators of the precision matrix can also be used. In general, Theorem 4 still holds if $\log p = o(n)$ and the estimator $\hat{\Omega}$ satisfies the following conditions:

\[
\|\hat{\Omega} - \Omega\|_{L_1} = o_p\left(\frac{1}{\log p}\right) \quad \text{and} \quad \max_{1 \leq i \leq p} |\hat{\omega}_{ii}^{(0)} - \omega_{ii}| = o_p\left(\frac{1}{\log p}\right),
\]

where $\hat{\omega}_{ii}^{(0)} = \frac{n_1}{n_1 + n_2} \hat{\omega}_{ii}^{(1)} + \frac{n_2}{n_1 + n_2} \hat{\omega}_{ii}^{(2)}$ and $\hat{\omega}_{ii}^{(j)}$ is defined in (2.1.6) for $i = 1, \ldots, p$ and $j = 1, 2$.

**Power Comparison of the Oracle Tests**

The tests $\Phi_\alpha(\Omega)$ and $\Phi_\alpha(\hat{\Omega})$ are shown in Sections 2.2.2 to be minimax rate optimal for testing against sparse alternatives. Under some additional regularity conditions, it can be shown that the test $\Phi_\alpha(\Omega)$ is uniformly at least as powerful as both $\Phi_\alpha(\Omega^{\frac{1}{2}})$ and $\Phi_\alpha(I)$, and the results are stated in Proposition 1 in Appendix A. Furthermore, we show that, under some special alternatives, the test $\Phi_\alpha(\Omega)$ is strictly more powerful than both $\Phi_\alpha(\Omega^{\frac{1}{2}})$ and $\Phi_\alpha(I)$. For further details, see Propositions 6 and 7 in Appendix A.
2.3 Extension to Non-Gaussian Distributions

We have so far focused on the Gaussian setting and studied the asymptotic null distributions and power of the tests. In this section, the results for the tests $\Phi_{\alpha}(\Omega)$ and $\Phi_{\alpha}(\hat{\Omega})$ are extended to non-Gaussian distributions.

We require some moment conditions on the distributions of $X$ and $Y$. Let $X$ and $Y$ be two $p$-dimensional random vectors satisfying

$$X = \mu_1 + U_1 \text{ and } Y = \mu_2 + U_2,$$

where $U_1$ and $U_2$ are independent and identical distributed random vectors with mean zero and covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$. Let $V_j = \Omega U_j =: (V_{1j}, \ldots, V_{pj})^T$ for $j = 1, 2$. The moment conditions are divided into two cases: the sub-Gaussian-type tails and polynomial-type tails.

(C6). (Sub-Gaussian-type tails) Suppose that $\log p = o(n^{1/4})$. There exist some constants $\eta > 0$ and $K > 0$ such that

$$E \exp(\eta V_{i1}^2/\omega_{ii}) \leq K \text{ and } E \exp(\eta V_{i2}^2/\omega_{ii}) \leq K \text{ for } 1 \leq i \leq p.$$

(C7). (Polynomial-type tails) Suppose that for some constants $\gamma_0, c_1 > 0$, $p \leq c_1 n^{\gamma_0}$, and for some constants $\epsilon > 0$ and $K > 0$

$$E|V_{i1}/\omega_{ii}^{\frac{1}{2}}|^{2\gamma_0+2+\epsilon} \leq K \text{ and } E|V_{i2}/\omega_{ii}^{\frac{1}{2}}|^{2\gamma_0+2+\epsilon} \leq K \text{ for } 1 \leq i \leq p.$$

Theorem 5. Suppose that (C1), (C3) and (C6) (or (C7)) hold. Then under the null
hypothesis \( H_0 \), for any \( x \in \mathbb{R} \),

\[
P_{H_0}(M_\Omega - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right), \quad \text{as } n, p \to \infty.
\]

Proof of Theorem 5: Let \((V_1, ..., V_p)'\) be a zero mean random vector with covariance matrix \( \Omega = (\omega_{ij}) \) and the diagonal \( \omega_{ii} = 1 \) for \( 1 \leq i \leq p \) satisfying moment conditions \((C6)\) or \((C7)\). Let \( \hat{V}_l = V_l I\{|V_l| \leq \tau_n\} \) for \( l = 1, ..., n \), where \( \tau_n = \eta^{-1/2} \sqrt{\log(p + n)} \) if \((C6)\) holds and \( \tau_n = \sqrt{n/(\log p)^8} \) if \((C7)\) holds. Let \( W_i = \sum_{l=1}^n V_{li} / \sqrt{n} \) and \( \hat{W}_i = \sum_{l=1}^n \hat{V}_{li} / \sqrt{n} \). Then

\[
P(\max_{1 \leq i \leq p} |W_i - \hat{W}_i| \geq \frac{1}{\log p}) \leq P(\max_{1 \leq i \leq p} |V_i| \geq \tau_n) \leq np \max_{1 \leq i \leq p} P(|V_i| \geq \tau_n) = O(p^{-1} + n^{-\epsilon/8}). \tag{2.3.1}
\]

Note that

\[
\max_{1 \leq i \leq p} W_i^2 - \max_{1 \leq i \leq p} \hat{W}_i^2 \leq 2 \max_{1 \leq i \leq p} |W_i| \max_{1 \leq i \leq p} |W_i - \hat{W}_i| + \max_{1 \leq i \leq p} |W_i - \hat{W}_i|^2. \tag{2.3.2}
\]

By (2.3.1) and (2.3.2), it is enough to prove that for any \( x \in \mathbb{R} \), as \( p \to \infty \)

\[
P(\max_{1 \leq i \leq p} \hat{W}_i^2 - 2 \log p + \log \log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right).
\]

It follows from Lemma 2 that for any fixed \( k \leq \lfloor p/2 \rfloor \),

\[
\sum_{t=1}^{2k} (-1)^{t-1} \sum_{1 \leq i_1 < ... < i_t \leq p} P(|\hat{W}_{i_1}| \geq x_p, ..., |W_{i_t}| \geq x_p) \leq P(\max_{1 \leq i \leq p} |\hat{W}_i| \geq x_p) \leq \sum_{t=1}^{2k} (-1)^{t-1} \sum_{1 \leq i_1 < ... < i_t \leq p} P(|\hat{W}_{i_1}| \geq x_p, ..., |W_{i_t}| \geq x_p). \tag{2.3.3}
\]

Define \(|\hat{W}|_{\text{min}} = \min_{1 \leq i \leq p} |\hat{W}_i|\). Then by Theorem 1 in Zaitsev (1987), we have

\[
P(|\hat{W}|_{\text{min}} \geq x_p)
\]
\[ P(|Z|_{\min} \geq x_p - \epsilon_n (\log p)^{-1/2}) + c_1 d^{5/2} \exp \left( - \frac{n^{1/2} \epsilon_n}{c_2 d^{3/2} (\log p)^{1/2}} \right), \quad (2.3.4) \]

where \( c_1 > 0 \) and \( c_2 > 0 \) are absolute constants, \( \epsilon_n \to 0 \) which will be specified later and \( Z = (Z_{t_1}, ..., Z_{t_t})' \) is a \( t \) dimensional normal vector as defined in Theorem 1. Because \( \log p = o(n^{1/4}) \), we can let \( \epsilon \to 0 \) sufficiently slow such that

\[ c_1 d^{5/2} \exp \left( - \frac{n^{1/2} \epsilon_n}{c_2 d^{3/2} (\log p)^{1/2}} \right) = O(p^{-M}) \quad (2.3.5) \]

for any large \( M > 0 \). It follows from (2.3.3), (2.3.4) and (2.3.5) that

\[ P(\max_{1 \leq i \leq p} |\hat{W}_i| \geq x_p) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} \sum_{1 \leq i_1 < \ldots < i_t \leq p} P(|Z|_{\min} \geq x_p - \epsilon_n (\log p)^{-1/2}) + o(1). \quad (2.3.6) \]

Similarly, using Theorem 1 in Zaïtsev (1987) again, we can get

\[ P(\max_{1 \leq i \leq p} |\hat{W}_i| \geq x_p) \geq \sum_{t=1}^{2k} (-1)^{t-1} \sum_{1 \leq i_1 < \ldots < i_t \leq p} P(|Z|_{\min} \geq x_p - \epsilon_n (\log p)^{-1/2}) - o(1). \quad (2.3.7) \]

So by (2.3.6), (2.3.7) and the proof of Theorem 1, the theorem is proved. \( \square \)

Theorem 5 shows that \( \Phi_\alpha(\Omega) \) is still an asymptotically \( \alpha \)-level test when the distribution is non-Gaussian. When \( \Omega \) is unknown, the CLIME estimator \( \hat{\Omega} \) in Cai, Liu and Luo (2011) satisfies

\[ \|\hat{\Omega} - \Omega\|_{L_1} = O_p \left( s_p M_p^{1-q} \left( \frac{\log p}{n} \right)^{(1-q)/2} \right) \quad (2.3.8) \]

under certain moment conditions on \( U_{i1} \) and \( U_{i2} \) for \( 1 \leq i \leq p \). The following theorem shows that the test \( \Phi_\alpha(\hat{\Omega}) \) shares the same optimality as \( \Phi_\alpha(\Omega) \) in the non-Gaussian setting.
Theorem 6. Suppose that (C1), (C3), (C6) (or (C7)) and (2.3.8) hold and \( \Omega \in \mathcal{U}_q(s_p) \) with

\[ s_p = o\left(\frac{n^{(1-q)/2}}{M_p^{1-q} (\log p)^{(3-q)/2}}\right), \]

(i). Under the null hypothesis \( H_0 \), for any \( x \in \mathbb{R} \),

\[ \mathbb{P}_{H_0}\left(M_{\hat{\Omega}} - 2 \log p + \log \log p \leq x\right) \to \exp\left(-\frac{1}{\sqrt{\pi}} \exp\left(-\frac{x}{2}\right)\right), \text{ as } n, p \to \infty. \]

(ii). Under \( H_1 \) and the conditions of Theorem 2, we have

\[ \mathbb{P}_{H_1}\left(\Phi_\alpha(\hat{\Omega}) = 1\right) \to 1, \text{ as } n, p \to \infty. \]

Proof of Theorem 6: (i). It can be shown that \( \hat{\Omega}(\bar{X} - \bar{Y}) = \Omega(\bar{X} - \bar{Y}) + o_p(1/\sqrt{n \log p}) \), as in the proof of Theorem 4. Theorem 6 (i) then follows from Theorem 5.

(ii). Note that \( \hat{\Omega}(\bar{X} - \bar{Y}) = (\hat{\Omega} - \Omega)(\bar{X} - \bar{Y} - \delta) + (\hat{\Omega} - \Omega)\delta + \Omega(\bar{X} - \bar{Y}) \). It suffices to prove \( \mathbb{P}\left(\max_{1 \leq i \leq p} \left| \frac{\sqrt{n}(\hat{\Omega}\delta)_i}{\sqrt{\omega_{ii}}} + \sqrt{n}\Omega(\bar{X} - \bar{Y} - \delta)_i/\sqrt{\omega_{ii}} \right| \geq \sqrt{\rho \log p}\right) \to 1 \) for some \( \rho > 2 \). To this end, we only need to show \( \mathbb{P}\left(\max_{1 \leq i \leq p} \left| (\hat{\Omega}\delta)_i/\sqrt{\omega_{ii}} \right| \geq \sqrt{(2 + \varepsilon/4) \log p/n}\right) \to 1 \). Note that

\[
\max_{1 \leq i \leq p} \left| (\hat{\Omega}\delta)_i/\sqrt{\omega_{ii}} \right| \\
\geq \max_{1 \leq i \leq p} \left| (\Omega\delta)_i/\sqrt{\omega_{ii}} \right| + o_p(1) \max_{1 \leq i \leq p} \left| \delta_i \right| \geq (1 + o_p(1)) \max_{1 \leq i \leq p} \left| (\Omega\delta)_i/\sqrt{\omega_{ii}} \right|.
\]

By the condition \( \max_i |\delta_i/\sigma_{ii}^{1/2}| \geq \sqrt{2\beta \log p/n} \) with \( \beta \geq 1/(\min_i \sigma_{ii} \omega_{ii}) + \varepsilon \) for some constant \( \varepsilon > 0 \), we can get \( \max_{1 \leq i \leq p} \left| (\Omega\delta)_i/\sqrt{\omega_{ii}} \right| \geq \sqrt{(2 + \varepsilon/2) \log p/n} \) with probability tending to one. This proves (ii). \( \square \)
2.4 Simulation Study

The tests $\Phi_\alpha(\Omega)$ and $\Phi_\alpha(\hat{\Omega})$ are easy to implement. In this section, we consider the numerical performance of the tests $\Phi_\alpha(\Omega)$ and $\Phi_\alpha(\hat{\Omega})$ and compare these tests with a number of other tests, including the three oracle tests $\Phi_\alpha(\Omega)$, $\Phi_\alpha(\Omega^{1/2})$ and $\Phi_\alpha(I)$, the tests based on the sum of squares type statistics in Bai and Saranadasa (1996), Srivastava and Du (2009), Chen and Qin (2010), and the commonly used Hotelling’s $T^2$ test. These last four tests are denoted respectively by BS, SD, CQ and $T^2$ respectively in the tables below.

In the simulations, we shall always take $\mu_2 = 0$. Under the null hypothesis, $\mu_1 = \mu_2 = 0$, while under the alternative hypothesis, we take $\mu_1 = (\mu_{11}, \ldots, \mu_{1p})'$ to have $m$ nonzero entries with the support $S = \{l_1, \ldots, l_m : l_1 < l_2 < \cdots < l_m\}$ uniformly and randomly drawn from $\{1, \ldots, p\}$. Two values of $m$ are considered: $m = \lfloor 0.05p \rfloor$ and $m = \lfloor \sqrt{p} \rfloor$. Here $\lfloor x \rfloor$ denote the largest integer that is no greater than $x$. For each of these two values of $m$, and for any $l_j \in S$, two settings of the magnitude of $\mu_{1l_j}$ are considered: $\mu_{1l_j} = \pm \sqrt{\log p/n}$ with equal probability and $\mu_{1l_j}$ has magnitude randomly uniformly drawn from the interval $[-\sqrt{8 \log p/n}, \sqrt{8 \log p/n}]$. We take $\mu_{1k} = 0$ for $k \in S^c$.

Three different scenarios about the precision matrix $\Omega$ are considered in the simulation: $\Omega$ is known, $\Omega$ is sparse and the case when the covariance matrix $\Sigma$ is sparse. In the case when $\Omega$ is known, we compare the oracle performance of three maximum-type test statistics with the sum of squares type statistics. When $\Omega$ is
unknown, we use the CLIME estimator in Cai, Liu and Luo (2011) to estimate it when $\Omega$ is sparse, while the inverse of the adaptive thresholding estimator in Cai and Liu (2011) is used to estimate it when $\Sigma$ is sparse.

Let $D = (d_{ij})$ be a diagonal matrix with diagonal elements $d_{ii} = \text{Unif}(1, 3)$ for $i = 1, \ldots, p$. Denote by $\lambda_{\text{min}}(A)$ the minimum eigenvalue of a symmetric matrix $A$.

In the case when the precision matrix $\Omega$ is known, the following two models for $\Sigma$ are considered:

- Model 1: $\Sigma^* = (\sigma^*_{ij})$ where $\sigma^*_{ii} = 1$, $\sigma^*_{ij} = 0.5$ for $i \neq j$. $\Sigma = D^{1/2} \Sigma^* D^{1/2}$.

- Model 2: $\Sigma^* = (\sigma^*_{ij})$ where $\sigma^*_{ii} = 1$, $\sigma^*_{ij} = \text{Unif}(0, 1)$ for $i < j$ and $\sigma^*_{ji} = \sigma^*_{ij}$. $\Sigma = D^{1/2}(\Sigma^* + \delta I)/(1 + \delta)D^{1/2}$ with $\delta = |\lambda_{\text{min}}(\Sigma^*)| + 0.05$.

In the case when the precision matrix $\Omega$ is sparse, we consider the following three models:

- Model 3: $\Sigma = (\sigma_{ij})$ where $\sigma_{ii} = 1$, $\sigma_{ij} = 0.8$ for $2(k - 1) + 1 \leq i \neq j \leq 2k$, where $k = 1, \ldots, [p/2]$ and $\sigma_{ij} = 0$ otherwise.

- Model 4: $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = 0.6|\ell_{i-j}|$ for $1 \leq i, j \leq p$.

- Model 5: $\Omega = (\omega_{ij})$ where $\omega_{ii} = 2$ for $i = 1, \ldots, p$, $\omega_{i+1} = 0.8$ for $i = 1, \ldots, p-1$, $\omega_{i+2} = 0.4$ for $i = 1, \ldots, p-2$, $\omega_{i+3} = 0.4$ for $i = 1, \ldots, p-3$, $\omega_{i+4} = 0.2$ for $i = 1, \ldots, p-4$, $\omega_{ij} = \omega_{ji}$ for $i, j = 1, \ldots, p$ and $\omega_{ij} = 0$ otherwise.

The following three models are considered in the last scenario when the covariance matrix $\Sigma$ is sparse:
• Model 6: $\Sigma^* = (\sigma^*_{ij})$ where $\sigma^*_{ii} = 1$, $\sigma^*_{ij} = 0.8$ for $2(k - 1) + 1 \leq i \neq j \leq 2k$, where $k = 1, \ldots, [p/2]$ and $\sigma^*_{ij} = 0$ otherwise. $\Sigma = D^{1/2}\Sigma^*D^{1/2}$.

• Model 7: $\Omega = (\omega_{ij})$ where $\omega_{ij} = 0.6^{|i-j|}$ for $1 \leq i, j \leq p$. $\Sigma = D^{1/2}\Omega^{-1}D^{1/2}$.

• Model 8: $\Omega^{1/2} = (a_{ij})$ where $a_{ii} = 1$, $a_{ij} = 0.8$ for $2(k - 1) + 1 \leq i \neq j \leq 2k$, where $k = 1, \ldots, [p/2]$ and $a_{ij} = 0$ otherwise. $\Omega = D^{1/2}\Omega^{1/2}\Omega^{1/2}D^{1/2}$ and $\Sigma = \Omega^{-1}$.

Under each model, two independent random samples $\{X_k\}$ and $\{Y_l\}$ are generated with the same sample size $n = 100$ from two multivariate normal distributions with the means $\mu_1$ and $\mu_2$ respectively and a common covariance matrix $\Sigma$. The dimension $p$ takes values $p = 50, 100$ and $200$. The power and significance level are calculated from 1000 replications. The numerical results are summarized in Tables 2.1-2.3.

It can be seen from Table 2.1 that the estimated sizes are reasonably close to the nominal level 0.05 for all the tests. Tables 2.2 and 2.3 show that the oracle test $\Phi_\alpha(\Omega)$ has the highest power in all eight models over all dimensions ranging from 50 to 200, and the performance of the test $\Phi_\alpha(\hat{\Omega})$ based on either the CLIME estimator or the inverse of the adaptive thresholding estimator is close to that of the oracle test $\Phi_\alpha(\Omega)$ in Models 3 - 8. The tests based on the sum of squares test statistics and the test $\Phi_\alpha(I)$ are not powerful against the sparse alternatives considered in the all models. The powers of these tests are significantly lower than that of $\Phi_\alpha(\Omega)$. The test $\Phi_\alpha(\Omega)$ also outperforms $\Phi_\alpha(\hat{\Omega}^2)$. In summary, the tests
\( \Phi_\alpha(\Omega) \) and \( \Phi_\alpha(\hat{\Omega}) \) perform well against the sparse alternatives and are significantly more powerful in comparison to the other tests in the simulation study.

We also carried out simulations for Gamma distribution. Similar phenomena as those in the Gaussian case are observed. For reasons of space, these simulation results are given in Appendix A.

### 2.5 Technical Lemmas

The proofs of some of the main theorems rely on a few additional technical lemmas. These technical results are collected in this section and they are proved in Appendix A.

**Lemma 2 (Bonferroni Inequality).** Let \( A = \cup_{t=1}^p A_t \). For any \( k < [p/2] \), we have

\[
\sum_{t=1}^{2k} (-1)^{t-1} E_t \leq P(A) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} E_t,
\]

where \( E_t = \sum_{1 \leq i_1 < \cdots < i_t \leq p} P(A_{i_1} \cap \cdots \cap A_{i_t}) \).

**Lemma 3 (Berman, 1962).** If \( X \) and \( Y \) have a bivariate normal distribution with expectation zero, unit variance and correlation coefficient \( \rho \), then

\[
\lim_{c \to \infty} \frac{P(X > c, Y > c)}{[2\pi(1 - \rho)^{1/2} c^2]^{-1} \exp \left( -\frac{c^2}{2(1+\rho)} \right) \left( 1 + \rho \right)^{1/2}} = 1,
\]

uniformly for all \( \rho \) such that \( |\rho| \leq \delta \), for any \( \delta, 0 < \delta < 1 \).
<table>
<thead>
<tr>
<th>$p$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T^2$</td>
<td>0.04 0.07</td>
<td>–</td>
<td>0.06 0.05</td>
<td>–</td>
<td>0.04 0.06</td>
<td>–</td>
<td>0.04 0.04</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BS</td>
<td>0.07 0.07 0.06</td>
<td>0.06 0.07 0.05</td>
<td>0.06 0.07 0.06</td>
<td>0.07 0.06 0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.07 0.07 0.06</td>
<td>0.06 0.07 0.06</td>
<td>0.06 0.07 0.06</td>
<td>0.07 0.06 0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQ</td>
<td>0.07 0.07 0.06</td>
<td>0.06 0.07 0.06</td>
<td>0.06 0.07 0.06</td>
<td>0.06 0.06 0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(I)$</td>
<td>0.03 0.02 0.02</td>
<td>0.03 0.04 0.04</td>
<td>0.06 0.05 0.04</td>
<td>0.03 0.04 0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\Omega^T)$</td>
<td>0.03 0.04 0.04</td>
<td>0.03 0.04 0.04</td>
<td>0.05 0.05 0.03</td>
<td>0.03 0.04 0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\Omega)$</td>
<td>0.04 0.04 0.04</td>
<td>0.03 0.03 0.05</td>
<td>0.05 0.04 0.03</td>
<td>0.03 0.04 0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\hat{\Omega})$</td>
<td>– – –</td>
<td>– – –</td>
<td>0.05 0.05 0.06</td>
<td>0.04 0.05 0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $T^2$ | 0.04 0.05 | – | 0.05 0.06 | – | 0.05 0.06 | – | 0.05 0.05 | – |
| BS   | 0.06 0.05 0.04 | 0.06 0.07 0.06 | 0.07 0.07 0.06 | 0.06 0.06 0.05 |
| SD   | 0.06 0.05 0.04 | 0.06 0.06 0.06 | 0.06 0.06 0.06 | 0.06 0.06 0.06 |
| CQ   | 0.06 0.05 0.02 | 0.06 0.06 0.06 | 0.07 0.07 0.06 | 0.06 0.06 0.05 |
| $\Phi_\alpha(I)$ | 0.03 0.03 0.02 | 0.03 0.03 0.04 | 0.03 0.04 0.04 | 0.03 0.03 0.03 |
| $\Phi_\alpha(\Omega^T)$ | 0.03 0.03 0.03 | 0.04 0.03 0.04 | 0.04 0.04 0.04 | 0.04 0.04 0.04 |
| $\Phi_\alpha(\Omega)$ | 0.03 0.03 0.03 | 0.03 0.04 0.04 | 0.03 0.03 0.04 | 0.03 0.03 0.03 |
| $\Phi_\alpha(\hat{\Omega})$ | 0.05 0.06 0.06 | 0.04 0.04 0.05 | 0.04 0.05 0.06 | 0.03 0.04 0.03 |

Table 2.1: Empirical sizes with $\alpha = 0.05$ and $n = 100$. Based on 1000 replications.
<table>
<thead>
<tr>
<th>$p$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T^2$</td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
<td>Model 5</td>
<td>Model 6</td>
<td>Model 7</td>
<td>Model 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T^2$</td>
<td>0.11</td>
<td>0.17</td>
<td>-</td>
<td>0.36</td>
<td>0.19</td>
<td>-</td>
<td>0.22</td>
<td>0.41</td>
<td>-</td>
<td>0.17</td>
<td>0.27</td>
<td>-</td>
<td>0.14</td>
<td>0.26</td>
<td>-</td>
<td>0.10</td>
<td>0.17</td>
<td>-</td>
<td>0.14</td>
<td>0.21</td>
<td>-</td>
</tr>
<tr>
<td>BS</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.10</td>
<td>0.11</td>
<td>0.11</td>
<td>0.19</td>
<td>0.27</td>
<td>0.11</td>
<td>0.13</td>
<td>0.24</td>
<td>0.12</td>
<td>0.28</td>
<td>0.44</td>
<td>0.08</td>
<td>0.11</td>
<td>0.14</td>
<td>0.10</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>SD</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.11</td>
<td>0.12</td>
<td>0.11</td>
<td>0.19</td>
<td>0.27</td>
<td>0.11</td>
<td>0.13</td>
<td>0.24</td>
<td>0.12</td>
<td>0.27</td>
<td>0.44</td>
<td>0.08</td>
<td>0.11</td>
<td>0.16</td>
<td>0.10</td>
<td>0.15</td>
<td>0.27</td>
</tr>
<tr>
<td>CQ</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.09</td>
<td>0.11</td>
<td>0.11</td>
<td>0.19</td>
<td>0.27</td>
<td>0.11</td>
<td>0.13</td>
<td>0.24</td>
<td>0.12</td>
<td>0.27</td>
<td>0.44</td>
<td>0.08</td>
<td>0.11</td>
<td>0.14</td>
<td>0.10</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>$\Phi_n(I)$</td>
<td>0.06</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0.08</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.21</td>
<td>0.09</td>
<td>0.12</td>
<td>0.21</td>
<td>0.11</td>
<td>0.21</td>
<td>0.32</td>
<td>0.05</td>
<td>0.06</td>
<td>0.10</td>
<td>0.07</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>$\Phi_n(\Omega^+)$</td>
<td>0.12</td>
<td>0.17</td>
<td>0.33</td>
<td>0.24</td>
<td>0.18</td>
<td>0.16</td>
<td>0.24</td>
<td>0.49</td>
<td>0.69</td>
<td>0.19</td>
<td>0.33</td>
<td>0.53</td>
<td>0.16</td>
<td>0.33</td>
<td>0.57</td>
<td>0.08</td>
<td>0.14</td>
<td>0.31</td>
<td>0.12</td>
<td>0.22</td>
<td>0.45</td>
</tr>
<tr>
<td>$\Phi_n(\Omega)$</td>
<td>0.12</td>
<td>0.18</td>
<td>0.33</td>
<td>0.87</td>
<td>0.42</td>
<td>0.23</td>
<td>0.34</td>
<td>0.66</td>
<td>0.86</td>
<td>0.24</td>
<td>0.40</td>
<td>0.61</td>
<td>0.19</td>
<td>0.44</td>
<td>0.68</td>
<td>0.11</td>
<td>0.21</td>
<td>0.44</td>
<td>0.17</td>
<td>0.28</td>
<td>0.63</td>
</tr>
<tr>
<td>$\Phi_n(\Omega^+)$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.36</td>
<td>0.70</td>
<td>0.89</td>
<td>0.26</td>
<td>0.41</td>
<td>0.66</td>
<td>0.21</td>
<td>0.40</td>
<td>0.62</td>
<td>0.12</td>
<td>0.23</td>
<td>0.47</td>
<td>0.18</td>
<td>0.28</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 2.2: Powers of tests with $\alpha = 0.05$ and $n = 100$. Signal strength keeps the same. Based on 1000 replications.
Table 2.3: Powers of tests with $\alpha = 0.05$ and $n = 100$. Signal Strength varies. Based on 1000 replications.
Lemma 4. Suppose (C1) holds and \( \Sigma \) has all diagonal elements equal to 1. Then for \( p^r \)-sparse \( \delta \), with \( r < 1/4 \) and nonzero locations \( l_1, ..., l_m, m = p^r \), randomly and uniformly drawn from \( \{1, ..., p\} \), we have, for any \( 2r < a < 1 - 2r \), as \( p \to \infty \),

\[
P\left( \max_{i \in H} \left| \frac{(\Omega \delta)_i}{\sqrt{\omega_{ii}}} - \sqrt{\omega_{ii}} \delta_i \right| = O(p^{r-a/2}) \max_{i \in H} |\delta_i| \right) \to 1, \tag{2.5.1} \]

and

\[
P\left( \max_{i \in H} \left| (\Omega^{1/2} \delta)_i - a_{ii} \delta_i \right| = O(p^{r-a/2}) \max_{i \in H} |\delta_i| \right) \to 1, \tag{2.5.2} \]

where \( \Omega^{1/2} = (a_{ij}) \) and \( H \) is the support of \( \delta \).

Lemma 5. Let \( Y_i \sim N(\mu_i, 1) \) be independent for \( i = 1, ..., n \). Let \( a_n = o((\log n)^{-1/2}) \). Then

\[
\sup_{x \in \mathbb{R}} \max_{1 \leq k \leq n} \left| P\left( \max_{1 \leq i \leq k} Y_i \geq x + a_n \right) - P\left( \max_{1 \leq i \leq k} Y_i \geq x \right) \right| = o(1) \tag{2.5.3} \]

uniformly in the means \( \mu_i, 1 \leq i \leq n \). If \( Y_i \) is replaced by \( |Y_i| \), then (2.5.3) still holds.

Lemma 6 (Baraud, 2002). Let \( \mathcal{F} \) be some subset of \( l_2(J) \). Let \( \mu_\rho \) be some probability measure on \( \mathcal{F}_\rho = \{ \theta \in \mathcal{F}, \|\theta\| \geq \rho \} \) and let \( P_{\mu_\rho} = \int P_\theta d\mu_\rho(\theta) \). Assuming that \( P_{\mu_\rho} \) is absolutely continuous with respect to \( P_0 \), we define \( L_{\mu_\rho}(y) = \frac{dP_{\mu_\rho}}{dP_0}(y) \). For all \( \alpha > 0 \), \( \nu \in [0, 1 - \alpha] \), if \( E_{\theta}(L^{2}_{\mu_\rho}(Y)) \leq 1 + 4(1 - \alpha - \nu)^2 \), then

\[
\forall \rho \leq \rho^*, \quad \inf_{\theta \in \mathcal{F}_\rho} \sup_{\Phi_\alpha(\Phi_\alpha = 0)} P_{\theta}(\Phi_\alpha) \geq \nu.
\]
Chapter 3

Testing High Dimensional Covariance Matrices

3.1 The testing procedure

Given two i.i.d. random samples \( \{ X_1, \ldots, X_{n_1} \} \) from a \( p \)-variate distribution with the covariance matrix \( \Sigma_1 = (\sigma_{ij1})_{p \times p} \) and \( \{ Y_1, \ldots, Y_{n_2} \} \) from a distribution with the covariance matrix \( \Sigma_2 = (\sigma_{ij2})_{p \times p} \), define the sample covariance matrices

\[
(\hat{\sigma}_{ij1})_{p \times p} := \hat{\Sigma}_1 = \frac{1}{n_1} \sum_{k=1}^{n_1} (X_k - \bar{X})(X_k - \bar{X})',
\]

\[
(\hat{\sigma}_{ij2})_{p \times p} := \hat{\Sigma}_2 = \frac{1}{n_2} \sum_{k=1}^{n_2} (Y_k - \bar{Y})(Y_k - \bar{Y})',
\]

where \( \bar{X} = \frac{1}{n_1} \sum_{k=1}^{n_1} X_k \) and \( \bar{Y} = \frac{1}{n_2} \sum_{k=1}^{n_2} Y_k \). The null hypothesis \( H_0 : \Sigma_1 = \Sigma_2 \) is equivalent to \( H_0 : \max_{1 \leq i \leq j \leq p} |\sigma_{ij1} - \sigma_{ij2}| = 0 \). A natural approach to testing this hypothesis is to compare the sample covariances \( \hat{\sigma}_{ij1} \) and \( \hat{\sigma}_{ij2} \) and to base the test
on the maximum differences. It is important to note that the sample covariances \( \hat{\sigma}_{ij1}'s \) and \( \hat{\sigma}_{ij2}'s \) are in general heteroscedastic and can possibly have a wide range of variability. It is thus necessary to first standardize \( \hat{\sigma}_{ij1} - \hat{\sigma}_{ij2} \) before making a comparison among different entries.

To be more specific, define the variances \( \theta_{ij1} = \text{Var}((X_i - \mu_{1i})(X_j - \mu_{1j})) \) and \( \theta_{ij2} = \text{Var}((Y_i - \mu_{2i})(Y_j - \mu_{2j})) \). Given the two samples \( \{X_1, \ldots, X_{n_1}\} \) and \( \{Y_1, \ldots, Y_{n_2}\} \), \( \theta_{ij1} \) and \( \theta_{ij2} \) can be respectively estimated by

\[
\hat{\theta}_{ij1} = \frac{1}{n_1} \sum_{k=1}^{n_1} [(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) - \hat{\sigma}_{ij1}]^2, \quad \bar{X}_i = \frac{1}{n_1} \sum_{k=1}^{n_1} X_{ki},
\]

and

\[
\hat{\theta}_{ij2} = \frac{1}{n_2} \sum_{k=1}^{n_2} [(Y_{ki} - \bar{Y}_i)(Y_{kj} - \bar{Y}_j) - \hat{\sigma}_{ij2}]^2, \quad \bar{Y}_i = \frac{1}{n_2} \sum_{k=1}^{n_2} Y_{ki}.
\]

Such an estimator of the variance have been used in Cai and Liu (2011) in the context of adaptive estimation of a sparse covariance matrix. Given \( \hat{\theta}_{ij1} \) and \( \hat{\theta}_{ij2} \), the variance of \( \hat{\sigma}_{ij1} - \hat{\sigma}_{ij2} \) can then be estimated by \( \hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2 \).

Define the standardized statistics

\[
M_{ij} := \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}, \quad 1 \leq i \leq j \leq p. \tag{3.1.1}
\]

We consider the following test statistic for testing the hypothesis \( H_0 : \Sigma_1 = \Sigma_2 \),

\[
M_n := \max_{1 \leq i \leq j \leq p} M_{ij} = \max_{1 \leq i \leq j \leq p} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}. \tag{3.1.2}
\]

The asymptotic behavior of the test statistic \( M_n \) will be studied in detail in Section 2.2. Intuitively, \( M_{ij} \) are approximately square of standard normal variables.
under the null $H_0$ and they are only “weakly dependent” under suitable conditions. The test statistic $M_n$ is the maximum of $p(p+1)/2$ such variables and so the value of $M_n$ is close to $2 \log p^2$ under $H_0$, based on the extreme values of normal random variables. More precisely, we show in Section 2.2 that, under certain regularity conditions, $M_n - 4 \log p + \log \log p$ converges to a type I extreme value distribution under the null hypothesis $H_0$. Based on this result, for a given significance level $0 < \alpha < 1$, we define the test $\Phi_\alpha$ by

$$\Phi_\alpha = I(M_n \geq q_\alpha + 4 \log p - \log \log p)$$

(3.1.3)

where $q_\alpha$ is the $1 - \alpha$ quantile of the type I extreme value distribution with the cumulative distribution function $\exp(-\frac{1}{\sqrt{8\pi}} \exp(-\frac{x^2}{2}))$, i.e.,

$$q_\alpha = - \log(8\pi) - 2 \log \log(1 - \alpha)^{-1}. \quad (3.1.4)$$

The hypothesis $H_0 : \Sigma_1 = \Sigma_2$ is rejected whenever $\Phi_\alpha = 1$

The test $\Phi_\alpha$ is particularly well suited for testing against sparse alternatives. It is consistent if one of the entries of $\Sigma_1 - \Sigma_2$ has a magnitude more than $C\sqrt{\log p/n}$ for some constant $C > 0$ and no other special structure of $\Sigma_1 - \Sigma_2$ is required. It will be shown that the test $\Phi_\alpha$ is an asymptotically $\alpha$-level test and enjoys certain optimality against sparse alternatives.

The standardized statistics $M_{ij}$ are also useful for identifying the support of $\Sigma_1 - \Sigma_2$. That is, they can be used to estimate the positions at which the two covariance matrices differ from each other. This is of interest in many applications
including gene selection. We show in Section 3.3.1 that under certain conditions, the support of $\Sigma_1 - \Sigma_2$ can be correctly recovered by thresholding the $M_{ij}$.

### 3.2 Theoretical analysis of size and power

We now turn to an analysis of the properties of the test $\Phi_\alpha$ including the asymptotic size and power. The asymptotic size of the test is obtained by deriving the limiting distribution of the test statistic $M_n$ under the null. We are particularly interested in the power of the test $\Phi_\alpha$ under the sparse alternatives where $\Sigma_1 - \Sigma_2$ is sparse in the sense that $\Sigma_1 - \Sigma_2$ only contains a small number of nonzero entries.

#### 3.2.1 Definitions and assumptions

We begin by introducing some basic definitions and technical assumptions. Throughout this chapter, denote $|a|_2 = \sqrt{\sum_{j=1}^p a_j^2}$ for the usual Euclidean norm of a vector $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p \times q}$, define the spectral norm $\|A\|_2 = \sup_{|x|_2 \leq 1} |Ax|_2$ and the Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant $C$ such that $|a_n| \leq C|b_n|$ holds for all sufficiently large $n$, write $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n / b_n = 0$, and write $a_n \asymp b_n$ if there exist constants $C > c > 0$ such that $c|b_n| \leq |a_n| \leq C|b_n|$ for all sufficiently large $n$. The two sample sizes are assumed to be comparable, i.e., $n_1 \asymp n_2$. Let $n = \max(n_1, n_2)$.

Let $R_1 =: (\rho_{ij1})$ and $R_2 =: (\rho_{ij2})$ be the correlation matrices of $X$ and $Y$.
respectively. For a fixed constant $\alpha_0 > 0$, define

$$s_j = s_j(\alpha_0) := \text{card}\{i : |\rho_{ij1}| \geq (\log p)^{-1-\alpha_0} \text{ or } |\rho_{ij2}| \geq (\log p)^{-1-\alpha_0}\}.$$ 

So $s_j$ is the cardinality of the set of indices $i$ such that either the $i$th variable of $X$ is highly correlated ($> (\log p)^{-1-\alpha_0}$) with the $j$th variable of $X$ or the $i$th variable of $Y$ is highly correlated with the $j$th variable of $Y$. For $0 < r < 1$, define the set

$$\Lambda(r) = \{1 \leq i \leq p : |\rho_{ij1}| > r \text{ or } |\rho_{ij2}| > r \text{ for some } j \neq i\}.$$ 

So $\Lambda(r)$ is the set of indices $i$ such that either the $i$th variable of $X$ is highly correlated ($> r$) with some other variable of $X$ or the $i$th variable of $Y$ is highly correlated ($> r$) with some other variable of $Y$.

To obtain the asymptotic distribution of $M_n$, we assume that $\max_{1 \leq j \leq p} s_j$ is not too large, and the set $\Lambda(r)$ is also not too large for some $r < 1$.

(C1). Suppose that there exists a subset $\Upsilon \subset \{1, 2, \ldots, p\}$ with $\text{Card}(\Upsilon) = o(p)$ and a constant $\alpha_0 > 0$ such that for all $\gamma > 0$, $\max_{1 \leq j \leq p, j \notin \Upsilon} s_j(\alpha_0) = o(p^\gamma)$. Moreover, there exist some constant $r < 1$ and a sequence of numbers $\Lambda_{p,r}$ such that $\text{Card}(\Lambda(r)) \leq \Lambda_{p,r} = o(p)$.

It is easy to check that if the eigenvalues of the correlation matrices $R_1$ and $R_2$ are bounded from above, then Condition (C1) is easily satisfied. In fact, if $\lambda_{\max}(R_1) \leq C_0$ and $\lambda_{\max}(R_2) \leq C_0$ for some constant $C_0 > 0$, then $\max_j s_j \leq C(\log p)^{2+2\alpha_0}$. The condition $\max_{j,j \notin \Upsilon} s_j = o(p^\gamma)$ is quite weak. It allows the largest eigenvalues to be of order $p/(\log p)^{2+2\alpha_0}$. It is also easy to see that the condition
$\text{Card}(\Lambda(r)) = o(p)$ for some $r < 1$ is trivially satisfied if all the correlations are bounded away from $\pm 1$, i.e.,

$$\max_{1 \leq i < j \leq p} |\rho_{ij}| \leq r < 1 \quad \text{and} \quad \max_{1 \leq i < j \leq p} |\rho_{ij}| \leq r < 1. \quad (3.2.1)$$

We do not require the distributions of $X$ and $Y$ to be Gaussian. Instead we shall impose more general moment conditions.

**Sub-Gaussian type tails:** Suppose that $\log p = o(n^{1/5})$ and $n_1 \asymp n_2$. There exist some constants $\eta > 0$ and $K > 0$ such that

$$E \exp(\eta (X_i - \mu_{i1})^2/\sigma_{i1}^2) \leq K, \quad E \exp(\eta (Y_i - \mu_{i2})^2/\sigma_{i2}^2) \leq K \quad \text{for all } i.$$  

Furthermore, we assume that for some constants $\tau_1 > 0$ and $\tau_2 > 0$,

$$\min_{1 \leq i < j \leq p} \frac{\theta_{ij}}{\sigma_{i1}\sigma_{j1}} \geq \tau_1 \quad \text{and} \quad \min_{1 \leq i < j \leq p} \frac{\theta_{ij}}{\sigma_{i2}\sigma_{j2}} \geq \tau_2. \quad (3.2.2)$$

**Polynomial-type tails:** Suppose that for some $\gamma_0, c_0 > 0$, $p \leq c_1 n^{\gamma_0}, n_1 \asymp n_2$ and for some $\epsilon > 0$

$$E|(X_i - \mu_{i1})/\sigma_{i1}^{1/2}|^{4\gamma_0+4+\epsilon} \leq K, \quad E|(Y_i - \mu_{i2})/\sigma_{i2}^{1/2}|^{4\gamma_0+4+\epsilon} \leq K \quad \text{for all } i.$$  

Furthermore, we assume (3.2.2) holds.

In addition to the moment conditions, we also need the following technical condition for some of the results. It holds if $X$ and $Y$ have elliptically contoured distributions, see, e.g., Anderson (2003, pp. 47-54).

**Suppose that there exist $\kappa_1, \kappa_2 \geq \frac{1}{3}$ such that for any $i, j, k, l \in \{1, 2, \ldots, p\}$,**

$$E(X_i - \mu_{i1})(X_j - \mu_{j1})(X_k - \mu_{k1})(X_l - \mu_{l1}) = \kappa_1(\sigma_{ij1}\sigma_{kl1} + \sigma_{ik1}\sigma_{jl1} + \sigma_{il1}\sigma_{jk1}),$$

$$E|X_i - \mu_{i1}|^2 = \kappa_2(\sigma_{i1}^2 + \sigma_{i2}^2 + \sigma_{i3}^2) \quad \text{for all } i.$$  

56
\[
E(Y_i - \mu_i_2)(Y_j - \mu_j_2)(Y_k - \mu_k_2)(Y_l - \mu_l_2) = \kappa_2(\sigma_{ij_2}\sigma_{kl_2} + \sigma_{ik_2}\sigma_{jl_2} + \sigma_{il_2}\sigma_{jk_2}).
\]

Remark 2. Conditions (C1) and (C3) are only needed for the limiting distribution of the test statistic. (C3) holds for the elliptically contoured distributions (Anderson (2003)) with \(\kappa_1 \equiv \frac{1}{3}E(X_i - \mu_1)^4/[E(X_i - \mu_1)^2]^2\) and \(\kappa_2 \equiv \frac{1}{3}E(Y_i - \mu_2)^4/[E(Y_i - \mu_2)^2]^2\). Results on the power of the test \(\Phi_\alpha\) does not rely on either (C1) or (C3). Conditions (C2) and (C2*) are moment conditions on \(X\) and \(Y\). We only require either (C2) or (C2*) to hold. This is much weaker than the Gaussian assumption required in the literature such as Schott (2007) and Srivastava and Yanagihara (2010). Condition (3.2.2) is satisfied with \(\tau_1 = \tau_2 = 1\) if \(X \sim N(\mu_1, \Sigma_1)\) and \(Y \sim N(\mu_2, \Sigma_2)\). In the non-Gaussian case, if (C3) and (3.2.1) hold, then (3.2.2) holds with \(\tau_1 = \kappa_1 - r^2/3\) and \(\tau_2 = \kappa_2 - r^2/3\).

3.2.2 Limiting null distribution and optimality

We are now ready to present the asymptotic null distribution of \(M_n\). The following theorem shows that \(M_n - 4 \log p + \log \log p\) converges weakly under \(H_0\) to an extreme value distribution of type I with the distribution function \(F(t) = \exp(-\frac{1}{\sqrt{8\pi}}e^{-t/2}).\)

Theorem 7. Suppose that (C1), (C2) (or (C2*)) and (C3) hold. Then under \(H_0\), for any \(t \in \mathbb{R}\),

\[
P\left(M_n - 4 \log p + \log \log p \leq t \right) \to \exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{t}{2}\right)\right), \quad (3.2.3)
\]

as \(n, p \to \infty\). Furthermore, under \(H_0\), the convergence in (3.2.3) is uniform for all \(X\) and \(Y\) satisfying (C1), (C2) (or (C2*)) and (C3).
Proof of Theorem 7: Without loss of generality, we assume that $\bm{\mu}_1 = 0$, $\bm{\mu}_2 = 0$, $\sigma_{ii1} = \sigma_{ii2} = 1$ for $1 \leq i \leq p$. Let
\[
\hat{M}_n = \max_{1 \leq i \leq j \leq p} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}, \quad \text{and} \quad \tilde{M}_n = \max_{1 \leq i \leq j \leq p} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}.
\]
Note that under the event $\{ |\hat{\theta}_{ij1}/\theta_{ij1} - 1| \leq C \varepsilon_n / \log p, |\hat{\theta}_{ij2}/\theta_{ij2} - 1| \leq C \varepsilon_n / \log p \}$,
\[
\left| M_n - \hat{M}_n \right| \leq C \frac{M_n \varepsilon_n}{\log p},
\]
\[
\left| \tilde{M}_n - M_n \right| \leq C(n_1 + n_2)(\max_{1 \leq i \leq p} \tilde{X}_i^4 + \max_{1 \leq i \leq p} \tilde{Y}_i^4)
\]
\[
+ C(n_1 + n_2)^{1/2} \tilde{M}_n^{1/2}(\max_{1 \leq i \leq p} \tilde{X}_i^2 + \max_{1 \leq i \leq p} \tilde{Y}_i^2).
\]
By the exponential inequality, $\max_{1 \leq i \leq p} |\tilde{X}_i| + \max_{1 \leq i \leq p} |\tilde{Y}_i| = O_p(\sqrt{\log p/n})$. Thus, by Lemma 10, it suffices to show that for any $x \in \mathbb{R}$,
\[
P\left( \tilde{M}_n - 4 \log p + \log \log p \leq x \right) \rightarrow \exp \left( -\frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right) \quad (3.2.4)
\]
as $n, p \rightarrow \infty$. Let $\rho_{ij} = \rho_{ij1} = \rho_{ij2}$ under $H_0$. Define
\[
A_j = \left\{ i : |\rho_{ij}| \geq (\log p)^{-1-\alpha_0} \right\},
\]
\[
A = \{(i, j) : 1 \leq i \leq j \leq p \},
\]
\[
A_0 = \{(i, j) : 1 \leq i \leq p, i \leq j, j \in A_1 \},
\]
\[
B_0 = \{(i, j) : i \in A(r) \cup \mathcal{Y}, i < j \leq p \} \cup \{(i, j) : j \in A(r) \cup \mathcal{Y}, 1 \leq i < j \},
\]
\[
D_0 = A_0 \cup B_0.
\]
By the definition of $D_0$, for any $(i_1, j_1) \in A \setminus D_0$ and $(i_2, j_2) \in A \setminus D_0$, we have $|\rho_{kl}| \leq r$ for any $k \neq l \in \{i_1, j_1, i_2, j_2\}$. Set $y_p = x + 4 \log p - \log \log p$,
\[
\tilde{M}_{A \setminus D_0} = \max_{(i, j) \in A \setminus D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}, \quad \text{and} \quad \tilde{M}_{D_0} = \max_{(i, j) \in D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}.
\]
Then
\[
\left| \Pr(\tilde{M}_n \geq y_p) - \Pr(\tilde{M}_{A \setminus D_0} \geq y_p) \right| \leq \Pr(\tilde{M}_{D_0} \geq y_p).
\]

Note that \(\text{Card}(\Lambda(r)) + \text{Card}(Y) = o(p)\) and for all \(\gamma > 0\), \(\max_{j \not\in T} s_j = o(p^\gamma)\). This implies that \(\text{Card}(D_0) \leq C_0 p^{1+\gamma} + o(p^2)\) for any \(\gamma > 0\). By Lemma 11, we obtain that for any fixed \(x \in R\),
\[
\Pr(\tilde{M}_{D_0} \geq y_p) \leq \text{Card}(D_0) \times C_p^{-2} + o(1) = o(1).
\]

Set
\[
\tilde{M}_{A \setminus D_0} = \max_{(i,j) \in A \setminus D_0} \frac{(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} =: \max_{(i,j) \in A \setminus D_0} U_{ij}^2,
\]
where \(U_{ij} = \frac{\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}}\). It suffices to show that for any \(x \in \mathbb{R}\),
\[
\Pr(\tilde{M}_{A \setminus D_0} - 4 \log p + \log \log p \leq x) \to \exp \left( -\frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right)
\]
as \(n, p \to \infty\). We arrange the two dimensional indices \(\{(i, j) : (i, j) \in A \setminus D_0\}\) in any ordering and set them as \(\{(i_m, j_m) : 1 \leq m \leq q\}\) with \(q = \text{Card}(A \setminus D_0)\). Let \(n_2/n_1 \leq K_1 \) with \(K_1 \geq 1\), \(\theta_{k1} = \theta_{ikx1}, \theta_{k2} = \theta_{ikx2}\) and define
\[
\begin{align*}
Z_{lk} &= \frac{n_2}{n_1}(X_{li_k} X_{lj_k} - \sigma_{ikj1}), & 1 \leq l \leq n_1, \\
\hat{Z}_{lk} &= -(Y_{li_k} Y_{lj_k} - \sigma_{ikj2}), & n_1 + 1 \leq l \leq n_1 + n_2, \\
V_k &= \frac{1}{\sqrt{n_2^2 \theta_{k1}/n_1 + n_2 \theta_{k2}}} \sum_{l=1}^{n_1 + n_2} Z_{lk}, \\
\hat{V}_k &= \frac{1}{\sqrt{n_2^2 \theta_{k1}/n_1 + n_2 \theta_{k2}}} \sum_{l=1}^{n_1 + n_2} \hat{Z}_{lk},
\end{align*}
\]
where \(\hat{Z}_{lk} = Z_{lk} I\{Z_{lk} \leq \tau_n\} - E Z_{lk} I\{|Z_{lk}| \leq \tau_n\}\), and \(\tau_n = \eta^{-1} 8 K_1 \log(p+n)\) if (C2) holds, \(\tau_n = \sqrt{n}/(\log p)^6\) if (C2*) holds. Note that \(\max_{(i,j) \in A \setminus D_0} U_{ij}^2 = \max_{1 \leq k \leq q} V_k^2\).
We have, if (C2) holds, then
\[\max_{1 \leq k \leq q} \frac{1}{\sqrt{n}} \sum_{l=1}^{n_1+n_2} E|Z_{lk}| I\{ |Z_{lk}| \geq \eta^{-1}8K_1 \log(p+n) \} \]
\[\leq C \sqrt{n} \max_{1 \leq l \leq n_1+n_2} \max_{1 \leq k \leq q} E|Z_{lk}| I\{ |Z_{lk}| \geq \eta^{-1}8K_1 \log(p+n) \} \]
\[\leq C \sqrt{n}(p+n)^{-4} \max_{1 \leq l \leq n_1+n_2} \max_{1 \leq k \leq q} E|Z_{lk}| \exp(\eta |Z_{lk}|/(2K_1)) \]
\[\leq C \sqrt{n}(p+n)^{-4},\]
and if (C2\(^*\)) holds, then
\[\max_{1 \leq k \leq q} \frac{1}{\sqrt{n}} \sum_{l=1}^{n_1+n_2} E|Z_{lk}| I\{ |Z_{lk}| \geq \sqrt{n}/(\log p)^8 \} \]
\[\leq C \sqrt{n} \max_{1 \leq l \leq n_1+n_2} \max_{1 \leq k \leq q} E|Z_{lk}| I\{ |Z_{lk}| \geq \sqrt{n}/(\log p)^8 \} \]
\[\leq C n^{-\gamma_0-\epsilon/8}.\]

Hence we have
\[P\left( \max_{1 \leq k \leq q} |V_k - \hat{V}_k| \geq (\log p)^{-1} \right) \leq P\left( \max_{1 \leq k \leq q} \max_{1 \leq l \leq n_1+n_2} |Z_{lk}| \geq \tau_n \right) \]
\[\leq np \max_{1 \leq j \leq p} \left[ P\left( X_j^2 \geq \tau_n/2 \right) + P\left( Y_j^2 \geq \tau_n/2 \right) \right] \]
\[= O(p^{-1} + n^{-\epsilon/8}). \quad (3.2.5)\]

Note that
\[\left| \max_{1 \leq k \leq q} V_k^2 - \max_{1 \leq k \leq q} \hat{V}_k^2 \right| \leq 2 \max_{1 \leq k \leq q} |\hat{V}_k| \max_{1 \leq k \leq q} |V_k - \hat{V}_k| + \max_{1 \leq k \leq q} |V_k - \hat{V}_k|^2. \quad (3.2.6)\]

By (3.2.5) and (3.2.6), it suffices to prove that for any \( x \in \mathbb{R} \), as \( n, p \to \infty \),
\[P\left( \max_{1 \leq k \leq q} \hat{V}_k^2 - 4 \log p + \log \log p \leq x \right) \to \exp \left( -\frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right). \quad (3.2.7)\]
By Bonferroni inequality (see Lemma 8), for any integer \( m \) with \( 0 < m < q/2 \),

\[
\sum_{d=1}^{2m} (-1)^{d-1} \sum_{1 \leq k_1 < \ldots < k_d \leq q} \mathbb{P}\left( \bigcap_{j=1}^{d} E_{k_j} \right) \leq \mathbb{P}\left( \max_{1 \leq k \leq q} \hat{V}_k^2 \geq y_p \right)
\]

\[
\leq \sum_{d=1}^{2m-1} (-1)^{d-1} \sum_{1 \leq k_1 < \ldots < k_d \leq q} \mathbb{P}\left( \bigcap_{j=1}^{d} E_{k_j} \right) \quad (3.2.8)
\]

where \( E_{k_j} = \{ \hat{V}_{k_j}^2 \geq y_p \} \). Let \( \tilde{Z}_{lk} = \tilde{Z}_{lk}/(n_2\theta_{k_1}/n_1 + \theta_{k_2})^{1/2} \) for \( 1 \leq k \leq q \) and \( W_l = (\tilde{Z}_{l,k_1}, \ldots, \tilde{Z}_{l,k_d}) \), for \( 1 \leq l \leq n_1 + n_2 \). Define \( |a|_{\min} = \min_{1 \leq i \leq d} |a_i| \) for any vector \( a \in \mathbb{R}^d \). Then we have

\[
\mathbb{P}\left( \bigcap_{j=1}^{d} E_{k_j} \right) = \mathbb{P}\left( \left| n_2^{-1/2} \sum_{l=1}^{n_1+n_2} W_l \right|_{\min} \geq y_n^{1/2} \right).
\]

By Theorem 1 in Zaitsev (1987), we have

\[
\mathbb{P}\left( \left| n_2^{-1/2} \sum_{l=1}^{n_1+n_2} W_l \right|_{\min} \geq y_n^{1/2} \right) \leq \mathbb{P}\left( |N_d|_{\min} \geq y_p^{1/2} - \epsilon_n (\log p)^{-1/2} \right)
\]

\[
+ c_1 d^{5/2} \exp \left( - \frac{n_2^{1/2} \epsilon_n}{c_2 d^3 T_n (\log p)^{1/2}} \right), \quad (3.2.9)
\]

where \( c_1 > 0 \) and \( c_2 > 0 \) are absolutely constants, \( \epsilon_n \to 0 \) which will be specified later and \( N_d =: (N_{k_1}, \ldots, N_{k_d}) \) is a normal vector with \( \mathbb{E} N_d = 0 \) and \( \text{Cov}(N_d) = \frac{n_1}{n_2} \text{Cov}(W_1) + \text{Cov}(W_{n_1+1}) \). Recall that \( d \) is a fixed integer not depending on \( n, p \).

Because \( \log p = o(n_1^{1/5}) \), we can let \( \epsilon_n \to 0 \) sufficiently slow such that

\[
c_1 d^{5/2} \exp \left( - \frac{n_2^{1/2} \epsilon_n}{c_2 d^3 T_n (\log p)^{1/2}} \right) = O(p^{-M}) \quad (3.2.10)
\]

for any large \( M > 0 \). It follows from (3.2.8), (3.2.9) and (3.2.10) that

\[
\mathbb{P}\left( \max_{1 \leq k \leq q} \hat{V}_k^2 \geq y_p \right)
\]
\[ \leq \sum_{d=1}^{2m-1} (-1)^{d-1} \sum_{1 \leq k_1 < \ldots < k_d \leq q} \mathbb{P} \left( |N_d|_{\min} \geq y_p^{1/2} - \epsilon_n (\log p)^{-1/2} \right) + o(1). \quad (3.2.11) \]

Similarly, using Theorem 1 in Zaitsev (1987) again, we can get
\[ \mathbb{P} \left( \max_{1 \leq k \leq q} \hat{V}_k^2 \geq y_p \right) \geq \sum_{d=1}^{2m} (-1)^{d-1} \sum_{1 \leq k_1 < \ldots < k_d \leq q} \mathbb{P} \left( |N_d|_{\min} \geq y_p^{1/2} + \epsilon_n (\log p)^{-1/2} \right) - o(1). \quad (3.2.12) \]

We need the following technical lemma to prove the result.

Lemma 7. For any fixed integer \( d \geq 1 \) and real number \( x \in \mathbb{R} \),
\[ \sum_{1 \leq k_1 < \ldots < k_d \leq q} \mathbb{P} \left( |N_d|_{\min} \geq y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2} \right) = \frac{1}{d!} \left( \frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right)^d (1 + o(1)). \quad (3.2.13) \]

Now submitting (3.2.13) into (3.2.11) and (3.2.12), we get
\[ \lim_{n,p \to \infty} \mathbb{P} \left( \max_{1 \leq k \leq q} \hat{V}_k^2 \geq y_p \right) \leq \sum_{d=1}^{2m} (-1)^{d-1} \frac{1}{d!} \left( \frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right)^d \]
and
\[ \lim_{n,p \to \infty} \mathbb{P} \left( \max_{1 \leq k \leq q} \hat{V}_k^2 \geq y_p \right) \geq \sum_{d=1}^{2m-1} (-1)^{d-1} \frac{1}{d!} \left( \frac{1}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right)^d \]
for any positive integer \( m \). Letting \( m \to \infty \), we obtain (3.2.7).

Now we prove Lemma 7. When \( d = 1 \), we have by the tail probabilities of normal distribution,
\[ \mathbb{P} \left( |N_1|_{\min} \geq y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2} \right) = (1 + o(1)) \frac{2p^{-2}}{\sqrt{8\pi}} \exp \left( -x/2 \right). \]
This implies (3.2.13) in Lemma 7. It remains to prove the lemma when \( d \geq 2 \). Note that for any \((i, j) \in A \setminus A_0\) and \((k, l) \in A \setminus A_0\), we have

\[
\text{Cov}(X_i X_j, X_k X_l) = E X_i X_j X_k X_l + O((\log p)^{-2-2\alpha_0}).
\]

Define graph \( G_{abcd} = (V_{abcd}, E_{abcd}) \), where \( V_{abcd} = \{a, b, c, d\} \) is the set of vertices and \( E_{abcd} \) is the set of edges. There is an edge between \( i \neq j \in \{a, b, c, d\} \) if and only if \( |\sigma_{ij}| \geq (\log p)^{-1-\alpha_0} \). We say \( G_{abcd} \) is a three vertices graph (3-G) if the number of different vertices in \( V_{abcd} \) is 3. Similarly, \( G_{abcd} \) is a four vertices graph (4-G) if the number of different vertices in \( V_{abcd} \) is 4. A vertex in \( G_{abcd} \) is said to be isolated if there is no edge connected to it. Note that for any \( 1 \leq m_1 \neq m_2 \leq q \), \( G_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} \) is 3-G or 4-G. We say a graph \( G := G_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} \) satisfy \((\ast)\) if

\[
(\ast): \quad \text{If } G \text{ is 4-G, then there is at least one isolated vertex in } G; \\
\text{otherwise } G \text{ is 3-G and } E_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} = \emptyset.
\]

For any \( G_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} \) satisfying \((\ast)\), by (C3),

\[
|E X_{i_{m_1}} X_{j_{m_1}} X_{i_{m_2}} X_{j_{m_2}}| = O((\log p)^{-1-\alpha_0}), \tag{3.2.14}
\]

where \( O(1) \) is uniformly for \( i_{m_1}, j_{m_1}, i_{m_2}, j_{m_2} \). We now define the following set

\[
\mathcal{I} &= \{1 \leq k_1 < \cdots < k_d \leq q\}, \\
\mathcal{I}_0 &= \{1 \leq k_1 < \cdots < k_d \leq q : \text{for some } m_1, m_2 \in \{k_1, \cdots, k_d\} \text{ with } m_1 \neq m_2 \} \\
G := G_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} \text{ does not satisfy } (\ast)\}, \\
\mathcal{I}_c &= \{1 \leq k_1 < \cdots < k_d \leq q : \text{for any } m_1, m_2 \in \{k_1, \cdots, k_d\} \text{ and } m_1 \neq m_2, \}
\]

63
\[ G \text{ satisfies } (\ast) \}.

It is easy to see that \( \mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_0^c \). For any subset \( S \) of \( \{k_1, \ldots, k_d\} \), we say that \( S \) satisfies (\ast\ast) if

\[
(\ast\ast) \quad \text{for any } m_1 \neq m_2 \in S, \ G_{i_{m_1}j_{m_1}}i_{m_2}j_{m_2} \text{ satisfies } (\ast).
\]

For \( 2 \leq l \leq d \), let

\[
\mathcal{I}_{0l} = \{ 1 \leq k_1 < \cdots < k_d \leq q : \text{the largest cardinality of } S \text{ is } l, \text{ where } S \text{ is any subset of } \{k_1 < \cdots, k_d\} \text{ satisfies } (\ast\ast) \},
\]

\[
\mathcal{I}_{01} = \{ 1 \leq k_1 < \cdots < k_d \leq q : \text{for any } m_1, m_2 \in \{k_1, \cdots, k_d\} \text{ with } m_1 \neq m_2 \}
\]

\[
\mathcal{G} := G_{i_{m_1}j_{m_1}i_{m_2}j_{m_2}} \text{ does not satisfy } (\ast) \}.
\]

Clearly \( \mathcal{I}_0^c = \mathcal{I}_0^d \) and \( \mathcal{I}_0 = \bigcup_{l=1}^{d-1} \mathcal{I}_{0l} \). We can prove that \( \text{Card}(\mathcal{I}_{0l}) \leq C_dq^{l+2\gamma(d-l)} \) and \( \text{Card}(\mathcal{I}_0^c) = (1 + o(1))C_q^d \). We claim that

\[
\sum_{\mathcal{I}_0^c} P\left( |N_{d}|_{\text{min}} \geq y_p^{1/2} \pm \varepsilon_n (\log p)^{-1/2} \right) = (1 + o(1))d! \left( \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{x}{2}\right) \right)^d (3.2.15)
\]

and

\[
\sum_{\mathcal{I}_0} P\left( |N_{d}|_{\text{min}} \geq y_p^{1/2} \pm \varepsilon_n (\log p)^{-1/2} \right) = o(1). \quad (3.2.16)
\]

By (3.2.15) and (A.4.19), Lemma 7 is proved.

We first prove (A.4.19). For \( 1 \leq a \neq b \leq q \), define the indicator function

\[
d( (i_a, j_a), (i_b, j_b)) = 1 \quad \text{if } G_{i_{ia}j_{ia}i_{jb}j_{jb}} \text{ does not satisfy } (\ast);
\]

\[
= 0 \quad \text{otherwise}.
\]

64
We further divide $\mathcal{I}_{0l}$ as follows. Let $(k_1, \ldots, k_d) \in \mathcal{I}_{0l}$ and let $S_* \subset (k_1, \ldots, k_d)$ be the largest cardinality subset satisfying $(\ast \ast)$. (If there are more than two subsets that attain the the largest cardinality, then we can choose any one of them.) Define

$$\mathcal{I}_{0l1} = \{(k_1, \ldots, k_d) \in \mathcal{I}_{0l} : \text{there exists an } a \notin S_* \text{ such that for some } b_1, b_2 \in S_* \text{ with } b_1 \neq b_2, d((i_a, j_a), (i_{b_1}, j_{b_1})) = 1, d((i_a, j_a), (i_{b_2}, j_{b_2})) = 1\},$$

$$\mathcal{I}_{0l2} = \mathcal{I}_{0l} \setminus \mathcal{I}_{0l1}.$$

Note that $\mathcal{I}_{0l1} = \emptyset$ and $\mathcal{I}_{0l2} = \mathcal{I}_{0l}$. Recall that $d$ is fixed and $l \leq d - 1$. We can prove that $\text{Card}(\mathcal{I}_{0l1}) \leq C_d l^{-1} + 2^\gamma (d - l + 1)$ and $\text{Card}(\mathcal{I}_{0l2}) \leq C_d l^{d-2} + 2^\gamma (d - l)$. Write $S_* = (b_1, \ldots, b_l)$ and $x_p = y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2}$. For any $(k_1, \ldots, k_d) \in \mathcal{I}_{0l}$,

$$P\left(|\mathbf{N}_d|_{\min} \geq y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2}\right) \leq P\left(|N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p\right) = \frac{1}{(2\pi)^{l/2} |\mathbf{U}_l|^{1/2}} \int_{|\mathbf{y}|_{\min} \geq x_p} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{U}_l^{-1} \mathbf{y}\right) d\mathbf{y},$$

where $\mathbf{U}_l$ is the covariance matrix of $(N_{b_1}, \ldots, N_{b_l})$. By (3.2.14), we have $||\mathbf{U}_l - \mathbf{I}_l||_2 = O((\log p)^{-1-\alpha_0})$. Let $|\mathbf{y}|_{\max} = \max_{1 \leq i \leq l} |y_i|$ for $\mathbf{y} = (y_1, \ldots, y_l)$. Then

$$\frac{1}{(2\pi)^{l/2} |\mathbf{U}_l|^{1/2}} \int_{|\mathbf{y}|_{\min} \geq x_p} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{U}_l^{-1} \mathbf{y}\right) d\mathbf{y} = \frac{1}{(2\pi)^{l/2} |\mathbf{U}_l|^{1/2}} \int_{|\mathbf{y}|_{\min} \geq x_p, |\mathbf{y}|_{\max} \leq (\log p)^{1/2 + \alpha_0/4}} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{U}_l^{-1} \mathbf{y}\right) d\mathbf{y} + O\left(\exp \left(- (\log p)^{1+\alpha_0/2}/4\right)\right)$$

$$= \frac{1}{(2\pi)^{l/2} |\mathbf{U}_l|^{1/2}} \int_{|\mathbf{y}|_{\min} \geq x_p, |\mathbf{y}|_{\max} \leq (\log p)^{1/2 + \alpha_0/4}} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{y}\right) d\mathbf{y} + O\left(\exp \left(- (\log p)^{1+\alpha_0/2}/4\right)\right)$$

$$= \frac{1}{(2\pi)^{l/2}} \int_{|\mathbf{y}|_{\min} \geq x_p, |\mathbf{y}|_{\max} \leq (\log p)^{1/2 + \alpha_0/4}} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{y}\right) d\mathbf{y} + O\left(\exp \left(- (\log p)^{1+\alpha_0/2}/4\right)\right)$$

$$= \frac{1}{(2\pi)^{l/2}} \int_{|\mathbf{y}|_{\min} \geq x_p} \exp \left(-\frac{1}{2} \mathbf{y}^T \mathbf{y}\right) d\mathbf{y} + O\left(\exp \left(- (\log p)^{1+\alpha_0/2}/4\right)\right)$$

65
For \(k\) for generality and for notation briefness, we can assume that \(d(k) \leq 1\) for \(1 \leq i \leq l\). That is, \(\|p\| = \|b\| = \|x\| = \|y\| = 1\). Let \(\bar{V}_{\alpha} \subseteq S(\alpha)\) be the covariance matrix of \((b_{\alpha}, d_{\alpha}, \ldots, k_{\alpha})\), where \(N, l \in \mathbb{R}\). It follows that \(\|\bar{V}_{\alpha} - V_{\alpha}\|_2 = O((\log p)^{-1-\alpha})\), where \(V_{\alpha} = \text{diag}(D, I_{l-1})\) and \(D\) is the covariance matrix of \((N_{\alpha}, l_{\alpha}, \ldots, k_{\alpha})\).

We say the graph \(G_{ia,jia,jb}\) is aG-bE if \(G_{ia,jia,jb}\) is a-G and there are \(b\) edges in \(E_{ia,jia,jb}\) for \(a = 3, 4\) and \(b = 0, 1, 2, 3, 4\). We further divide \(\mathcal{I}_{0l2}\) into two parts:

\[
\mathcal{I}_{0l2,1} = \left\{(k_1, \ldots, k_d) \in \mathcal{I}_{0l2} : G_{ia,jia,jb}\right\}, \\
\mathcal{I}_{0l2,2} = \left\{(k_1, \ldots, k_d) \in \mathcal{I}_{0l2} : G_{ia,jia,jb}\right\}.
\]

Note that \(\mathcal{I}_{0l2} = \mathcal{I}_{0l2,1} \cup \mathcal{I}_{0l2,2}\). We can prove that \(\text{Card}(\mathcal{I}_{0l2,2}) \leq C p^{2l^2-2+2(d-1)\gamma} \times p^{1+3\gamma}\), where \(C\) is a constant depending only on the fixed number \(d\). This, together

\[
+ O\left(\exp\left(- (\log p)^{1+\alpha_0/2}/4\right)\right) \\
= O\left(p^{-2l}\right). \tag{3.2.17}
\]

This implies that \(\sum_{\mathcal{I}_{0l2}} P\left(|N_{d}|_{\min} \geq y_p^{1/2} \pm \epsilon_n(\log p)^{-1/2}\right) \leq C p^{-2+4\gamma(d-l+1)} = o(1). \tag{3.2.18}\)

For \((k_1, \ldots, k_d) \in \mathcal{I}_{0l2}\), let \(\bar{a} = \min\{a : a \in (k_1, \ldots, k_d), a \notin S_\ast\}\). Without loss of generality and for notation briefness, we can assume that \(d((i_{\bar{a}}, j_{\bar{a}}), (i_b, j_b)) = 1\).

Then we have

\[
\sum_{\mathcal{I}_{0l2}} P\left(|N_{d}|_{\min} \geq y_p^{1/2} \pm \epsilon_n(\log p)^{-1/2}\right) \\
\leq \sum_{\mathcal{I}_{0l2}} P\left(|N_{\bar{a}}| \geq x_p, |N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p\right).
\]

Because \((k_1, \ldots, k_d) \in \mathcal{I}_{0l2}\), by (C3), we can show \(\text{Cov}(N_{\bar{a}}, N_{b_j}) = O((\log p)^{-1-\alpha})\) for \(2 \leq j \leq l\). Recall, \(S_\ast = (b_1, \ldots, b_l)\). We have \(\text{Cov}(N_{b_i}, N_{b_j}) = O((\log p)^{-1-\alpha})\) for \(1 \leq i \neq j \leq l\). Let \(\bar{V}_{\alpha}\) be the covariance matrix of \((N_{\bar{a}}, N_{b_1}, \ldots, N_{b_l})\). It follows that \(\|\bar{V}_{\alpha} - \bar{V}_{\alpha}\|_2 = O((\log p)^{1-\alpha})\), where \(\bar{V}_{\alpha} = \text{diag}(D, I_{l-1})\) and \(D\) is the covariance matrix of \((N_{\bar{a}}, N_{b_1})\).
with (3.2.17), implies that

\[
\sum_{I_{0\ell,2}} P\left( |N_{\bar{a}}| \geq x_p, |N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p \right) \\
\leq \sum_{I_{0\ell,2}} P\left( |N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p \right) \\
\leq C p^{2l-1+(2d-2l+1)\gamma} \times p^{-2l} = o(1).
\] (3.2.19)

For \((k_1, \ldots, k_d) \in I_{0\ell,1}\), by (C3), we have for large \(p\),

\[
\frac{1}{\theta_{i_{a}j_{a}}^{1/2} \theta_{i_{b}j_{b}}^{1/2}} |EX_{i_{a}}X_{j_{a}}X_{i_{b}}X_{j_{b}}| \\
\leq \max\{|\rho_{i_{a}i_{a}}\rho_{j_{a}j_{a}}|, |\rho_{i_{a}j_{a}}\rho_{j_{a}i_{b}}|\} + O((\log p)^{-\alpha_0}) \\
\leq r + O((\log p)^{-\alpha_0}) < (r + 1)/2
\]
uniformly for all \(\bar{a}\) and \(b_1\). Recall the covariance matrix \(\text{Var}_I\) of \((N_{\bar{a}}, N_{b_1}, \ldots, N_{b_l})\) satisfying \(\|\text{Var}_I - \text{diag}(D, I_{l-1})\|_2 = O((\log p)^{-\alpha_0})\). Using the similar argument as that in (3.2.17), we can obtain that

\[
\sum_{I_{0\ell,1}} P\left( |N_{\bar{a}}| \geq x_p, |N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p \right) \\
\leq C \sum_{I_{0\ell,1}} \left[ P\left( |N_{\bar{a}}| \geq x_p, |N_{b_1}| \geq x_p \right) \times p^{-2l+2} + \exp\left( - (\log p)^{1+\alpha_0}/4 \right) \right] \\
\leq C \sum_{I_{0\ell,1}} \left[ p^{-2-(2-2r)/(3+r)} \times p^{-2l+2} + \exp\left( - (\log p)^{1+\alpha_0}/4 \right) \right],
\] (3.2.20)

where the last inequality follows from Lemma 8. We can show that \(\text{Card}(I_{0\ell,1}) \leq C p^{2l-2+2(d-l-1)\gamma} \times p^{2+2\gamma}\). Hence it follows from (3.2.20) that

\[
\sum_{I_{0\ell,1}} P\left( |N_{\bar{a}}| \geq x_p, |N_{b_1}| \geq x_p, \ldots, |N_{b_l}| \geq x_p \right) \leq C p^{-(2-2r)/(3+r)+2(d-l)\gamma} = o(1).
\] (3.2.21)

Combining (3.2.18), (3.2.19) and (3.2.21) yields (A.4.19).
It remains for us to show that (3.2.15). By (3.2.14), we have \[ \| \text{Cov}(N_d) - I_d \|_2 = O((\log p)^{-1-\alpha}) \] uniformly for \((k_1, \ldots, k_d) \in I_0^c\). Then by exactly the same proof as that in (3.2.17), we can get

\[
P\left( |N_d|_{\text{min}} \geq y_p^{1/2} \pm \varepsilon_n (\log p)^{-1/2} \right) = (1 + o(1)) \left( \frac{2}{\sqrt{8\pi}} \exp \left( -\frac{x}{2} \right) \right)^d p^{-2d}
\]

uniformly for \((k_1, \ldots, k_d) \in I_0^c\). This, together with the fact \(\text{Card}(I_0^c) = (1 + o(1)) C_d^d q^d\), proves (3.2.15).

The limiting behavior of \(M_n\) is similar to that of the largest off-diagonal entry \(L_n\) of the sample correlation matrix in Jiang (2004), wherein the paper derived the asymptotic distribution of \(L_n\) under the assumption that the components \(X_1, \ldots, X_p\) are independent. Further extensions and improvements are given in Zhou (2007), Liu, Lin and Shao (2008) and Cai and Jiang (2011, 2012). A key assumption in these papers is the independence between the components. The techniques in their proofs cannot be used to obtain (3.2.3), since dependent components are allowed in Theorem 7. The proof of (3.2.3) requires different techniques.

The limiting null distribution given in (3.2.3) shows that the test \(\Phi_\alpha\) defined in (3.1.3) is an asymptotically level \(\alpha\) test. That is,

\[
P(\text{Type I error}) = P_{H_0}(\Phi_\alpha = 1) \to \alpha.
\]

The limiting null distribution in Theorem 7 requires Conditions (C1) and (C3). However, without (C1) and (C3), the size of the test can still be effectively controlled.
Proposition 1. Under (C2) (or (C2)*), for $0 < \alpha < 1$,

$$P(\text{Type I error}) = P_{H_0}(\Phi_\alpha = 1) \leq -\log(1 - \alpha) + o(1).$$  \hspace{1cm} (3.2.23)

Proof of Proposition 1: According to the proof of Theorem 7 and letting $d = 1$ in (3.2.11), we have

$$P_{H_0}(M_{ij} \geq q_\alpha + 4 \log p - \log \log p) = (1 + o(1))P(|N(0, 1)| \geq q_\alpha + 4 \log p - \log \log p).$$

Then we can get, under (C2) (or (C2)*), for $0 < \alpha < 1,$

$$\text{Type I error} = P_{H_0}(\Phi_\alpha = 1) \leq \sum_{1 \leq i \leq j \leq p} P_{H_0}(M_{ij} \geq q_\alpha + 4 \log p - \log \log p) \leq -\log(1 - \alpha) + o(1).$$

Note that for commonly used significant level $\alpha = 0.05$, $-\log(1 - \alpha) = 0.05129$, which is close to 0.05, and for $\alpha = 0.01$, $-\log(1 - \alpha) = 0.01005$, which is even closer to the nominal level. It is easy to see that $-\log(1 - \alpha) \approx \alpha$ for small $\alpha$. So, the test $\Phi_\alpha$ effectively controls the probability of type I error without (C1) and (C3).

We now turn to an analysis of the power of the test $\Phi_\alpha$ given in (3.1.3). We shall define the following class of matrices:

$$U(c) = \left\{ (\Sigma_1, \Sigma_2) : \max_{1 \leq i \leq j \leq p} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}} \geq c\sqrt{\log p} \right\}. \hspace{1cm} (3.2.24)$$

The next theorem shows that the null parameter set in which $\Sigma_1 = \Sigma_2$ is asymptotically distinguishable from $U(4)$ by the test $\Phi_\alpha$. That is, $H_0$ is rejected by $\Phi_\alpha$ with overwhelming probability if $(\Sigma_1, \Sigma_2) \in U(4).$
Theorem 8. Suppose that (C2) or (C2*) holds. Then as \( n, p \to \infty \),

\[
\inf_{(\Sigma_1, \Sigma_2) \in U(4)} P\left( \Phi_\alpha = 1 \right) \to 1.
\] (3.2.25)

Proof of Theorem 8: Recall that

\[
M_n^1 = \max_{1 \leq i \leq j \leq p} \left( \frac{\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2} - \sigma_{ij1} + \sigma_{ij2}}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2} \right)^2.
\]

By Lemmas 10 and 11,

\[
P\left( M_n^1 \leq 4 \log p - \frac{1}{2} \log \log p \right) \to 1
\] (3.2.26)
as \( n, p \to \infty \). By Lemma 10, the inequalities

\[
\max_{1 \leq i \leq j \leq p} \frac{(\sigma_{ij1} - \sigma_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} \leq 2M_n^1 + 2M_n
\] (3.2.27)

and

\[
\max_{1 \leq i \leq j \leq p} \frac{(\sigma_{ij1} - \sigma_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} \geq 16 \log p,
\]

which yields that \( P\left( M_n \geq q_\alpha + 4 \log p - \log \log p \right) \to 1 \) as \( n, p \to \infty \).

It can be seen from Theorem 8 that it only requires one of the entries of \( \Sigma_1 - \Sigma_2 \) having a magnitude more than \( C \sqrt{\log p/n} \) in order for the test \( \Phi_\alpha \) to correctly reject \( H_0 \). This lower bound is rate-optimal. Denote by \( \mathcal{P} \) the collection of distributions satisfying (C2) or (C2*). Let \( \mathcal{T}_\alpha \) be the set of \( \alpha \)-level tests over \( \mathcal{P} \), i.e., \( P(T_\alpha = 1) \leq \alpha \) under \( H_0 \) over all distributions in \( \mathcal{P} \) for any \( T_\alpha \in \mathcal{T}_\alpha \).

Theorem 9. Suppose that (C2) or (C2*) holds. Let \( \alpha, \beta > 0 \) and \( \alpha + \beta < 1 \). Then there exists a constant \( c_0 > 0 \) such that for all large \( n \) and \( p \),

\[
\inf_{(\Sigma_1, \Sigma_2) \in U(c_0)} \sup_{T_\alpha \in \mathcal{T}_\alpha} P\left( T_\alpha = 1 \right) \leq 1 - \beta.
\] (3.2.28)
The difference between $\Sigma_1$ and $\Sigma_2$ is measured by $\max_{1 \leq i \leq j \leq p} |\sigma_{ij1} - \sigma_{ij2}|$ in Theorem 9. Another measurement is based on the Frobenius norm $\|\Sigma_1 - \Sigma_2\|_F$.

Suppose that $\Sigma_1 - \Sigma_2$ is a sparse matrix with $c_0(p)$ nonzero entries. That is,

$$c_0(p) = \sum_{i=1}^{p} \sum_{j=1}^{p} I\{\sigma_{ij1} - \sigma_{ij2} \neq 0\}.$$ 

We introduce the following class of matrices for $\Sigma_1 - \Sigma_2$:

$$\mathcal{V}(c) = \left\{ (\Sigma_1, \Sigma_2) : \max_i |\sigma_{ii1}| \leq K, \max_i |\sigma_{ii2}| \leq K, \|\Sigma_1 - \Sigma_2\|_F^2 \geq c c_0(p) \frac{\log p}{n} \right\}.$$ 

Note that on $\mathcal{V}(c)$, we have $\max_{1 \leq i \leq j \leq p} |\sigma_{ij1} - \sigma_{ij2}| \geq \sqrt{c \log p/n}$. Thus, for sufficiently large constant $c$,

$$\inf_{(\Sigma_1, \Sigma_2) \in \mathcal{V}(c)} \mathbb{P}\left( \Phi_{\alpha} = 1 \right) \to 1$$

as $n, p \to \infty$. The following theorem shows that the lower bound for $\|\Sigma_1 - \Sigma_2\|_F^2$ in $\mathcal{V}(c)$ is rate-optimal. That is, no $\alpha$-level test can reject $H_0$ with overwhelming probability uniformly over $\mathcal{V}(c_0)$ for some $c_0 > 0$.

**Theorem 10.** Suppose that (C2) or (C2*) holds. Assume that $c_0(p) \leq p^r$ for some $0 < r < 1/2$. Let $\alpha, \beta > 0$ and $\alpha + \beta < 1$. There exists a constant $c_0 > 0$ such that for all large $n$ and $p$,

$$\inf_{(\Sigma_1, \Sigma_2) \in \mathcal{V}(c_0)} \sup_{T_{\alpha} \in \mathcal{T}_{\alpha}} \mathbb{P}\left( T_{\alpha} = 1 \right) \leq 1 - \beta.$$ 

Proof of Theorem 10: It suffices to take $\mathcal{T}_{\alpha}$ to be the set of $\alpha$-level tests over the normal distributions, since it contains all the $\alpha$-level tests over the collection
of distributions satisfying (C2) or (C2*). Let \( M \) denote the set of all subsets of \( \{1,\ldots,p\} \) with cardinality \( c_0(p) \). Let \( \hat{m} \) be a random subset of \( \{1,\ldots,p\} \), which is uniformly distributed on \( M \). We construct a class of \( \Sigma_1, \mathcal{N} = \{ \Sigma_{\hat{m}}, \hat{m} \in M \} \), such that

\[
\sigma_{ij} = 0 \text{ for } i \neq j \text{ and } \sigma_{ii} - 1 = \rho 1_{i \in \hat{m}},
\]

for \( i, j = 1,\ldots,p \) and \( \rho = c \sqrt{\log p/n} \), where \( c > 0 \) will be specified later. Let \( \Sigma_2 = I \) and \( \Sigma_1 \) be uniformly distributed on \( \mathcal{N} \). Let \( \mu_\rho \) be the distribution of \( \Sigma_1 - I \). Note that \( \mu_\rho \) is a probability measure on \( \{ \Delta \in S(c_0(p)) : \|\Delta\|_F^2 = c_0(p)\rho^2 \} \), where \( S(c_0(p)) \) is the class of matrices with \( c_0(p) \) nonzero entries. Let \( dP_1(\{X_n, Y_n\}) \) be the likelihood function given \( \Sigma_1 \) being uniformly distributed on \( \mathcal{N} \) and

\[
L_{\mu_\rho} := L_{\mu_\rho}(\{X_n, Y_n\}) = E_{\mu_\rho}(dP_1(\{X_n, Y_n\}))/dP_0(\{X_n, Y_n\}),
\]

where \( E_{\mu_\rho}(\cdot) \) is the expectation on \( \Sigma_1 \). By the arguments in Baraud (2002, p. 595), it suffices to show that \( EL_{\mu_\rho}^2 \leq 1 + o(1) \). It is easy to see that

\[
L_{\mu_\rho} = E_{\hat{m}} \left( \prod_{i=1}^n \frac{1}{|\Sigma_{\hat{m}}|^1/2} \exp \left( -\frac{1}{2} Z_i^T (\Omega_{\hat{m}} - I) Z_i \right) \right),
\]

where \( \Omega_{\hat{m}} = \Sigma_{\hat{m}}^{-1} \) and \( Z_1,\ldots,Z_n \) are i.i.d multivariate normal vectors with mean vector 0 and covariance matrix \( I \). Thus,

\[
\begin{align*}
EL_{\mu_\rho}^2 &= E \left( \frac{1}{\binom{p}{k}} \sum_{m \in M} \left( \prod_{i=1}^n \frac{1}{|\Sigma_m|^{1/2}} \exp \left( -\frac{1}{2} Z_i^T (\Omega_m - I) Z_i \right) \right) \right)^2 \\
&= \frac{1}{\binom{p}{k}^2} \sum_{m,m' \in M} E \left( \prod_{i=1}^n \frac{1}{|\Sigma_m|^{1/2}} \frac{1}{|\Sigma_{m'}|^{1/2}} \exp \left( -\frac{1}{2} Z_i^T (\Omega_m + \Omega_{m'} - 2I) Z_i \right) \right) \).
\end{align*}
\]

Set \( \Omega_m + \Omega_{m'} - 2I = (a_{ij}) \). Then \( a_{ij} = 0 \) for \( i \neq j \), \( a_{jj} = 0 \) if \( j \in (m \cup m')^c \), \( a_{jj} = 2(\frac{1}{1+\rho} - 1) \) if \( j \in m \cap m' \) and \( a_{jj} = \frac{1}{1+\rho} - 1 \) if \( j \in m \setminus m' \) and \( m' \setminus m \). Let
\( t = |m \cap m'|. \) Then we have

\[
\mathbb{E}L^2_{\mu_{\rho}} = \frac{1}{(p)} \sum_{t=0}^{k_p} \binom{k_p}{t} \left( \frac{p - k_p}{t} \right) \left( \frac{1}{1 + \rho} \right)^{k_p n} (1 + \rho)^{(k_p - t)n} \left( \frac{1}{1 - \rho} \right)^{\frac{tn}{2}} \\
\leq \ p^{k_p} (p - k_p)! / p! \sum_{t=0}^{k_p} \binom{k_p}{t} \left( \frac{k_p}{t} \right) \left( \frac{1}{1 - \rho^2} \right)^{\frac{tn}{2}} \\
= \ (1 + o(1)) \left( 1 + \frac{k_p}{p(1 - \rho^2)^{n/2}} \right)^{k_p},
\]

for \( r < 1/2. \) So

\[
\mathbb{E}L^2_{\mu_{\rho}} \leq \exp \left( k_p \log (1 + k_p p^2 - 1) \right) (1 + o(1)) \leq \exp \left( k_p^2 p^2 - 1 \right) (1 + o(1)) = 1 + o(1)
\]

by letting \( c \) be sufficiently small. Theorem 10 is proved. \( \square \)

Note that for every \( c > 0, \) there exists some constant \( K(c) > 0 \) such that for any \( 0 < c_0 < K(c), \mathcal{V}(c) \subset U(c_0). \) Thus, Theorem 9 follows from Theorem 10 directly.

### 3.2.3 Comparison with other tests

We now compare the power of the test \( \Phi_\alpha \) with those of the tests in Li and Chen (2012), Srivastava and Yanagihara (2010), and Schott (2007) under the sparse alternatives. Define the following class of matrices:

\[
\mathcal{S}(s_p, c_{n,p}) = \left\{ (\Sigma_1, \Sigma_2) : \sum_{j=1}^{p} \sum_{i=1}^{p} I\{\sigma_{ij1} - \sigma_{ij2} \neq 0\} \leq s_p, \right. \\
K^{-1} \leq \min_{1 \leq i \leq p} |\sigma_{iil}| \leq \max_{1 \leq i \leq p} |\sigma_{iil}| \leq K \text{ for } l = 1, 2, \\
4\sqrt{\log p} \leq \max_{1 \leq i \leq p} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}} \leq c_{n,p}, \left. \right\},
\]

where we assume \( s_p = o(\min\{p/c_{n,p}^2, p/\sqrt{nc_{n,p}}\}). \) Note that \( \mathcal{S}(s_p, c_{n,p}) \subset U(4). \) Hence the power of the test \( \Phi_\alpha \) converges to one by Theorem 2.
We now turn to an analysis of the powers of the tests given in Li and Chen (2012), Schott (2007) and Srivastava and Yanagihara (2010). We first recall the conditions used in these three papers.

**D1.** For any $i, j, k, l \in \{1, 2\}$,

\[
\text{tr}(\Sigma_k \Sigma_l) \to \infty \quad \text{and} \quad \text{tr}\{\Sigma_i \Sigma_j \Sigma_k \Sigma_l\} = o\{\text{tr}(\Sigma_i \Sigma_j)\text{tr}(\Sigma_k \Sigma_l)\}.
\]

**D2.** $\lim_{p \to \infty} \frac{1}{p} \text{tr}(\Sigma_1^i) = \gamma_{i1} \in (0, \infty)$ and $\lim_{p \to \infty} \frac{1}{p} \text{tr}(\Sigma_2^i) = \gamma_{i2} \in (0, \infty)$.

**D3.** $p/n_1 \to b_1$ and $p/n_2 \to b_2$, where $b_1, b_2 \in (0, \infty)$.

Let $\beta_{LC}(\alpha)$, $\beta_{SY}(\alpha)$ and $\beta_{Sc}(\alpha)$ be respectively the powers of the tests given in Li and Chen (2012), Srivastava and Yanagihara (2010) and Schott (2007) with the significance level being controlled at level $\alpha$. We shall make the comparisons under the normality condition since two of these tests require this assumption.

**Proposition 2.** Suppose that $X \sim N(\mu_1, \Sigma_1)$ and $Y \sim N(\mu_2, \Sigma_2)$. Assume that $(\Sigma_1, \Sigma_2) \in \mathcal{S}(s_p, c_{n,p})$.

(i). For the proposed test $\Phi_\alpha$ defined in (3.1.3),

\[
\lim_{n,p \to \infty} P(\Phi_\alpha = 1) = 1.
\]

(ii). Suppose (D1) holds, then

\[
\lim_{n,p \to \infty} \beta_{LC}(\alpha) = \alpha.
\]

(iii). Suppose (D2) holds for $i = 1, 2, 3, \text{and} 4$, then

\[
\lim_{n,p \to \infty} \beta_{SY}(\alpha) = \alpha.
\]
Proof of Proposition 2: Proposition 2 (i) is a directly result of Theorem 8. For (ii), by the analysis in Li and Chen (2012), pages 914 and 915, if

\[
\text{DIST}_{LC}(\Sigma_1, \Sigma_2) := \frac{\|\Sigma_1 - \Sigma_2\|_F^2}{\sqrt{\frac{1}{n_1}\|\Sigma_1\|_F^4 + \frac{1}{n_2}\|\Sigma_2\|_F^4}} = o(1),
\]  

(3.2.29)

then \(\beta_{LC}(\alpha) \to \alpha\). By some elementary calculations, we have \(\max_{i,j}\{\theta_{ij1} + \theta_{ij2}\} = O(1)\) and \(\max_{i,j}|\sigma_{ij1} - \sigma_{ij2}| = O(c_{n,p}/\sqrt{n})\) for \((\Sigma_1, \Sigma_2) \in \mathcal{S}(s_p, c_{n,p})\). Hence, \(\text{DIST}_{LC}(\Sigma_1, \Sigma_2) = O(s_p c_{n,p}/p)\). This shows that if \(s_p = o(p/c_{n,p}^2)\), the asymptotic power \(\lim_{(n,p) \to \infty} \beta_{LC}(\alpha) = \alpha\). For (iii), Lemma 2.3 in Srivastava and Yanagihara (2010) shows that, if

\[
\text{DIST}_{SY}(\Sigma_1, \Sigma_2) = n \left| \frac{\text{tr}(\Sigma_1^2)/p}{(\text{tr}(\Sigma_1)/p)^2} - \frac{\text{tr}(\Sigma_2^2)/p}{(\text{tr}(\Sigma_2)/p)^2} \right| = o(1),
\]  

(3.2.30)

then \(\beta_{SY}(\alpha) \to \alpha\). It can be shown that, for \((\Sigma_1, \Sigma_2) \in \mathcal{S}(s_p, c_{n,p})\), \(\text{DIST}_{SY}(\Sigma_1, \Sigma_2) = O(\sqrt{n}s_p c_{n,p}/p)\). This shows that, if \(s_p = o(p/(\sqrt{n}c_{n,p}))\), then the asymptotic power \(\lim_{(n,p) \to \infty} \beta_{SY}(\alpha) = \alpha\).

Proposition 2 shows that under the class of sparse alternatives in \(\mathcal{S}(s_p, c_{n,p})\), the tests given in Li and Chen (2012) and Srivastava and Yanagihara (2010) would suffer from trivial power while the proposed test \(\Phi_\alpha\) enjoys the full power. This fact is also illustrated in some of the simulation results given in Section 3.4.

For the test given in Schott (2007), there is no result on the limiting distribution under the alternative. But from the theoretical analysis given in Schott (2007), under Condition (D2) for \(i = 1, \ldots, 8\) and Condition (D3), one can see that in order
to ensure the power tending to one, it is required that

$$\text{DIST}_S(\Sigma_1, \Sigma_2) = \| \Sigma_1 - \Sigma_2 \|_F^2 \to \infty.$$  \hfill (3.2.31)

Note that, under (D2) for \(i = 1, \ldots, 8\) and (D3), since \(s_p = o(p/c_{n,p}^2)\), we have \(\text{DIST}_S(\Sigma_1, \Sigma_2) = o(1)\) for \((\Sigma_1, \Sigma_2) \in S(s_p, c_{n,p})\) and (3.2.31) is thus not satisfied.

Condition (C1) is generally much weaker than (D2) for \(i = 1, \ldots, 4\). For simplicity, let us assume the diagonal entries \(\sigma_{ii} = \sigma_{ii}^2 = 1\). Put \(\Upsilon\) in (C1) as

$$\Upsilon = \{1 \leq i \leq p : \sum_{j=1}^{p} \sigma_{ij1}^2 \geq \log p \text{ or } \sum_{j=1}^{p} \sigma_{ij2}^2 \geq \log p\}.$$  

Since \(\text{tr}(\Sigma_i^2) = \| \Sigma_i \|_F^2 = O(p)\) for \(i = 1\) and 2, we have \(\text{Card}(\Upsilon) = O(p/\log p)\). So \(\sum_{j=1}^{p} (\sigma_{ij1}^2 + \sigma_{ij2}^2) \leq 2 \log p\) uniformly for \(i \notin \Upsilon\). This implies that, uniformly for \(i \notin \Upsilon\), \(\sigma_{i(k)1}^2 \leq k^{-1} \sum_{j=1}^{p} \sigma_{ij1}^2 = O(k^{-1} \log p)\) for \(1 \leq k \leq p\), where \(\sigma_{i(1)}^2 \geq \ldots \geq \sigma_{i(p)}^2\).

Similar inequality holds for \(\sigma_{ij2}\). Hence (D2) implies \(\max_{j \notin \Upsilon} s_j(\alpha) = O(p^\gamma)\) for any \(\alpha > 0\) and \(\gamma > 0\), which is stronger than the condition on \(\max_{j \notin \Upsilon} s_j(\alpha_0)\) in (C1).

Additionally, Condition (C1) allows for \(\text{tr}(\Sigma_i^2) \asymp (p^2/(\log p)^{2+2\alpha})\).

### 3.3 Support recovery of \(\Sigma_1 - \Sigma_2\) and application to gene selection

We have focused on testing the equality of two covariance matrices \(\Sigma_1\) and \(\Sigma_2\) in Sections 3.1 and 3.2. As mentioned in the introduction, if the null hypothesis \(H_0 : \Sigma_1 = \Sigma_2\) is rejected, further exploring in which ways they differ is also of significant
interest in practice. Motivated by applications in gene selection, we consider in this section two related problems, one is recovering the support of $\Sigma_1 - \Sigma_2$ and another is identifying the rows on which the two covariance matrices differ from each other.

### 3.3.1 Support recovery of $\Sigma_1 - \Sigma_2$

The goal of support recovery is to find the positions at which the two covariance matrices differ from each other. The problem can also be viewed as simultaneous testing of equality of individual entries between the two covariance matrices. Denote the support of $\Sigma_1 - \Sigma_2$ by

$$\Psi = \Psi(\Sigma_1, \Sigma_2) = \{ (i, j) : \sigma_{ij}^1 \neq \sigma_{ij}^2 \}. \quad (3.3.1)$$

In certain applications, the variances along the diagonal play a more important role than the covariances. For example, in a differential variability analysis of gene expression Ho, et al (2008) proposed to select genes based on the differences in the variances. In this section we shall treat the variances along the diagonal differently from the off-diagonal covariances for support recovery. Since there are $p$ diagonal elements, $M_{ii}$ along the diagonal are thresholded at $2 \log p$ based on the extreme values of normal variables. The off-diagonal entries $M_{ij}$ with $i \neq j$ are thresholded at a different level. More specifically, set

$$\hat{\Psi}(\tau) = \{ (i, i) : M_{ii} \geq 2 \log p \} \cup \{ (i, j) : M_{ij} \geq \tau \log p, \ i \neq j \}, \quad (3.3.2)$$

where $M_{ij}$ are defined in (3.1.1) and $\tau$ is the threshold constant for the off-diagonal entries.
The following theorem shows that with $\tau = 4$ the estimator $\hat{\Psi}(4)$ recovers the support $\Psi$ exactly with probability tending to 1 when the magnitudes of nonzero entries are above certain thresholds. Define

$$W_0(c) = \left\{ (\Sigma_1, \Sigma_2) : \min_{(i,j) \in \Psi} \frac{|\sigma_{ij1} - \sigma_{ij2}|}{\sqrt{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}} \geq c\sqrt{\log p} \right\}.$$ 

We have the following result.

**Theorem 11.** Suppose that (C2) or (C2$^*$) holds. Then as $n, p \to \infty$,

$$\inf_{(\Sigma_1, \Sigma_2) \in W_0(4)} P(\hat{\Psi}(4) = \Psi) \to 1.$$ 

Proof of Theorem 11: The proof of Theorem 11 is similar to that of Theorem 8. In fact, by (3.2.26) and a similar inequality as (3.2.27), we can get

$$P\left( \min_{(i,j) \in \Psi} M_{ij} \geq 4 \log p \right) \to 1$$

uniformly for $(\Sigma_1, \Sigma_2) \in W_0(4)$. \hfill \Box

The choice of the thresholding constant $\tau = 4$ is optimal for the exact recovery of the support $\Psi$. Consider the class of $s_0(p)$-sparse matrices,

$$S_0 = \left\{ A = (a_{ij})_{p \times p} : \max_i \sum_{j=1}^p I\{a_{ij} \neq 0\} \leq s_0(p) \right\}.$$ 

We show in the following theorem that, for any threshold constant $\tau < 4$, the probability of recovering the support exactly tends to 0 for all $\Sigma_1 - \Sigma_2 \in S_0$. This is mainly due to the fact that the threshold level $\tau \log p$ is too small to ensure that the zero entries of $\Sigma_1 - \Sigma_2$ are estimated by zero. We assume that
(C1*). There exists a subset $\Theta \subset \{1, 2, \ldots, p\}$ with $\text{Card}(\Theta) \geq p^\gamma$ for all $0 < \gamma < 1$ such that $|\rho_{ij1}| \leq C(\log p)^{-1-\alpha_0}$ and $|\rho_{ij2}| \leq C(\log p)^{-1-\alpha_0}$ for all $i \neq j \in \Theta$ and some $\alpha_0 > 0$.

It is easy to see that (C1*) is much weaker than (C1) because it allows $p - p^\gamma$ variables to be arbitrarily correlated.

**Theorem 12.** Suppose that (C1*), (C2) (or (C2*)) and (C3) hold. Let $0 < \tau < 4$. If $s_0(p) = O(p^{1-\tau_1})$ with some $\tau_1 < 1$, then as $n, p \to \infty$,

$$
\sup_{\Sigma_1, \Sigma_2 \in \mathcal{S}_0} P\left(\hat{\Psi}(\tau) = \Psi\right) \to 0.
$$

Proof of Theorem 12: Without loss of generality, we assume $\Theta = \{1, 2, \ldots, p_0\}$ with $p_0 = \text{Card}(\theta) \geq p^\gamma$ for all $0 < \gamma < 1$. Let $A_1$ be the largest subset of $\Theta \setminus \{1\}$ such that $\sigma_{1k1} = \sigma_{1k2}$ for all $k \in A_1$. Let $i_1 = \min\{j : j \in A_1, j > 1\}$. Then we have $|i_1 - 1| \leq s_0(p)$. Also, $\text{Card}(A_1) \geq p_0 - s_0(p)$. Similarly, let $A_l$ be the largest subset of $A_{l-1} \setminus \{i_{l-1}\}$ such that $\sigma_{i_{l-1}k1} = \sigma_{i_{l-1}k2}$ for all $k \in A_l$ and $i_l = \min\{j : j \in A_l, j > i_{l-1}\}$. We can see that $i_l - i_{l-1} \leq s_0(p)$ for $l < p_0/s_0(p)$ and $\text{Card}(A_l) \geq \text{Card}(A_{l-1}) - s_0(p) \geq p_0 - (s_0(p) + 1)l$. Let $l = \lceil p^{\tau_2} \rceil$ with $\tau/4 < \tau_2 < \tau_1$. Let $\Sigma_{1l}$ and $\Sigma_{2l}$ be the covariance matrices of $(X_{i_0}, \ldots, X_{i_l})$ and $(Y_{i_0}, \ldots, Y_{i_l})$. Then the entries of $\Sigma_{1l}$ and $\Sigma_{2l}$ are the same except for the diagonal. Hence it follows from the proof of Theorem 7 that

$$
P\left(\max_{0 \leq j < k \leq l} M_{ij,ik} - 4 \log l + \log \log l \leq x\right) \to \exp\left(-\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{x^2}{2}\right)\right)
$$
uniformly for all $\Sigma_1 - \Sigma_2 \in S_0$. This implies that

$$\inf_{\Sigma_1 - \Sigma_2 \in S_0} \mathbb{P}\left( \max_{0 \leq j < k \leq l} M_{ij,ik} \geq c \log p \right) \to 1$$

for all $\tau < c < 4\tau_2$. By the definition of $\hat{\Psi}(\tau)$ and the fact $\sigma_{ij,ik1} = \sigma_{ij,ik2}$ for all $0 \leq j < k \leq l$, Theorem 12 is proved. 

In addition, the minimum separation of the magnitudes of the nonzero entries must be at least of order $\sqrt{\log p/n}$ to enable the exact recovery of the support. Consider, for example, $s_0(p) = 1$ in Theorem 10. Then Theorem 10 indicates that no test can distinguish $H_0$ and $H_1$ uniformly on the space $\mathcal{V}(c_0)$ with probability tending to one. It is easy to see that $\mathcal{V}(c_0) \subseteq \mathcal{W}_0(c_1)$ for some $c_1 > 0$. Hence, there is no estimate that can recover the support $\Psi$ exactly uniformly on the space $\mathcal{W}_0(c_1)$ with probability tending to one.

We have so far focused on the exact recovery of the support of $\Sigma_1 - \Sigma_2$ with probability tending to 1. It is sometimes desirable to recover the support under a less stringent criterion such as the family-wise error rate (FWER), defined by $\text{FWER} = \mathbb{P}(V \geq 1)$, where $V$ is the number of false discoveries. The goal of using the threshold level $4 \log p$ is to ensure $\text{FWER} \to 0$. To control the FWER at a pre-specified level $\alpha$ for some $0 < \alpha < 1$, a different threshold level is needed. For this purpose, we shall set the threshold level at $4 \log p - \log \log p + q_\alpha$ where $q_\alpha$ is given in (3.1.4). Define

$$\hat{\Psi}^* = \{(i, i) : M_{ii} \geq 2 \log p\} \cup \{(i, j) : M_{ij} \geq 4 \log p - \log \log p + q_\alpha, \ i \neq j\}.$$

We have the following result.
Proposition 3. Suppose that (C1), (C2) (or (C2∗)) and (C3) hold. Under $\Sigma_1 - \Sigma_2 \in \mathcal{S}_0 \cap \mathcal{W}_0(4)$ with $s_0(p) = o(p)$, we have as $n, p \to \infty$,

$$\mathbb{P}(\hat{\Psi}^* \neq \Psi) \to \alpha.$$ 

Proof of Proposition 3

By (3.3.3) in the proof of Theorem 11, we only need to show that

$$\mathbb{P}\left(\max_{(i,j) \in A \setminus E_0} M_{ij} \geq 4 \log p - \log \log p + q_\alpha\right) \to \alpha,$$

where the set $A$ is defined in the proof of Theorem 7 and

$$E_0 = \{(i, j) : 1 \leq i \leq j \leq p, \sigma_{ij1} \neq \sigma_{ij2}\} \cup \{(i, i) : 1 \leq i \leq p\}.$$ 

Since $s_0(p) = o(p)$, we have $\text{Card}(E_0) = o(p^2)$. The rest proof follows exactly the same as that of Theorem 7.

Proposition 3 assumes Conditions (C1) and (C3). As in Proposition 1, it can be shown that $\mathbb{P}(\hat{\Psi}^* \neq \Psi) \leq -\log(1 - \alpha) + o(1)$ under Condition (C2) (or (C2∗)) only, without assuming (C1) or (C3).

3.3.2 Testing rows of two covariance matrices

As discussed in the introduction, the standard methods for gene selection are based on the comparisons of the means and thus lack the ability to select genes that change their relationships with other genes. It is of significant practical interest to develop methods for gene selection which capture the changes in the gene’s dependence structure.
Several methods have been proposed in the literature. It was noted in Ho, et al. (2008) that the changes of variances are biologically interesting and are associated with changes in coexpression patterns in different biological states. Ho, et al. (2008) proposed to test $H_{0i}: \sigma_{ii1} = \sigma_{ii2}$ and select the $i$-th gene if $H_{0i}$ is rejected. Hu, et al. (2009) and Hu, et al. (2010) introduced methods which are based on simultaneous testing of the equality of the joint distributions of each row between two sample correlation matrices/covariance distance matrices.

The covariance provides a natural measure of the association between two genes and it can also reflect the changes of variances. Motivated by these applications, in this section we consider testing the equality of two covariance matrices row by row. Let $\sigma_{i1}$ and $\sigma_{i2}$ be the $i$-th row of $\Sigma_1$ and $\Sigma_2$ respectively. We consider testing simultaneously the hypotheses

$$H_{0i}: \sigma_{i1} = \sigma_{i2}, \ 1 \leq i \leq p.$$ 

We shall use the family-wise error rate (FWER) to measure the type I errors for the $p$ tests. The support estimate $\hat{\Psi}(4)$ defined in (3.3.2) can be used to test the hypotheses $H_{0i}$ by rejecting $H_{0i}$ if the $i$-th row of $\hat{\Psi}(4)$ is nonzero. Suppose that (C1) and (C2) (or (C2*)) hold. Then it can be shown easily that for this test, the FWER $\rightarrow 0$.

A different test is needed to simultaneously test the hypotheses $H_{0i}, 1 \leq i \leq p$, at a pre-specified FWER level $\alpha$ for some $0 < \alpha < 1$. Define

$$M_i = \max_{1 \leq j \leq p, j \neq i} M_{ij}.$$ 

82
The null distribution of $M_i$ can be derived similarly under the same conditions as in Theorem 7.

**Theorem 13.** Suppose the null hypothesis $H_{0i}$ holds. Then under the conditions in Theorem 7, for any given $x \in \mathbb{R}$, as $n, p \to \infty$,

$$P\left(M_i - 2 \log p + \log \log p \leq x\right) \to \exp \left(- \frac{1}{\sqrt{\pi}} \exp \left(- \frac{x^2}{2}\right)\right). \quad (3.3.4)$$

Based on the limiting null distribution of $M_i$ given in (3.3.4), we introduce the following test for testing a single hypothesis $H_{0i}$,

$$\Phi_{i, \alpha} = I(M_i \geq \alpha_p \text{ or } M_{ii} \geq 2 \log p) \quad (3.3.5)$$

with $\alpha_p = 4 \log p - \log \log p + q_\alpha$, where $q_\alpha$ is given in (3.1.4).

$H_{0i}$ is rejected and the $i$-th gene is selected if $\Phi_{i, \alpha} = 1$. It can be shown that the tests $\Phi_{i, \alpha}$, $1 \leq i \leq p$ together controls the overall FWER at level $\alpha$ asymptotically.

**Theorem 14.** Let $F = \{i : \sigma_{i, 1} \neq \sigma_{i, 2}, 1 \leq i \leq p\}$ and suppose Card($F$) = $o(p)$. Then under the conditions in Theorem 7,

$$\text{FWER} = P\left(\max_{i \in F} \Phi_{i, \alpha} = 1\right) \to \alpha.$$ 

Proof of Theorems 13 and 14: The proofs of Theorems 13 and 14 are similar to that of Theorem 7. Note that by Lemmas 10 and 11,

$$\left|P\left(\max_{i \in F} \Phi_{i, \alpha} = 1\right) - P\left(\max_{(i,j) \in A \setminus E_1} M_{ij} \geq \alpha_p\right)\right| \leq CC\text{ard}(F)p \times p^{-2} + o(1) = o(1),$$

where $E_1$ is defined by

$$E_1 = \{(i, j) : i \in F, i \leq j \leq p\} \cup \{(i, i) : 1 \leq i \leq p\}. $$
By replacing $A$, $A_0$ and $B_0$ in the proof of Theorem 7 with $A \setminus E_1$, $A_0 \setminus E_1$ and $B_0 \setminus E_1$ respectively, the rest proof of Theorem 14 follows exactly the same as that of Theorem 7. The proof of Theorem 13 is much simpler than that of Theorem 7. Due to the close similarity, the proof is omitted here.

Theorem 14 requires Conditions (C1) and (C3). However, similar to Proposition 1, it can be shown that the FWER can still be effectively controlled without (C1) and (C3) in the sense that $\text{FWER} \leq -\log(1 - \alpha) + o(1)$ under Condition (C2) (or (C2*)) only.

### 3.4 Numerical results

In this section, we turn to the numerical performance of the proposed methods. The goal is to first investigate the numerical performance of the test $\Phi_\alpha$ and the support recovery procedure through simulation studies and then illustrate our methods by applying them to the analysis of a prostate cancer dataset. The test $\Phi_\alpha$ is compared with four other tests, the likelihood ratio test (LRT), the tests given in Schott (2007) and Li and Chen (2012) which are both based on an unbiased estimator of $\|\Sigma_1 - \Sigma_2\|_F^2$, and the test proposed in Srivastava and Yanagihara (2010) which is based on a measure of distance by $\text{tr}(\Sigma_1^2)/(\text{tr}(\Sigma_1))^2 - \text{tr}(\Sigma_2^2)/(\text{tr}(\Sigma_2))^2$. The LRT is not well defined when $p > n$, so the comparison with the LRT is made only for the case $p \leq n$.

We first introduce the matrix models used in the simulations. Let $D = (d_{ij})$ be
a diagonal matrix with diagonal elements $d_{ii} = \text{Unif}(0.5, 2.5)$ for $i = 1, ..., p$. Denote by $\lambda_{\min}(A)$ the minimum eigenvalue of a symmetric matrix $A$. The following four models under the null, $\Sigma_1 = \Sigma_2 = \Sigma^{(i)}$, $i = 1, 2, 3$ and 4, are used to study the size of the tests.

- Model 1: $\Sigma^{*(1)} = (\sigma^{*(1)}_{ij})$ where $\sigma^{*(1)}_{ii} = 1$, $\sigma^{*(1)}_{ij} = 0.5$ for $5(k-1) + 1 \leq i \neq j \leq 5k$, where $k = 1, ..., [p/5]$ and $\sigma^{*(1)}_{ij} = 0$ otherwise. $\Sigma^{(1)} = D^{1/2} \Sigma^{*(1)} D^{1/2}$.

- Model 2: $\Sigma^{*(2)} = (\sigma^{*(2)}_{ij})$ where $\omega^{*(2)}_{ij} = 0.5 \left| i - j \right|$ for $1 \leq i, j \leq p$. $\Sigma^{(2)} = D^{1/2} \Sigma^{*(2)} D^{1/2}$.

- Model 3: $\Sigma^{*(3)} = (\sigma^{*(3)}_{ij})$ where $\sigma^{*(3)}_{ii} = 1$, $\sigma^{*(3)}_{ij} = 0.5 \ast \text{Bernoulli}(1, 0.05)$ for $i < j$ and $\sigma^{*(3)}_{ji} = \sigma^{*(3)}_{ij}$. $\Sigma^{(3)} = D^{1/2} (\Sigma^{*(3)} + \delta I) / (1 + \delta) D^{1/2}$ with $\delta = |\lambda_{\min}(\Sigma^{*(3)})| + 0.05$.

- Model 4: $\Sigma^{(4)} = O \Delta O$, where $O = \text{diag}(\omega_1, ..., \omega_p)$ and $\omega_1, ..., \omega_p \overset{iid}{\sim} \text{Unif}(1, 5)$ and $\Delta = (a_{ij})$ and $a_{ij} = (-1)^{i+j}0.4 \left| i - j \right|^{1/10}$. This model was used in Srivastava and Yanagihara (2009).

To evaluate the power of the tests, let $U = (u_{kl})$ be a matrix with 8 random nonzero entries. The locations of 4 nonzero entries are selected randomly from the upper triangle of $U$, each with a magnitude generated from $\text{Unif}(0, 4) \ast \max_{1 \leq j \leq p} \sigma_{jj}$. The other 4 nonzero entries in the lower triangle are determined by symmetry. We use the following four pairs of covariance matrices $(\Sigma^{(i)}_1, \Sigma^{(i)}_2)$, $i = 1, 2, 3$ and 4, to
compare the power of the tests, where $\Sigma_1^{(i)} = \Sigma^{(i)} + \delta I$ and $\Sigma_2^{(i)} = \Sigma^{(i)} + U + \delta I$, with $\delta = |\min\{\lambda_{\min}(\Sigma^{(i)} + U), \lambda_{\min}(\Sigma^{(i)})\}| + 0.05$.

The sample sizes are taken to be $n_1 = n_2 = n$ with $n = 60$ and $100$, while the dimension $p$ varies over the values $50, 100, 200, 400$ and $800$. For each model, data are generated from multivariate normal distributions with mean zero and covariance matrices $\Sigma_1$ and $\Sigma_2$. The nominal significant level for all the tests is set at $\alpha = 0.05$. The actual sizes and powers in percents for the four models, reported in Table 3.1, are estimated from 5000 replications.

It can be seen from Table 3.1 that the estimated sizes of our test $\Phi_\alpha$ are close to the nominal level $0.05$ in all the cases. This reflects the fact that the null distribution of the test statistic $M_n$ is well approximated by its asymptotic distribution. For Models 1-3, the estimated sizes of the tests in Schott (2007) and Li and Chen (2012) are also close to $0.05$. But both tests suffer from the size distortion for Model 4, the actual sizes are around 0.10 for both tests. The likelihood ratio test has serious size distortion (all is equal to 1). Srivastava and Yanagihara (2010)'s test has actual sizes close to the nominal significance level in all the cases.

The power results in Table 3.1 show that the proposed test has much higher power than the other tests in all settings. The number of nonzero off-diagonal entries of $\Sigma_1 - \Sigma_2$ does not change when $p$ grows, so the estimated powers of all tests tend to decrease when the dimension $p$ increases. It can be seen in Table 3.1 that the powers of Schott (2007), Li and Chen (2012) and Srivastava and Yanagihara (2010)'s
tests decrease extremely fast as \( p \) grows. However, the power of the proposed test \( \Phi_\alpha \) remains high even when \( p = 800 \), especially in the case of \( n = 100 \). Overall, for the sparse models, the new test significantly outperforms all the other three tests.

We also carried out simulations for non-Gaussian distributions including Gamma distribution, \( t \) distribution and normal distribution contaminated with exponential distribution. Similar phenomena as those in the Gaussian case are observed. For reasons of space, these simulation results are given in Appendix B.

### 3.4.1 Support recovery

We now consider the simulation results on recovering the support of \( \Sigma_1 - \Sigma_2 \) in the first three models with \( D = I \) and the fourth model with \( O = I \) under the normal distribution. For \( i = 1, 2, 3 \) and 4, let \( U^{(i)} \) be a matrix with 50 random nonzero entries, each with a magnitude of 2 and let \( \Sigma_1^{(i)} = (\Sigma^{(i)} + \delta I)/(1 + \delta) \) and \( \Sigma_2^{(i)} = (\Sigma^{(i)} + U^{(i)} + \delta I)/(1 + \delta) \) with \( \delta = |\min(\lambda_{\text{min}}(\Sigma^{(i)} + U^{(i)}), \lambda_{\text{min}}(\Sigma^{(i)}))| + 0.05 \). After normalization, the nonzero elements of \( \Sigma_2^{(i)} - \Sigma_1^{(i)} \) have magnitude between 0.74 and 0.86 for \( i = 1, 2, 3 \) and 4 in our generated models.

The accuracy of the support recovery is evaluated by the similarity measure \( s(\hat{\Psi}, \Psi) \) defined by

\[
s(\hat{\Psi}, \Psi) = \frac{|\hat{\Psi} \cap \Psi|}{\sqrt{|\hat{\Psi}| \cdot |\Psi|}},
\]

where \( \hat{\Psi} = \hat{\Psi}(4) \) and \( \Psi \) is the support of \( \Sigma_1^{(i)} - \Sigma_2^{(i)} \) and \(|\cdot|\) denotes the cardinality. Note that \( 0 \leq s(\hat{\Psi}, \Psi) \leq 1 \) with \( s(\hat{\Psi}, \Psi) = 0 \) indicating disjointness and \( s(\hat{\Psi}, \Psi) = 1 \)
<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>100.0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>4.6</td>
<td>5.0</td>
<td>6.1</td>
</tr>
<tr>
<td>Φₐ</td>
<td>5.5</td>
<td>5.4</td>
<td>5.0</td>
<td>5.3</td>
</tr>
<tr>
<td>likelihood</td>
<td>6.0</td>
<td>5.5</td>
<td>5.6</td>
<td>5.9</td>
</tr>
<tr>
<td>Schott</td>
<td>6.5</td>
<td>6.1</td>
<td>5.4</td>
<td>4.8</td>
</tr>
<tr>
<td>Li-Chen</td>
<td>6.6</td>
<td>6.0</td>
<td>5.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Srivastava</td>
<td>6.6</td>
<td>5.2</td>
<td>5.3</td>
<td>4.9</td>
</tr>
<tr>
<td>100</td>
<td>100.0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>4.8</td>
<td>4.1</td>
<td>4.4</td>
<td>4.9</td>
</tr>
<tr>
<td>Φₐ</td>
<td>4.5</td>
<td>4.2</td>
<td>4.3</td>
<td>5.0</td>
</tr>
<tr>
<td>likelihood</td>
<td>4.9</td>
<td>4.6</td>
<td>4.5</td>
<td>4.3</td>
</tr>
<tr>
<td>Schott</td>
<td>5.5</td>
<td>5.0</td>
<td>4.9</td>
<td>5.0</td>
</tr>
<tr>
<td>Li-Chen</td>
<td>5.5</td>
<td>5.0</td>
<td>5.2</td>
<td>5.2</td>
</tr>
<tr>
<td>Srivastava</td>
<td>4.9</td>
<td>5.1</td>
<td>5.0</td>
<td>5.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Empirical power</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>87.9</td>
<td>74.1</td>
<td>47.6</td>
<td>40.4</td>
</tr>
<tr>
<td>Φₐ</td>
<td>86.7</td>
<td>56.2</td>
<td>60.9</td>
<td>46.3</td>
</tr>
<tr>
<td>Schott</td>
<td>25.8</td>
<td>20.3</td>
<td>17.6</td>
<td>9.3</td>
</tr>
<tr>
<td>Li-Chen</td>
<td>8.9</td>
<td>8.3</td>
<td>6.7</td>
<td>5.2</td>
</tr>
<tr>
<td>Srivastava</td>
<td>7.8</td>
<td>6.7</td>
<td>5.1</td>
<td>3.7</td>
</tr>
<tr>
<td>100</td>
<td>99.9</td>
<td>99.4</td>
<td>94.1</td>
<td>91.6</td>
</tr>
<tr>
<td>Φₐ</td>
<td>99.9</td>
<td>99.6</td>
<td>98.1</td>
<td>95.9</td>
</tr>
<tr>
<td>Schott</td>
<td>93.7</td>
<td>90.0</td>
<td>95.0</td>
<td>92.9</td>
</tr>
<tr>
<td>Li-Chen</td>
<td>12.9</td>
<td>14.1</td>
<td>12.8</td>
<td>8.9</td>
</tr>
<tr>
<td>Srivastava</td>
<td>14.3</td>
<td>12.8</td>
<td>10.2</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Table 3.1: $N(0, 1)$ random variables. Model 1-4 Empirical sizes and powers in percents. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications.
indicating exact recovery. We summarize the average (standard deviation) values of $s(\hat{\Psi}, \Psi)$ in percents for all models with $n = 100$ in Table 3.2 based on 100 replications. The values are close to one, and hence the supports are well recovered by our procedure.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 50$</td>
<td>97.4(2.5)</td>
<td>89.8(4.3)</td>
<td>91.3(4.1)</td>
<td>93.0(3.8)</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>95.3(3.1)</td>
<td>86.8(5.1)</td>
<td>86.2(5.3)</td>
<td>89.7(4.4)</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>78.8(5.6)</td>
<td>80.1(5.9)</td>
<td>82.4(5.8)</td>
<td>84.0(6.4)</td>
</tr>
<tr>
<td>$p = 400$</td>
<td>75.6(6.5)</td>
<td>73.2(7.7)</td>
<td>79.4(6.1)</td>
<td>77.8(7.0)</td>
</tr>
<tr>
<td>$p = 800$</td>
<td>64.0(7.5)</td>
<td>61.8(8.8)</td>
<td>72.9(6.1)</td>
<td>68.1(7.6)</td>
</tr>
</tbody>
</table>

Table 3.2: Average (standard deviation) of $s(\hat{\Psi}, \Psi)$ in percents.

To better illustrate the elementwise recovery performance, heat maps of the nonzeros identified out of 100 replications for $p = 50$ and 100 are shown in Figures 1 and 2. These heat maps suggest that, in all models, the estimated support by $\hat{\Psi}(4)$ is close to the true support.

### 3.4.2 Real data analysis

We now illustrate our methods by applying them to the analysis of a prostate cancer dataset (Singh et al. (2002)) which is available at http://www.broad.mit.edu/cgi-
Figure 3.1: Heat maps of the frequency of the 0s identified for each entry of $\Sigma_1 - \Sigma_2$ ($n = 100$, $p = 50$ for the top row and $p = 100$ for the second row) out of 100 replications. White indicates 1000s identified out of 100 runs; black, 0/100.

The dataset consists of two classes of gene expression data that came from 52 prostate tumor patients and 50 prostate normal patients. This dataset has been analyzed in several papers on classification in which the two covariance matrices are assumed to be equal; see, for example, Xu, Brock and Parrish (2009). The equality of the two covariance matrices is an important assumption for the validity of these classification methods. It is thus interesting to test whether such an assumption is valid.

There are a total of 12600 genes. To control the computational costs, only the 5000 genes with the largest absolute values of the $t$-statistics are used. Let $\Sigma_1$ and $\Sigma_2$ be respectively the covariance matrices of these 5000 genes in tumor and normal samples. We apply the test $\Phi_\alpha$ defined in (3.1.3) to test the hypotheses.
$H_0 : \Sigma_1 = \Sigma_2$ versus $H_1 : \Sigma_1 \neq \Sigma_2$. Based on the asymptotic distribution of
the test statistic $M_n$, the $p$-value is calculated to be 0.0058 and the null hypothesis
$H_0 : \Sigma_1 = \Sigma_2$ is thus rejected at commonly used significant levels such as $\alpha = 0.05$
or $\alpha = 0.01$. Based on this test result, it is therefore not reasonable to assume
$\Sigma_1 = \Sigma_2$ in applying a classifier to this dataset.

We then apply three different methods to select genes with changes in variances/covariances between the two classes. The first is the differential variability analysis (Ho, et al., 2008) which chooses the genes with different variances between
two classes. In our procedure the $i$-th gene is selected if $M_{ii} \geq 2 \log p$. As a result,21 genes are selected. The second and third methods are based on the differential
covariance analysis, which is similar to the differential covariance distance vector
analysis in Hu, et al. (2010), but replacing the covariance distance matrix in their
paper by the covariance matrix. The second method selects the $i$-th gene if the $i$-th
row of $\hat{\Psi}(4)$ is nonzero. This leads to the selection of 43 genes. The third method,
which tests the covariance matrices row by row and is defined in (3.3.5), controls
the family-wise error rate at $\alpha = 0.1$, and is able to find 52 genes. As expected,
the gene selection based on the covariances could be more powerful than the one
that is based only on the variances. The tests provide valuable information as these
identified genes can be selected for a follow-up study.

Finally, we apply the support recovery procedure $\hat{\Psi}(4)$ to $\Sigma_1 - \Sigma_2$. For a visual
comparison between $\Sigma_1$ and $\Sigma_2$, Figure 3.2 plots the heat map of $\Sigma_1 - \Sigma_2$ of the
200 largest absolute values of two sample $t$ statistics. It can be seen from Figure 3.2 that the estimated support of $\Sigma_1 - \Sigma_2$ is quite sparse.

![Figure 3.2: Heat map of the selected genes by exactly recovery.](image)

### 3.5 Technical Lemmas

In this section, we collect some technical lemmas that are used in the proofs of the main results. The proofs of Lemmas 10 and 11 are given in Appendix B.

The first lemma is the classical Bonferroni’s inequality.

**Lemma 8 (Bonferroni inequality).** Let $B = \cup_{t=1}^p B_t$. For any $k < [p/2]$, we have

$$\sum_{t=1}^{2k} (-1)^{t-1} E_t \leq P(B) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} E_t,$$

where $E_t = \sum_{1 \leq i_1 < \cdots < i_t \leq p} P(B_{i_1} \cap \cdots \cap B_{i_t})$.

The second lemma comes from Berman (1962).
Lemma 9. [Berman (1962)] If $X$ and $Y$ have a bivariate normal distribution with expectation zero, unit variance and correlation coefficient $\rho$, then
\[
\lim_{c \to \infty} \frac{\Pr(X > c, Y > c)}{\sqrt{2\pi(1 - \rho)^{1/2}c^2}} \exp\left(-\frac{c^2}{1 + \rho}\right)(1 + \rho)^{1/2} = 1,
\]
uniformly for all $\rho$ such that $|\rho| \leq \delta$, for any $\delta$, $0 < \delta < 1$.

The next lemma is on the large deviations for $\hat{\theta}_{ij}$ and $\hat{\theta}_{ij}$.

Lemma 10. Under the conditions of (C2) or (C2*), there exists some constant $C > 0$ such that
\[
\Pr\left(\max_{i,j} |\hat{\theta}_{ij} - \theta_{ij}|/\sigma_{ii}\sigma_{jj} \geq C \frac{\varepsilon_n}{\log p}\right) = O(p^{-1} + n^{-\epsilon/8}), \quad \text{(3.5.1)}
\]
and
\[
\Pr\left(\max_{i,j} |\hat{\theta}_{ij} - \theta_{ij}|/\sigma_{ii}\sigma_{jj} \geq C \frac{\varepsilon_n}{\log p}\right) = O(p^{-1} + n^{-\epsilon/8}), \quad \text{(3.5.2)}
\]
where $\varepsilon_n = \max((\log p)^{1/6}/n^{1/2}, (\log p)^{-1}) \to 0$ as $n, p \to \infty$.

Define
\[
\tilde{\Sigma}_1 = (\tilde{\sigma}_{ij})_{p \times p} = \frac{1}{n_1} \sum_{k=1}^{n_1} (X - \mu_1)(X - \mu_1)^T,
\]
\[
\tilde{\Sigma}_2 = (\tilde{\sigma}_{ij})_{p \times p} = \frac{1}{n_2} \sum_{k=1}^{n_2} (Y - \mu_2)(Y - \mu_2)^T.
\]

Let $\Lambda$ be any subset of $\{(i, j) : 1 \leq i \leq j \leq p\}$ and $|\Lambda| = \text{Card}(\Lambda)$.

Lemma 11. Under the conditions of (C2) or (C2*), we have for some constant $C > 0$ that
\[
\Pr\left(\max_{(i,j) \in \Lambda} \frac{(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2} - \sigma_{ij1} + \sigma_{ij2})^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2} \geq x^2\right) \leq C|\Lambda|(1 - \Phi(x)) + O(p^{-1} + n^{-\epsilon/8})
\]
uniformly for $0 \leq x \leq (8 \log p)^{1/2}$ and $\Lambda \subseteq \{(i, j) : 1 \leq i \leq j \leq p\}$. 

Chapter 4

Adaptive Confidence Intervals For Regression Functions Under Shape Constraints

4.1 Benchmark and Lower Bound on Expected Length

The focus in this chapter is the construction of confidence intervals which have expected length that adapts to the unknown function. The evaluation of these procedures depends on lower bounds which are given here in terms of a local modulus of continuity first introduced by Cai and Low (2011) in the context of point estimation of convex functions under mean squared error loss. These lower bounds provide a
natural benchmark for our problems.

### 4.1.1 Benchmark and Lower Bound

We focus in this thesis on estimating the function $f$ at 0 since estimation at other points away from the boundary is similar. For a given function class $\mathcal{F}$, write $\mathcal{I}_\alpha(\mathcal{F})$ for the collection of all confidence intervals which cover $f(0)$ with guaranteed coverage probability of $1 - \alpha$ for all functions in $\mathcal{F}$. For a given confidence interval $CI$, denote by $L(CI)$ the length of $CI$ and $L(CI, f) = E_f(L(CI))$ the expected length of $CI$ at a given function $f$. The minimum expected length at $f$ of all confidence intervals with guaranteed coverage probability of $1 - \alpha$ over $\mathcal{F}$ is then given by

$$L^*_\alpha(f, \mathcal{F}) = \inf_{CI \in \mathcal{I}_\alpha(\mathcal{F})} L(CI, f). \tag{4.1.1}$$

A natural goal is to construct a confidence interval with expected length close to the minimum $L^*_\alpha(f, \mathcal{F})$ for every $f \in \mathcal{F}$ while maintaining the coverage probability over $\mathcal{F}$. However although $L^*_\alpha(f, \mathcal{F})$ is a natural benchmark for the expected length of confidence intervals it is not easy to evaluate exactly. Instead as a first step towards our goal, we provide a lower bound for the benchmark $L^*_\alpha(f, \mathcal{F})$ in terms of a local modulus of continuity $\omega(\epsilon, f, \mathcal{F})$ introduced by Cai and Low (2011). The local modulus is a quantity that is more easily computable and techniques for its analysis are similar to those given in Donoho and Liu (1991) and Donoho (1994) where a global modulus of continuity was introduced in the study of minimax theory.
for estimating linear functionals. See the examples in Section 4.1.2.

For a parameter space $\mathcal{F}$ and function $f \in \mathcal{F}$, the local modulus of continuity is defined by

$$\omega(\varepsilon, f, \mathcal{F}) = \sup \{ |g(0) - f(0)| : \|g - f\|_2 \leq \varepsilon, \ g \in \mathcal{F} \}. \quad (4.1.2)$$

where $\| \cdot \|_2$ is the $L_2(-\frac{1}{2}, \frac{1}{2})$ function norm. The following theorem gives a lower bound for the minimum expected length $L^*_\alpha(f, \mathcal{F})$ in terms of the local modulus of continuity $\omega(\varepsilon, f, \mathcal{F})$. In this theorem and throughout this chapter we write $\Phi$ for the cumulative distribution function and $\phi$ for the density function of a standard normal density and set $z_\alpha = \Phi^{-1}(1 - \alpha)$.

Theorem 15. Suppose $\mathcal{F}$ is a nonempty convex set. Let $0 < \alpha < \frac{1}{2}$ and $f \in \mathcal{F}$. Then for confidence intervals based on (1.3.1),

$$L^*_\alpha(f, \mathcal{F}) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha \sigma^2}} + \frac{\phi(z_\alpha)}{z_\alpha} - \alpha) \omega(\sqrt{\frac{z_\alpha}{n}}, f, \mathcal{F}). \quad (4.1.3)$$

In particular,

$$L^*_\alpha(f, \mathcal{F}) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha \sigma^2}}) \omega(\sqrt{\frac{z_\alpha}{n}}, f, \mathcal{F}). \quad (4.1.4)$$

Proof of Theorem 15: Suppose that $X \sim N(\theta, \sigma^2)$ where it is known that $\theta \in [0, a\sigma]$.

The confidence interval for $\theta$ which has guaranteed coverage over the interval $\theta \in [0, a\sigma]$ and which minimizes the expected length when $\theta = 0$ is given by

$$[0, \max(0, \min(X + z_\alpha \sigma, a\sigma))] \quad (4.1.5)$$
It follows that

\[ L = \sigma \int_{-z_\alpha}^{a-z_\alpha} z \phi(z) dz + \sigma (z_\alpha P(-z_\alpha \leq Z \leq a - z_\alpha) + aP(Z \geq a - z_\alpha)) \]  

(4.1.6)

and hence

\[ \frac{L}{\sigma} = (\phi(z_\alpha) - \phi(a - z_\alpha)) + z_\alpha (\Phi(a - z_\alpha) - \Phi(-z_\alpha)) + a(1 - \Phi(a - z_\alpha)) \]  

(4.1.7)

In particular when \( a = z_\alpha \)

\[ \frac{L}{\sigma} \geq z_\alpha (1 - \frac{\phi(0)}{z_\alpha} + \frac{\phi(z_\alpha)}{z_\alpha} - a) \]  

(4.1.8)

In particular we have

\[ \frac{L}{\sigma} \geq z_\alpha (1 - \frac{\phi(0)}{z_\alpha}) \]  

(4.1.9)

Write \( L^*_\alpha(f, \mathcal{F}) \) for the smallest expected length at \( f \) when we have guaranteed coverage over \( \mathcal{F} \). In particular let \( P_\theta \) be a subfamily of \( \mathcal{F} \) then \( L^*_\alpha(f, \mathcal{F}) \geq L^*_\alpha(f, P_\theta) \).

Now suppose that \( f_0 \) is the "true" function. Fix \( E > 0 \). There is a function \( f_1 \in \mathcal{F} \) such that

\[ |f_1(0) - f_0(0)| = \omega(\frac{\epsilon}{\sqrt{n}}, f, \mathcal{F}) \]

and such that

\[ \|f_1 - f_0\|_2 = \frac{\epsilon}{\sqrt{n}}. \]

Now for \( 0 \leq \theta \leq 1 \), let \( f_\theta = f_0 + \theta(f_1 - f_0) \). Let \( P_\theta \) be this collection of \( f_\theta \). Now for the process

\[ dY(t) = f_\theta(t) dt + \frac{1}{\sqrt{n}} dW(t), \quad -\frac{1}{2} \leq t \leq \frac{1}{2} \]
there is a sufficient statistic \( S_n \) given by
\[
S_n = f_0(0) + (f_1(0) - f_0(0)) \frac{1}{\int (f_1 - f_0)^2} \int (f_1(t) - f_0(t))(dY(t) - f_0(t)dt).
\]
Note that \( S_n \) has a normal distribution \( S_n \sim N(f_\theta(0), \frac{1}{n} \int (f_1 - f_0)^2) \) or more specifically \( S_n \sim N(f_\theta(0), \frac{1}{n} \omega^2(\frac{1}{\sqrt{n}}, f_0, \mathcal{F})) \).

Note that \( a = \epsilon \). Now take \( \epsilon = z_\alpha \). It then follows that
\[
L^*_\alpha(f_0, P_\theta) \geq \omega \left( \frac{z_\alpha}{\sqrt{n}}, f_0, \mathcal{F} \right) (1 - \frac{\phi(0)}{z_\alpha} + \frac{\phi(z_\alpha)}{z_\alpha} - \alpha). \quad \square
\]

The lower bounds given in Theorem 15 can be viewed as benchmarks for the evaluation of the expected length of confidence intervals when the true function is \( f \) for confidence intervals which have guaranteed coverage probability over all of \( \mathcal{F} \). The bound depends on the underlying true function \( f \) as well as the parameter space \( \mathcal{F} \).

The bounds from Theorem 15 are general. In some settings they can be used to rule out the possibility of adaptation, whereas in other settings they provide bounds on how much adaptation is possible. In particular the result ruling out adaptation over Lipschitz classes mentioned in the introduction easily follows from this theorem. For example, consider the Lipschitz class \( \Lambda(\beta, M) \) and suppose that \( f \) is in the interior of \( \Lambda(\beta, M) \). Straightforward calculations similar to those given in Section 4.1.2 show that
\[
\omega(E, f, \Lambda(\beta, M)) \sim C E^{\frac{2\beta}{2\beta+1}}. \quad (4.1.10)
\]
Now consider two Lipschitz classes $\Lambda(\beta_1, M_1)$ and $\Lambda(\beta_2, M_2)$ with $\beta_1 > \beta_2$. A fully adaptive confidence interval in this setting would have guaranteed coverage of $1 - \alpha$ over $\Lambda(\beta_1, M_1) \cup \Lambda(\beta_2, M_2)$ and maximum expected length over $\Lambda(\beta_i, M_i)$ of order $n^{\frac{\beta_i}{2\beta_i+1}}$ for $i = 1$ and 2. However, it follows from Theorem 15 and (4.1.10) that for all confidence intervals with coverage probability of $1 - \alpha$ over $\Lambda(\beta_2, M_2)$, for every $f \in \Lambda(\beta_2, M')$ with $M' < M_2$,

$$L_\alpha^*(f, \Lambda(\beta_2, M_2)) \geq C(\alpha)n^{-\frac{\beta_2}{2\beta_2+1}}$$

for some constant $C(\alpha)$ not depending on $f$. In particular this holds for all $f \in \Lambda(\beta_1, M_1) \cap \Lambda(\beta_2, M')$ and hence

$$\sup_{f \in \Lambda(\beta_1, M_1)} \inf_{CI \in \mathcal{L}_\alpha(\Lambda(\beta_1, M_1) \cup \Lambda(\beta_2, M_2))} L(CI, f) \geq C(\alpha)n^{-\frac{\beta_2}{2\beta_2+1}} \gg n^{-\frac{\beta_1}{2\beta_1+1}}.$$ 

Therefore it is not possible to have confidence intervals with adaptive expected length over two Lipschitz classes with different smoothness parameters.

In this thesis, Theorem 15 will be used to provide benchmarks in the setting of shape constraints. Denote by $F_m$ and $F_c$ respectively the collection of all monotonically nondecreasing functions and the collection of all convex functions on $[-\frac{1}{2}, \frac{1}{2}]$. We shall now show that in these cases the modulus and the associated lower bounds vary significantly from function to function.
4.1.2 Examples of Bounds For Monotone Functions and Convex Functions

We now turn to the application of the lower bound given in Theorem 15 in the case of monotone functions and convex functions. Here we shall evaluate the lower bound for four particular families of functions yielding different rates at which the expected length decreases to zero as the noise level decreases in contrast to the situation just described where the parameter space did not have an order constraint. Two of the functions will be both monotonically nondecreasing and convex. In this case the lower bound can also be quite different depending on whether we assume the knowledge that \( f \) is convex or monotonically nondecreasing.

The key quantity that is needed in any application of Theorem 15 is the local modulus. We follow the same approach as given in Donoho (1994) where a global modulus of continuity is considered for minimax estimation. In each case, for a given function \( f \), we first minimize the \( L_2 \) norm between a function \( g \in \mathcal{F} \) and the function \( f \) subject to the constraint that \( |g(0) - f(0)| = a \) for some given value \( a > 0 \). From here it is easy to invert and thus maximize \( |g(0) - f(0)| \) given a constraint on the \( L_2 \) norm between \( f \) and \( g \).

Example 1. As a first example consider the linear function \( f_k(t) = kt \) where \( k \geq 0 \) is a constant. This function is both monotonically nondecreasing and convex.

First consider the collection of monotonically nondecreasing functions \( F_m \). We shall treat separately the case \( k > 0 \) and the case \( k = 0 \). For the moment we shall
take $k > 0$. Suppose that $0 < a \leq \frac{k}{2}$. In this case $f_k \in F_m$ and a function $g$ that
minimizes $\|g - f_k\|_2$ subject to the constraint that $|g(0) - f_k(0)| = a$ is given by
$g(t) = f_k(t)$ if $t < 0$, $g(t) = a$ if $0 \leq t \leq b$, and $g(t) = f_k(t)$ if $t > b$, where $b$
satisfies $f_k(b) = a$. The assumption that $a \leq \frac{k}{2}$ guarantees $b \leq \frac{1}{2}$. We then have
$\|g - f_k\|_2 = \frac{a}{2}/(3k)^{\frac{1}{2}}$. It follows that if $\epsilon^2 \leq \frac{1}{24}k^2$
$$
\omega(\epsilon, f_k, F_m) = (3k)^{\frac{1}{2}}\epsilon^{\frac{1}{2}}
$$
and consequently for $n \geq \frac{24\epsilon^2}{k^2}$, if $k > 0$
$$
L^*_\alpha(f_k, F_m) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(3k)^{\frac{1}{2}}z_\alpha n^{-\frac{1}{2}}.
$$
In the case that $k = 0$ a function $g$ that minimizes $\|g - f_0\|_2$ subject to the
constraint that $|g(0) - f_0(0)| = a$ is given by $g(t) = f_0(t)$ if $t < 0$, $g(t) = a$ if $0 \leq t \leq \frac{1}{2}$. In this case it is easy to check that $\|g - f_0\|_2 = \frac{1}{\sqrt{2}}\epsilon$ and hence
$$
\omega(\epsilon, f_0, F_m) = \sqrt{2}\epsilon
$$
and hence
$$
L^*_\alpha(f_0, F_m) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})\sqrt{2}z_\alpha n^{-\frac{1}{2}}.
$$
We now consider the bound for the length of the confidence interval for $f_k$
belonging to the collection of convex functions. In this case we do not need to treat
the cases $k > 0$ and $k = 0$ separately. The function $g$ that minimize $\|g - f_k\|_2$ subject
to the constraint that $g$ is convex and $|g(0) - f_k(0)| = a$ is given by $g(t) = (k + 3a)t - a$
if $t \geq 0$ and $g(t) = (k - 3a)t - a$ if $t < 0$. In this case $\|g - f\|_2 = \frac{1}{2}a$. It then
immediately follows that
\[ \omega(\epsilon, f_k, F_c) = 2\epsilon \]
and so
\[ L^*_\alpha(f_k, F_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})2z_\alpha n^{-\frac{1}{2}}. \]

It is important to note that for \( k > 0 \) the minimum expected lengths \( L^*_\alpha(f_k, F_m) \) and \( L^*_\alpha(f_k, F_c) \) are different, one of order \( n^{-\frac{1}{3}} \) and another of order \( n^{-\frac{1}{2}} \), although the function \( f_k \) is the same. It is also interesting to note that the expected length of the confidence for monotone functions is an increasing function of \( k \) whereas the expected length of the confidence for convex functions does not depend on \( k \). Since we shall show that these bounds are achievable within a constant factor it follows that the minimum expected length of the confidence interval when \( f_k \) is the true function depends strongly on whether we specify that the underlying collection of functions is convex or monotone. Plots illustrating shapes of functions \( f_k \) and a least favorable function \( g \) are shown as below in Fig 4.1.

Example 2. As a second example which is also both monotonically nondecreasing and convex consider the function \( f(t) = k_1 t + k_2 t^r I(0 < t \leq \frac{1}{2}) \) where \( r \geq 1 \) and \( k_1 \geq 0 \) and \( k_2 > 0 \) are constants.

We consider the cases \( r = 1 \) and \( r > 1 \) separately. When \( r = 1 \) the function is piecewise linear with the change of slope at 0. In this case suppose \( 0 < a \leq \frac{k_1 + k_2}{2} \). A monotonically nondecreasing function \( g \in F_m \) that minimize \( \|g - f\|_2 \) subject to the constraint that \( |g(0) - f(0)| = a \) is given by \( g(t) = f(t) \) if \( t < 0 \), \( g(t) = a \) if \( 0 \leq t \leq b \),
and $g(t) = f(t)$ if $t > b$, where $b$ satisfies $f(b) = a$. The constraint $a \leq \frac{k_1 + k_2}{2}$ is to guarantee that such a $b$ exists with $b \leq \frac{1}{2}$. Then we have $\|g - f\|_2 = a^\frac{3}{2} (3(k_1 + k_2))^{-\frac{3}{2}}$ and it follows that if $\epsilon^2 \leq \frac{1}{27}(k_1 + k_2)^2$

$$\omega(\epsilon, f, F_m) = (3(k_1 + k_2))^{\frac{3}{4}} \epsilon^2$$

and consequently for $n \geq \frac{24\varepsilon^2}{(k_1 + k_2)^2}$:

$$L_\alpha^*(f, F_m) \geq (1 - \frac{1}{\sqrt{2\pi \varepsilon \alpha}})(3(k_1 + k_2))^{\frac{3}{4}} \varepsilon^2 n^{-\frac{1}{2}}.$$

We can also give a lower bound on the expected length for this same function for confidence intervals which guarantee coverage over the class of convex functions. Suppose $0 < a \leq \frac{k_2}{4}$. Here we need to find the convex $h$ that minimizes $\|h - f\|_2$ subject to the constraints that $|h(0) - f(0)| = a$. It is given by $h(t) = f(t)$ if $t \leq -\frac{2a}{k_2}$, $h(t) = (\frac{k_2}{2} + k_1)t + a$ if $-\frac{2a}{k_2} \leq t \leq \frac{2a}{k_2}$ and $h(t) = f(t)$ if $t \geq \frac{2a}{k_2}$. Then
\[ \|f - g\|_2 = 2a^{\frac{2}{3}}/(3k_2)^{\frac{1}{3}} \] and it follows that if \( \epsilon^2 \leq \frac{k_2^2}{48} \),

\[ \omega(\epsilon, f, F_c) = (3k_2/4)^{\frac{1}{3}} \epsilon^{\frac{2}{3}}. \]

Hence, for \( n \geq \frac{48\epsilon^2}{k_2^2} \),

\[ L_\alpha^*(f, F_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(3k_2/4)^{\frac{1}{3}} z_\alpha^{-\frac{2}{3}} n^{-\frac{1}{3}}. \]

We now turn to the case where \( r > 1 \). Suppose \( 0 < a \leq \frac{k_1^2}{2} + k_2(\frac{1}{2})^r \). In this case the monotonically nondecreasing function \( g \) that minimizes \( \|g - f\|_2 \) subject to the constraints that \( |g(0) - f(0)| = a \) is given by \( g(t) = f(t) \) if \( t < 0 \), \( g(t) = a \) if \( 0 \leq t \leq b \), and \( g(t) = f(t) \) if \( t > b \), where \( b \) satisfies \( f(b) = a \). As before the condition \( 0 < a \leq \frac{k_1^2}{2} + k_2(\frac{1}{2})^r \) guarantees that \( b \) exists with \( b < \frac{1}{2} \). In this case

\[ a^{\frac{2}{3}}(3k_1)^{-\frac{1}{3}} - ca^s \leq \|g - f\|_2 \leq a^{\frac{2}{3}}(3k_1)^{-\frac{1}{3}} + ca^s \] for some constant \( c > 0 \) and \( s > 3/2 \).

It follows that if \( \epsilon^2 \leq \frac{1}{2r} k_1^2 + (1 + \frac{1}{2r+1} - \frac{2}{r+1})(\frac{1}{2})^{2r+1} k_2^2 + (\frac{1}{2} - \frac{1}{r+1} + \frac{1}{r+2})(\frac{1}{2})^{r+1} k_1 k_2 \), then

\[ \omega(\epsilon, f, F_m) = (3k_1)^{\frac{1}{3}} \epsilon^{\frac{2}{3}} (1 + o(1)). \]

Hence,

\[ L_\alpha^*(f, F_m) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(3k_1)^{\frac{1}{3}} z_\alpha^{-\frac{2}{3}} n^{-\frac{1}{3}} (1 + o(1)). \]

For a bound on the expected length of this same function for confidence intervals with coverage guaranteed over the collection of convex functions we suppose \( 0 < a \leq k_2(\frac{1}{2})^{r+1} \). In this case the convex function \( h \) that minimizes \( \|h - f\|_2 \) subject to the constraints that \( |h(0) - f(0)| = a \), is given by \( h(t) = kt + a \), \( k > k_1 \), if \( x_0 \leq t \leq x_1 \) and \( h(t) = f(t) \) otherwise, where \( (x_0, cx_0) \) and \( (x_1, cx_1 + x_1^r) \) are the
intersection points of \( f(t) \) and the line \( kt + a \). Then the function \( h \) with slope \( k_0 \) that minimize \( \| h - f \|_2 \) would be the least favorable function. It follows that, if
\[
\epsilon^2 \leq \frac{k_0^2}{24} \left( \frac{1}{2} \right)^{2r},
\]
\[
\omega(\epsilon, f, F_c) = C(r) \frac{1}{k_0^{2r+1}} \epsilon^{2r+1}
\]
and consequently for \( n \geq \frac{24\epsilon^2 2^{2r}}{k_0^2} \),
\[
L^*_\alpha(f, F_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}}) C(r) \frac{1}{k_0^{2r+1}} \frac{2r}{z_\alpha^{2r+1}} n \frac{r}{2r+1},
\]
where \( C(r) > 0 \) is a constant depending on \( r \) only.

It is interesting that in this example the rates of convergence for \( L^*_\alpha(f, F_m) \) and \( L^*_\alpha(f, F_c) \) are the same for the case \( r = 1 \), and are different when \( r > 1 \). Plots illustrating shapes of functions \( f \) and a least favorable function \( g \) are shown as below in Fig 4.2.

Next we consider a function which is monotonically nondecreasing but not convex.

**Example 3.** Let \( f(t) = kt^r \) for some constant \( k > 0 \) and \( r = 2l + 1 \) or \( r = \frac{1}{2l+1} \) for \( l = 0, 1, 2, \cdots \). Suppose that \( a < \left( \frac{1}{2} \right)^r k \). In this case a function \( g \) that minimizes \( \| g - f \|_2 \) subject to the constraint that \( | g(0) - f(0) | = a \) is given by \( g(t) = f(t) \) if \( t < 0 \), \( g(t) = a \) if \( 0 \leq t \leq b \), and \( g(t) = f(t) \) if \( t > b \), where \( b \) satisfies \( f(b) = a \). As before the condition \( a < \left( \frac{1}{2} \right)^r k \) guarantees that \( b \) exists with \( b < \frac{1}{2} \). Then \( \| g - f \|_2 = a^{1+\frac{1}{2r}}/k^{\frac{1}{2r}}(2r^2/(r+1)(2r+1))^{\frac{1}{2r}} \) and it follows that when \( \epsilon^2 \leq \left( \frac{1}{2} \right)^{2r+1} k^2 \frac{2r^2}{(r+1)(2r+1)}, \)
\[
\omega(\epsilon, f, F_m) = \left( \frac{(r+1)(2r+1)k}{2r^2} \right)^{\frac{1}{2r+1}} \epsilon^{\frac{2r}{2r+1}}.
\]
Figure 4.2: Plots of $f$ and a least favorable function $g$ in Example 2 with the constraints $|g(0) - f(0)| = a$. 
Hence for $n \geq \frac{2^{2r+1}(r+1)(2r+1)z^2}{2^{r+1}k^2}$,

$$L^*_\alpha(f, \mathcal{F}_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(\frac{(r + 1)(2r + 1)k^2}{2^{r+1}z_\alpha^{2r+1}})^{\frac{r}{2r+1}} n^{-\frac{r}{2r+1}}$$

and once again it is clear that the rate at which the expected length decreases to zero depends strongly on the value of $r$.

As a final example we consider a function which is convex but not monotonically nondecreasing.

**Example 4.** Let $f(t) = t^2$. Suppose that $a < 1/2$. In this case the function $g$ that minimizes $\|g - f\|_2$ subject to the constraint that $|g(0) - f(0)| = a$ is given by $g(t) = -3\sqrt{a/2}t - a$ if $-\sqrt{2a} \leq t \leq 0$, $g(t) = 3\sqrt{a/2}t - a$ if $0 \leq t \leq \sqrt{2a}$, and $g(t) = f(t)$ otherwise. Then $\|g - f\|_2 = 2^{5/4}/\sqrt{15}a^{5/4}$ and it follows that when $\epsilon^2 \leq 1/\sqrt{15}$,

$$\omega(\epsilon, f, \mathcal{F}_m) = \frac{15^{2/5}}{2} \epsilon^{\frac{4}{5}}.$$ 

Hence for $n \geq \sqrt{15}z^2_\alpha$,

$$L^*_\alpha(f, \mathcal{F}_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}}) \frac{15^{2/5}}{2} z_\alpha n^{-\frac{2}{5}}$$

A similar minimization problem is solved in Dümbgen (2003).

Plots illustrating shapes of functions $f$ and a least favorable function $g$ for both Examples 3 and 4 are shown as below in Fig 4.3.
Figure 4.3: Plots of $f$ and a least favorable function $g$ in Examples 3 and 4 with the constraints $|g(0) - f(0)| = a$.

4.2 Confidence Procedures

In this section we both construct and give an analysis of adaptive confidence intervals for monotone functions and convex functions. The procedures are easily implementable. We consider the class of monotonically nondecreasing functions and the class of convex functions. Concave functions and monotonically nonincreasing functions can be handled similarly.

4.2.1 Construction

The construction is split into two steps. In the first step a countable collection of confidence intervals is created each of which has guaranteed coverage probability. These intervals are based on a collection of pairs of linear estimators. For each
interval one of the estimators has nonnegative bias and the other nonpositive bias. The one sided control of the bias of these estimators is a key special feature in these problems and an important part of what makes it possible to adapt to every individual function. Moreover for each function \( f \) this collection has at least one interval with expected length within a constant factor of the local modulus bound given in Theorem 15. The second step is to select from this collection a particular interval.

In the case of monotonically nondecreasing functions we take for each \( j \geq 2 \), pairs of estimators \( \delta^R_j = 2^j(Y(2^{-j}) - Y(0)) \) and \( \delta^L_j = 2^j(Y(0) - Y(-2^{-j})) \). Then for estimating \( f(0) \) it is easy to check that \( \delta^R_j \) has nonnegative and monotonically nonincreasing biases while \( \delta^L_j \) have nonpositive and monotonically nondecreasing biases. The one sided control of the biases of these estimators over the class of all monotonically nondecreasing functions easily allows for the construction of a confidence interval. For that we shall need the standard deviation of \( \delta^R_j \) and \( \delta^L_j \). In order to give a unified treatment in both the monotone and convex case it is useful to establish a common notation. Here we shall set \( \sigma^2_j = \frac{2^{j-1}}{n} \). It is then easy to check that both \( \delta^R_j \) and \( \delta^L_j \) have a standard deviation of \( \sqrt{2}\sigma_j \). It is then also easy to see that for each \( j \geq 2 \), the confidence interval \( CI^m_j(\alpha) \) given by

\[
CI^m_j(\alpha) = [\delta^L_j - z_{\alpha/2}\sqrt{2}\sigma_j; \delta^R_j + z_{\alpha/2}\sqrt{2}\sigma_j]
\]  

(4.2.1)

has guaranteed coverage of \( 1 - \alpha \). We should however note that in (4.2.1) the left endpoint of the interval may be larger than the right end point in which case we
adopt the convention that the confidence interval is just the empty set. The length of this confidence interval is then \( \max(\delta_j^R - \delta_j^L + 2\sqrt{2}z_{\alpha/2}\sigma_j, 0) \).

In the case of convex functions for \( j \geq 1 \) let \( \delta_j = 2^{j-1}(Y(2^{-j}) - Y(-2^{-j})) \) and let \( \tilde{\delta}_j = 2\delta_{j+1} - \delta_j \).

The following Lemma shows that for convex functions \( \delta_j \) have nonnegative and monotonically nonincreasing biases and that \( \tilde{\delta}_j \) have nonpositive and monotonically nondecreasing biases.

**Lemma 12.** For any convex function \( f \),

\[
0 \leq \text{Bias}(\delta_{j+1}) \leq \frac{1}{2} \text{Bias}(\delta_j) \tag{4.2.2}
\]

\[
E \delta_j - 3E \delta_{j+1} + 2E \delta_{j+2} \geq 0. \tag{4.2.3}
\]

It is also easy to check that the standard deviation of \( \delta_j \) is equal to \( \sigma_j \) where \( \sigma_j^2 = \frac{2^{j-1}}{n} \) and that \( 2\delta_{j+1} - \delta_j \) has a standard deviation of \( \sqrt{5}\sigma_j \). It then follows from the signs of the biases of \( \delta_{j+1} \) and \( 2\delta_{j+1} - \delta_j \) that for any given \( j \)

\[
CI_j^c(\alpha) = \left[ 2\delta_{j+1} - \delta_j - z_{\alpha/2}\sqrt{5}\sigma_j, \, \delta_{j+1} + z_{\alpha/2}\sigma_{j+1} \right] \tag{4.2.4}
\]
gives a confidence interval with coverage probability of at least \( 1 - \alpha \). We should also note once again that the left endpoint of the interval may be larger than the right end point in which case the confidence interval is taken to be the empty set and so in this case the length of this confidence interval is \( \max(\delta_j - \delta_{j+1} + (\sqrt{5} + \sqrt{2})z_{\alpha/2}\sigma_j, 0) \).

These results, for which a more formal proof is given in Section 2.5 are summarized in the following proposition.
Proposition 4. For every \( j \geq 2 \), the confidence interval \( CI_j^m \) defined in (4.2.1) has coverage probability of at least \( 1 - \alpha \) for all monotonically nondecreasing function \( f \in F_m \) and for every \( j \geq 1 \), the confidence interval \( CI_j^c \) defined in (4.2.4) has coverage probability of at least \( 1 - \alpha \) for all convex functions \( f \in F_c \).

Proof of Proposition 4: For monotone functions, we have

\[
P(f(0) \in CI_j^m) = P(\delta_j^L - z_{\frac{\alpha}{2}}\sqrt{2}\sigma_j \leq f(0) \leq \delta_j^R + z_{\frac{\alpha}{2}}\sqrt{2}\sigma_j)
\]

\[
\geq 1 - P(\delta_j^R < f(0) - z_{\frac{\alpha}{2}}\sqrt{2}\sigma_j) - P(\delta_j^L > f(0) + z_{\frac{\alpha}{2}}\sqrt{2}\sigma_j)
\]

\[
= 1 - P(Z < \frac{f(0) - E(\delta_j^R)}{\sqrt{2}\sigma_j} - z_{\frac{\alpha}{2}}) - P(Z > \frac{f(0) - E(\delta_j^L)}{\sqrt{2}\sigma_j} + z_{\frac{\alpha}{2}}),
\]

where \( Z \) is a standard normal random variable. Because \( f(0) - E(\delta_j^R) \leq 0 \) and \( f(0) - E(\delta_j^L) \geq 0 \), we have

\[
P(f(0) \in CI_j^m) \geq 1 - P(Z < -z_{\frac{\alpha}{2}}) - P(Z > z_{\frac{\alpha}{2}}) = 1 - \alpha.
\]

For convex functions, let \( b_j = \text{Bias}(\delta_j) \). It follows from Lemma 12 that \( b_j - 2b_{j+1} > 0 \), and hence we have

\[
P(f(0) \in CI_j^c) \geq P(2\delta_{j+1} - \delta_j - z_{\frac{\alpha}{2}}\sqrt{5}\sigma_j \leq f(0) \leq \delta_{j+1} + z_{\frac{\alpha}{2}}\sigma_{j+1})
\]

\[
\geq 1 - P(\delta_{j+1} < f(0) - z_{\frac{\alpha}{2}}\sigma_{j+1})
\]

\[
- P(2\delta_{j+1} - \delta_j > f(0) + z_{\frac{\alpha}{2}}\sqrt{5}\sigma_j)
\]

\[
= 1 - P(\frac{\delta_{j+1} - E\delta_{j+1}}{\sigma_{j+1}} < \frac{b_{j+1}}{\sigma_{j+1}} - z_{\frac{\alpha}{2}})
\]

\[
- P(\frac{2\delta_{j+1} - \delta_j - E(2\delta_{j+1} - \delta_j)}{\sqrt{5}\sigma_j} > \frac{b_j - 2b_{j+1}}{\sqrt{5}\sigma_j} + z_{\frac{\alpha}{2}})
\]

\[
\geq 1 - P(Z < -z_{\frac{\alpha}{2}}) - P(Z > z_{\frac{\alpha}{2}})
\]

\[
= 1 - \alpha. \quad \square
\]
The second stage in the construction is that of selecting from these collections of intervals the one to be used. First note that one should not select the shortest interval since the collections defined in (4.2.1) and (4.2.4) will always contain one which corresponds to the empty set. A more sensible goal is to try to select the interval with the smallest expected length or at least one which has expected length close to the smallest expected length.

The approach we take here is to choose an interval for which the expected length is of the same order of magnitude as the standard deviation of the length. Such an interval will always have expected length close to the shortest expected length. For the case of monotonically nondecreasing functions the selection of the interval from the countable collection in (4.2.1) can be done by creating another collection of estimators which can be used to estimate the expected length of the intervals.

More specifically set \( \xi_j = 2^{j-1}(Y(2^{-j+1}) - Y(2^{-j})) - 2^{j-1}(Y(-2^{-j}) - Y(-2^{-j+1})) \). Then for \( j \geq 2 \), \( \xi_j \)'s are independent of each other and both \( \delta^R_j \) and \( \delta^L_j \) are independent of \( \xi_k \) for every \( k \leq j \). We should note that the estimators \( \xi_j \) are similar to \( \delta^R_j - \delta^L_j \) in that they are both differences of averages of \( Y \) to the left and right of the origin and thus estimate the average local change of the function. However \( \delta^R_j - \delta^L_j \) are not independent for different \( j \) whereas the \( \xi_j \) are independent. It is thus natural to view the \( \xi_j \) as a surrogate for \( \delta^R_j - \delta^L_j \) with the technical advantage that they are independent. The selection of a \( j \) for which \( \xi_j \) has expected value close to \( \sigma_j \) will then result in a confidence interval \( CI_j^m \) close to the one with the smallest
expected length. The independence properties of the \( \xi_j \) allows us to guarantee a \( 1 - \alpha \) coverage probability while making this selection.

More specifically the construction proceeds as follows. Let

\[
\hat{j} = \inf_j \left\{ j : \xi_j \leq \frac{3}{2} z_\alpha \sigma_j \right\} (4.2.5)
\]

and define the final confidence interval by

\[
CI^m_\star = CI^m_j(\alpha) (4.2.6)
\]

Before we turn to the analysis of this procedure we also introduce here a related confidence procedure in the convex case. Here rather than introducing an independent estimate of the difference between the two estimators used in constructing the confidence interval we proceed more directly. The basic idea is similar but the dependence between the estimates of \( j \) and the confidence interval constructed from this estimate requires that we adjust the original coverage level of our \( CI^c_j \).

More specifically let \( T_j = \delta_j - \delta_{j+1} \). When the expected value of \( T_j \) is the same order as \( \sigma_j \), the confidence interval \( CI^c_j \) will then be close to the one with the smallest expected length. Our estimate of \( j \) is given by an empirical version, namely

\[
\hat{j} = \inf_j \{ j : T_j \leq z_\alpha \sigma_j \} (4.2.7)
\]

Although this estimate can be used to select the appropriate \( CI^c_j \) to use, as just mentioned, care also needs to be taken to make sure that the resulting selected interval maintains the required coverage probability. The analysis given below shows
that a choice of $\alpha/6$ in the construction of the original collection of intervals guarantees an overall coverage probability of $\alpha$. Thus in the case of convex functions we define our interval by

$$CI^c_\star = CI^c_j\left(\frac{\alpha}{6}\right)$$

(4.2.8)

### 4.2.2 Analysis of the Confidence Intervals

In this section we present the properties of the confidence intervals $CI^m_\star$ and $CI^c_\star$ defined by (4.2.6) and (4.2.8) focusing on the coverage and the expected length of these intervals.

We begin with the confidence interval $CI^m_\star$. In this case it is easy to check the coverage probability of $CI^m_\star$ by the independence of the interval $CI^m_j$ and $\xi_k$ for every $k$ satisfying $2 \leq k \leq j$.

The key to the analysis of the expected length is the introduction of $j^m_\star$ where

$$j^m_\star = \arg\min_j \{ j : E\xi_j \leq z_\alpha \sigma_j \}.$$  

(4.2.9)

The analysis of the expected length relies on showing that $\hat{j}$ is highly concentrated around $j^m_\star$. The concentration of $\hat{j}$ around $j^m_\star$ then provides a bound on the expected length of $CI^\star$. These results, for which a proof is given in Section 2.5 are summarized in the following theorem.

**Theorem 16.** Let $0 < \alpha \leq 0.2$. The confidence interval $CI^m_\star$ defined in (4.2.6) has coverage probability of at least $1 - \alpha$ for all monotonically nondecreasing functions
\( f \in F_m \) and satisfies
\[
E_f L(\text{CI}_m^*) \leq 1.21(3z_\alpha + 2\sqrt{2}z_\alpha^2)\sigma_{j^*} \leq c_0 z_\alpha \sigma_{j^*}, \tag{4.2.10}
\]
where \( c_0 \) is a constant and can be taken to be 8.85 for all \( 0 < \alpha \leq 0.2 \).

Proof of Theorem 16: We shall first prove that the confidence interval \( \text{CI}_m^* \) has guaranteed coverage probability of \( 1 - \alpha \) over \( F_m \) and then prove the upper bound for the expected length.

Note that
\[
P(f(0) \in \text{CI}_m^*) = \sum_{j=2}^{\infty} P(f(0) \in \text{CI}_j^m | \hat{j} = j)P(\hat{j} = j).
\]
Because both \( \delta_{j}^R \) and \( \delta_{j}^L \) are independent of \( \xi_k \) for \( k \leq j \) and the event \( \{ \hat{j} = j \} \) depends only on \( \xi_k \) for \( k \leq j \), then by Proposition 4 we have
\[
P(f(0) \in \text{CI}_m^*) = \sum_{j=2}^{\infty} P(f(0) \in \text{CI}_j^m)P(\hat{j} = j) \geq \sum_{j=2}^{\infty} (1 - \alpha)P(\hat{j} = j) = 1 - \alpha.
\]

We now turn to the upper bound for the expected length. Note that for \( s \geq 0 \),
\[
E\xi_{j^m + s} \leq z_\alpha \sigma_{j^m + s}, \text{ and so we have}
\]
\[
P(\hat{j} \geq j^m + k) \leq \prod_{s=0}^{k-1} P(\xi_{j^m + s} < \frac{3}{2} z_\alpha \sigma_{j^m + s}) \leq \prod_{s=0}^{k-1} P(Z > z_\alpha(\frac{3}{2} - \frac{1}{2^s})).
\]
It follows from \( E(\delta_{j}^R - \delta_{j}^L) \leq 2E\xi_j \) that \( E(\delta_{j}^R - \delta_{j}^L) \leq 2E\xi_j \), and hence we have
\[
E_f L(\text{CI}^*) = E_f (\delta_{j}^R - \delta_{j}^L + 2\sqrt{2}z_\alpha \sigma_j) \leq E_f (2\xi_j + 2\sqrt{2}z_\alpha \sigma_j)
\leq E_f ((3z_\alpha + 2\sqrt{2}z_\alpha \sigma_j) \cdot P(\hat{j} = j)).
\]
Thus,

$$E_f L(CI^*) \leq (3z_\alpha + 2\sqrt{2}z_\alpha^2)\sigma_{j^m} \left( P(\hat{j} \leq j^m_\alpha) + \sum_{k=1}^{\infty} 2^{k/2}P(\hat{j} = j^m_\alpha + k) \right). \quad (4.2.11)$$

Set $w_k = 2^{k/2} - 2^{(k-1)/2}$ for $k \geq 1$. Then it is easy to see that

$$S = P(\hat{j} \leq j^m_\alpha) + \sum_{k=1}^{\infty} 2^{k/2}P(\hat{j} = j^m_\alpha + k) = 1 + \sum_{k=1}^{\infty} w_kP(\hat{j} \geq j^m_\alpha + k).$$

Thus,

$$S = 1 + \sum_{k=1}^{\infty} w_k \prod_{s=0}^{k-1} P(Z > z_\alpha \left( \frac{3}{2} - \frac{1}{2^s} \right)).$$

The right hand side is increasing in $\alpha$. Through numerical calculations, we can see that, for $\alpha = 0.2$,

$$\sum_{k=1}^{\infty} w_k \prod_{s=0}^{k-1} P(Z > z_\alpha \left( \frac{3}{2} - \frac{1}{2^s} \right)) \leq 0.21.$$ 

Thus, by equation (4.2.11), we have

$$E_f L(CI^*) \leq 1.21(3z_\alpha + 2\sqrt{2}z_\alpha^2)\sigma_{j^m}.$$ 

**Remark 3.** The constant $c_0$ in Theorem 16 depends on the upper limit of $\alpha$. $c_0$ can be smaller if the upper limit on $\alpha$ is reduced. For example, for common choices of $\alpha = 0.05$ or $0.01$, $c_0 \leq 7.71$ for $\alpha = 0.05$, and $c_0 \leq 7.42$ for $\alpha = 0.01$.

Theorem 16 shows that the coverage probability is attained and also provides an upper bound on the expected length in terms of $\sigma_{j^m}$. In order to to establish that this expected length is within a constant factor of the lower bound given in Theorem 15, we need to provide a lower bound for $L^*_\alpha(f, F_m)$ in terms of $z_\alpha \sigma_{j^m}$. This connection is given in the following theorem.
Theorem 17. Let $0 < \alpha \leq 0.2$ and let $f \in F_m$. Then

$$L^*_\alpha(f, F_m) \geq (1 - \frac{1}{\sqrt{2\pi \alpha}}) \frac{1}{\sqrt{2}} z_\alpha \sigma_j^{\alpha}. \quad (4.2.12)$$

Proof of Theorem 17: Note that if $j_*^{m} > 2$, then $E\xi_{j_*^{m}-1} \geq z_\alpha \sigma_{j_*^{m}-1} = \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}}$ and hence there is a $t_* \leq 2^{-j_*^{m}+2}$ such that we have either $f(t_*) - f(0) \geq \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}}$ or $f(0) - f(-t_*) \geq \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}}$. If $f(t_*) \geq \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}} + f(0)$, let

$$g(t) = \begin{cases} \max\{\frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}} + f(0), f(t)\}, & \text{if } t \geq 0; \\ f(t), & \text{otherwise,} \end{cases}$$

and if $f(-t_*) \leq -\frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}} + f(0)$, let

$$g(t) = \begin{cases} \min\{-\frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}} + f(0), f(t)\}, & \text{if } t \leq 0; \\ f(t), & \text{otherwise.} \end{cases}$$

Then we have

$$\int_{-1/2}^{1/2} (f(t) - g(t))^2 dt \leq \frac{1}{2} z_\alpha^2 2^{j_*^{m}-1} \cdot 2^{-j_*^{m}+2} = \frac{z_\alpha^2}{n}.$$ 

If $j_*^{m} = 2$, let

$$g(t) = \begin{cases} \max\{\frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}} + f(0), f(t)\}, & \text{if } t \geq 0; \\ f(t), & \text{otherwise,} \end{cases}$$

then we have

$$\int_{-1/2}^{1/2} (f(t) - g(t))^2 dt \leq \frac{1}{2} z_\alpha^2 2^{j_*^{m}-1} \cdot \frac{1}{2} = \frac{z_\alpha^2}{n}.$$ 

It then follows that

$$\omega\left(\frac{z_\alpha}{\sqrt{n}}, f, F_m\right) \geq \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_*^{m}}.$$
and so
\[ L_\alpha^*(f, F_m) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}}) \frac{1}{\sqrt{2}} z_\alpha \sigma_{j_m^*}. \]

Combining Theorems 16 and 17, we have
\[ E_f L(CT^m) \leq c_1 L_\alpha^*(f, F_m) \] (4.2.13)

for all monotonically nondecreasing function \( f \in F_m \), where \( c_1 \) is a constant depending on \( \alpha \) only. For example, \( c_1 \) can be taken to be 14.40 for \( \alpha = 0.05 \) and 12.67 for \( \alpha = 0.01 \). Hence, the confidence interval \( CI^m \) is uniformly within a constant factor of the benchmark \( L_\alpha^*(f, F_m) \) for all monotonically nondecreasing functions \( f \) and all confidence level \( 1 - \alpha \geq 0.8 \).

We now turn to an analysis of the properties of the confidence interval \( CI^c \)

defined in (4.2.8). The key to this analysis is the introduction of \( j^c \) where
\[ j^c = \arg \min_j \left\{ j : ET_j \leq \frac{2}{3} z_\alpha \sigma_j \right\}. \] (4.2.14)

The analysis of both the coverage probability and the expected length relies on showing that \( \hat{j} \) is highly concentrated around \( j^c \). The probability of not covering \( f(0) \) can be bounded by
\[ P(f(0) \notin CI^c) \leq P(\hat{j} \leq j^c - 3) + P(\hat{j} \geq j^c + 3) + \sum_{l=-2}^{2} P(f(0) \notin CI_{j^c+l}). \] (4.2.15)

The first two terms are controlled by the high concentration of \( \hat{j} \) around \( j^c \) and the last term is controlled by Proposition 4 which bounds the coverage probability of any given \( j \). The concentration of \( \hat{j} \) around \( j^c \) also allows control on the expected length of \( CI^c \) which leads to the following theorem.
Theorem 18. Let $0 < \alpha \leq 0.2$. The confidence interval $CI_c^*$ defined in (4.2.8) has coverage probability of at least $1 - \alpha$ for all convex $f$ and satisfies

$$E_f L(CI_c^*) \leq 1.25(z_\alpha + (\sqrt{5} + \sqrt{2})z_{\frac{\alpha}{2}})\sigma_{j^*} \leq c_0 z_\alpha \sigma_{j^*} \tag{4.2.16}$$

where $c_0$ is a constant and can be taken to be 12.79 for all $0 < \alpha \leq 0.2$.

Proof of Theorem 18: We shall first prove that the confidence interval $CI_c^*$ has guaranteed coverage probability of $1 - \alpha$ over $F_c$ and then prove the upper bound for the expected length.

Note that if $j^*_c > 1$, then $ET_{j^*_c-1} \geq 2^{k-1/2} z_\alpha \sigma_{j^*_c-1} = \frac{\sqrt{2}}{3} z_\alpha \sigma_{j^*_c}$. It follows that for $k \geq 1$, $ET_{j^*_c-k} \geq 2^{k-1/2} \frac{2}{3} z_\alpha \sigma_{j^*_c-k} = 2^{(3k-1)/2} \frac{2}{3} z_\alpha \sigma_{j^*_c-k}$. Hence

$$P(\hat{j} = j^*_c - k) \leq P(T_{j^*_c-k} \leq z_\alpha \sigma_{j^*_c-k}) \leq P(Z \geq \frac{2^{(3k-1)/2}}{3} - 1)z_\alpha) \tag{4.2.17}$$

Also for $m \geq 0$, $ET_{j^*_c+m} \leq 2^{-m} \cdot \frac{2}{3} z_\alpha \sigma_{j^*_c} = 2^{-m/2} \cdot \frac{2}{3} z_\alpha \sigma_{j^*_c+m}$ and hence

$$P(\hat{j} \geq j^*_c + k) \leq \prod_{m=0}^{k-1} P(T_{j^*_c+m} > z_\alpha \sigma_{j^*_c+m}) \leq \prod_{m=0}^{k-1} P(Z > z_\alpha (1 - \frac{2}{3} 2^{-m/2})). \tag{4.2.18}$$

To bound the coverage probability note that

$$P(f(0) \notin CI_c^*) \leq \sum_{m=3}^{2} P(\hat{j} = j^*_c - m) + P(\hat{j} \geq j^*_c + 3) + \sum_{k=-2}^{2} P(f(0) \notin CI_{j^*_c+k}). \tag{4.2.19}$$

It then follows from Equation (4.2.17) that

$$P(\hat{j} = j^*_c - 3) \leq P(Z \geq \frac{13}{3} z_\alpha) \leq \frac{7\alpha}{10000}$$

for all $0 < \alpha \leq 0.2$. It is easy to verify directly that for all $z \geq 1$, $P(Z \geq 2z) \leq (1/6)P(Z \geq z)$. Furthermore, it is easy to see that for $k \geq 1$, $\frac{2^{(3k+3)-1)/2}}{3} - 1 \geq 2^{k+\frac{13}{3}}$.
and so
\[
P(\hat{j} = j^c - 3 - k) \leq P(Z \geq \frac{2^{3(k+3)-1/2}}{3} - 1) \alpha \leq P(Z \geq 2^k \frac{13}{3} z_\alpha)
\leq 6^{-k} P(Z \geq \frac{13}{3} z_\alpha) \leq 6^{-k} \frac{7\alpha}{10000}.
\]

Hence,
\[
\sum_{m \geq 3} P(\hat{j} = j^c - m) = \sum_{k \geq 0} P(\hat{j} = j^c - 3 - k) \leq \frac{7\alpha}{10000} \sum_{k \geq 0} 6^{-k} \leq \frac{7\alpha}{5000}.
\]

Note that (4.2.18) yields that
\[
P(\hat{j} \geq j^c + 3) \leq P(Z \geq \frac{1}{3} z_\alpha) \cdot P(Z \geq (1 - \frac{1}{3\sqrt{2}}) z_\alpha) \cdot P(Z \geq \frac{11}{12} z_\alpha) \leq \frac{\alpha}{6.4}
\]
for all $0 < \alpha \leq 0.3$. It now follows from (4.2.19) that
\[
P(f(0) \in CI^c) = 1 - P(f(0) \notin CI^c) \geq 1 - \left(\frac{7\alpha}{5000} + \frac{\alpha}{6.4} + 5 \times \frac{\alpha}{6}\right) \geq 1 - \alpha. \quad \Box
\]

We now turn to the upper bound for the expected length. Note that
\[
E_f L(CI^c) \leq \sum_{j=1}^{\infty} (z_\alpha + (\sqrt{5} + \sqrt{2}) z_{\frac{\alpha}{4}}) \sigma_j \cdot P(\hat{j} = j)
\]
(4.2.20)

Hence
\[
E_f L(CI^c) \leq (z_\alpha + (\sqrt{5} + \sqrt{2}) z_{\frac{\alpha}{4}}) \sigma_{j^c} \left( P(\hat{j} \leq j^c) + \sum_{k=1}^{\infty} 2^{k/2} P(\hat{j} = j^c + k) \right)
\]
(4.2.21)

Set $w_k = 2^{k/2} - 2^{(k-1)/2}$ for $k \geq 1$. Then it is easy to see that
\[
S = P(\hat{j} \leq j^c) + \sum_{k=1}^{\infty} 2^{k/2} P(\hat{j} = j^c + k) = 1 + \sum_{k=1}^{\infty} w_k P(\hat{j} \geq j^c + k).
\]

It then follows from (4.2.18) that
\[
S \leq 1 + \sum_{k=1}^{\infty} w_k \prod_{m=0}^{k-1} P(Z > z_\alpha(1 - \frac{2}{3^{2m/2}}))
\]

120
The right hand side is clearly increasing in $\alpha$. Direct numerical calculations show that for $\alpha = 0.2$
\[
\sum_{k=1}^{\infty} w_k \prod_{m=0}^{k-1} P(Z > z_\alpha (1 - \frac{2}{3} \frac{1}{2^{3m/2}})) \leq 0.25.
\]
It then follows directly from (4.2.21) that
\[EL(CI^c_s) \leq 1.25(z_\alpha + (\sqrt{5} + \sqrt{2})z_\alpha)\sigma_{j^c}. \]

Remark 4. The constant $c_0$ in Theorem 18 depends on the upper limit of $\alpha$. $c_0$ can be smaller if the upper limit on $\alpha$ is reduced. For example, for common choices of $\alpha = 0.05$ or 0.01, $c_0 \leq 8.57$ for $\alpha = 0.05$, and $c_0 \leq 7.42$ for $\alpha = 0.01$.

Theorem 18 shows that the coverage probability is attained and also provides an upper bound on the expected length in terms of $\sigma_{j^c}$. As was the case for monotone functions, in order to to establish that this expected length for convex functions is within a constant factor of the lower bound given in Theorem 15, we need to provide a lower bound for $L^*_\alpha(f, F_c)$ in terms of $z_\alpha \sigma_{j^c}$. This connection is given in the following theorem.

**Theorem 19.** Let $0 < \alpha \leq 0.2$ and let $f \in F_c$. Then
\[L^*_\alpha(f, F_c) \geq (1 - \frac{1}{\sqrt{2\pi} z_\alpha}) \frac{\sqrt{2}}{3} z_\alpha \sigma_{j^c}. \] (4.2.22)

Proof of Theorem 19: Note that if $j^c > 1$, then $ET_{j^c-1} \geq \frac{2}{3} z_\alpha \sigma_{j^c-1} = \frac{\sqrt{2}}{3} z_\alpha \sigma_{j^c}$ and hence there is a $t_*$ satisfying $0 < t_* \leq 2^{-j^c+1}$ such that $f_s(t_*) = \frac{\sqrt{2}}{3} z_\alpha \sigma_{j^c}$, where $f_s(t) = \frac{f(t) + f(-t)}{2} - f(0)$. Let $g$ be defined by
\[g(t) = f(t)1(|t| > t_*) + (f_s(t_*) + \frac{f(t_*) - f(-t_*)}{2t_*}t)1(|t| \leq t_*)\]
with \( g(0) = f_s(t_*) \) for which
\[
\int_{-1/2}^{1/2} (g(t) - f(t))^2 dt \leq \frac{9}{4} f_s^2(t_*) t_* \leq \frac{z_\alpha^2}{n}.
\]
If \( j_\alpha^* = 1 \), then let
\[
g(t) = f(t) + \frac{\sqrt{2}}{3} z_\alpha \sigma_{j_\alpha^*},
\]
then we have
\[
\int_{-1/2}^{1/2} (g(t) - f(t))^2 dt \leq \frac{2}{9} z_\alpha^2 \sigma_1^2 \leq \frac{z_\alpha^2}{n}.
\]
It then follows that
\[
\omega \left( \frac{z_\alpha}{\sqrt{n}}, f, F_c \right) \geq \frac{\sqrt{2}}{3} z_\alpha \sigma_{j_\alpha^*},
\]
and so
\[
L_{\alpha}^*(f, F_c) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}}) \frac{\sqrt{2}}{3} z_\alpha \sigma_{j_\alpha^*}. \quad \Box
\]

Theorems 18 and 19 together yield
\[
E_f L(CI_c^\alpha) \leq c_2 L_{\alpha}^*(f, F_c) \quad (4.2.23)
\]
for all convex function \( f \in F_c \), where \( c_2 \) is a constant depending on \( \alpha \) only. For example, \( c_2 \) can be taken to be 24 for \( \alpha = 0.05 \) and 19 for \( \alpha = 0.01 \). Hence, the confidence interval \( CI_c^\alpha \) is uniformly within a constant factor of the benchmark \( L_{\alpha}^*(f, F_c) \) for all convex functions \( f \) and all confidence level \( 1 - \alpha \geq 0.8 \).

### 4.3 Nonparametric Regression

We have so far focused on the white noise model. The theory presented in the earlier sections can also easily be extended to nonparametric regression. Consider
the regression model

\[ y_i = f(x_i) + \sigma z_i, \quad i = -n, -(n - 1), -1, 0, 1, \ldots, n \]  

(4.3.1)

where \( x_i = \frac{i}{2n} \) and \( z_i \sim iid \sim N(0, 1) \) and where for notational convenience we index the observations from \(-n\) to \(n\). Note that the noise level \( \sigma \) can be accurately estimated easily, as in Hall, et al (1990) or Munk, et al (2005). See also Wang, et al (2008). We shall thus assume it is known in this section. Then under the assumption that \( f \) is convex or monotone, we wish to provide a confidence interval for \( f(0) \).

### 4.3.1 Monotone Regression

Let \( J = \lfloor \log_2 n \rfloor \). For \( 1 \leq j \leq J \) define the local average estimators

\[
\bar{\delta}_j^R = 2^{-j+1} \sum_{k=1}^{2^{j-1}} y_k \quad \text{and} \quad \bar{\delta}_j^L = 2^{j+1} \sum_{k=1}^{2^{j-1}} y_{-k}.
\]

(4.3.2)

We should note that the indexing scheme is the reverse of that given for the white noise with drift process. Here estimators \( \bar{\delta}_j^R \) (or \( \bar{\delta}_j^L \)) with small values of \( j \) have smaller bias (or larger bias) and larger variance than those with larger values of \( j \).

As in the white noise model it is easy to check that \( \bar{\delta}_j^R \) has nonnegative bias and \( \bar{\delta}_j^L \) has nonpositive bias. Simple calculations show that the variance of \( \bar{\delta}_j^R \) and \( \bar{\delta}_j^L \) are both \( 2\sigma_j^2 \), where \( \sigma_j^2 = 2^{-j} \sigma^2 \). It is also important to introduce \( \bar{\xi}_j \) as in the white noise case, where

\[ \bar{\xi}_j = 2^{-j} \sum_{k=2j+1}^{2j} (y_k - y_{-k}) \].

It is easy to check that \( E \bar{\xi}_j \leq E \bar{\xi}_{j+1} \), \( \bar{\xi}_j \)'s are independent with each other and both \( \bar{\delta}_j^R \) and \( \bar{\delta}_j^L \) are independent with \( \bar{\xi}_k \) for every \( k \geq j \).
It then follows that $CI^m_j = [\bar{\delta}^L_j - z_{\bar{\alpha}}\sqrt{2}\sigma_j, \bar{\delta}^R_j + z_{\bar{\alpha}}\sqrt{2}\sigma_j]$ has guaranteed coverage probability of at least $1 - \alpha$ over all monotonically nondecreasing functions.

Now set

$$\hat{j} = \begin{cases} \max_j \{ j : \bar{\xi}_j \leq \frac{3}{2}z_{\alpha}\sigma_j \}, & \text{if } \bar{\xi}_1 \leq \frac{3}{2}z_{\alpha}\sigma_1; \\ 1, & \text{otherwise.} \end{cases}$$

and define the confidence interval to be

$$CI^*_m = CI^m_{\hat{j}}.$$  \hspace{1cm} (4.3.4)

The properties of this confidence interval can then be analyzed in the same way as before and can be shown to be similar to those for the white noise model. In particular, the following result holds.

**Theorem 20.** Let $0 < \alpha \leq 0.2$. The confidence interval $CI^*_m$ defined in (4.3.4) has coverage probability of at least $1 - \alpha$ for all monotone function $f$ and satisfies

$$E_f L(CI^*_m) \leq C_1 L^*_\alpha(f, F_m)$$  \hspace{1cm} (4.3.5)

for all monotonically nondecreasing function $f \in F_m$, where $C_1 > 0$ is a constant depending on $\alpha$ only.

### 4.3.2 Convex Regression

As in the monotone case, set $J = \lfloor \log_2 n \rfloor$. For $1 \leq j \leq J$ define the local average estimators

$$\bar{\delta}_j = 2^{-j} \sum_{k=1}^{2^{j-1}} (y_{-k} + y_k).$$  \hspace{1cm} (4.3.6)
We should note that this indexing scheme is the reverse of that given for the white noise with drift process. Here estimators \( \tilde{\delta}_j \) with small values of \( j \) have smaller bias and larger variance than those with larger values of \( j \).

As in the white noise model it is easy to check that \( \tilde{\delta}_j \) has nonnegative bias. It is also important to introduce an estimate which has a similar variance but is guaranteed to have nonpositive bias. The key step is to introduce

\[
T_j = \tilde{\delta}_j - \tilde{\delta}_{j-1},
\]

as an estimate of the bias of \( \tilde{\delta}_j \). The following lemma gives the required properties of \( \tilde{\delta}_j \) and \( T_j \).

**Lemma 13.** For any convex function \( f \),

\[
2ET_j \leq ET_{j+1}
\]

(4.3.8)

\[
0 \leq \text{Bias}(\tilde{\delta}_j) \leq \frac{2^{j-1} + 1}{2^j + 1}\text{Bias}(\tilde{\delta}_{j+1})
\]

(4.3.9)

From (4.3.9) it is clear that the biases of the estimators \( \tilde{\delta}_j \) are nonnegative and monotonically nondecreasing. In addition straightforward calculations using both (4.3.8) and (4.3.9) show that the estimators

\[
\delta_j^L = (2 + 2^{-(-j-1)})\tilde{\delta}_j - (1 + 2^{-(-j-1)})\tilde{\delta}_{j+1} = \tilde{\delta}_j - (1 + 2^{-(-j-1)})T_{j+1}
\]

have a nonpositive and monotonically nonincreasing biases. Simple calculations show that the variance of \( \delta_j^L \) is \( \tau_j^2 = (5 + 2^{-j+3} + 2^{-2j+2})2^{-j-1}\sigma^2 \).

It then follows that \( CI_j^c = [\tilde{\delta}_j - (1 + 2^{-(-j-1)})T_{j+1} - z_{\alpha/12}\tau_j, \tilde{\delta}_j + z_{\alpha/12}\sigma_j] \) has coverage over all convex functions.
Now set
\[ \hat{j} = \begin{cases} \max_j \{ j : T_j \leq z_0 \sigma_j \} , & \text{if } T_2 \leq z_0 \sigma_2 ; \\ 1 , & \text{otherwise.} \end{cases} \] (4.3.10)

and define the confidence interval to be
\[ CI_c^* = CI_j^c . \] (4.3.11)

This confidence interval shares similar properties as the one for the white noise model. In particular, the following result holds.

**Theorem 21.** Let \( 0 < \alpha \leq 0.2 \). The confidence interval \( CI_c^* \) defined in (4.3.11) has coverage probability of at least \( 1 - \alpha \) for all convex function \( f \) and satisfies
\[ E_f L(CI_c^*) \leq C_2 L_\alpha^c (f, F_c) \] (4.3.12)

for all convex function \( f \in F_c \), where \( C_2 > 0 \) is a constant depending on \( \alpha \) only.

We omit the proofs for Theorems 20 and 21 as they are analogous to those for the corresponding results in the white noise model.

### 4.4 Technical Lemmas

We prove the technical lemmas in this section.

#### 4.4.1 Proof of Lemma 12

Set \( f_s(t) = \frac{f(t) + f(-t)}{2} - f(0) \). Now note that \( f_s(tx) \) is convex in \( x \) for all \( 0 \leq t \leq 1 \).

Hence \( g(x) = \int_0^1 f_s(tx) dt \) is also convex with \( g(0) = 0 \). For \( x > 0 \) set \( z = xt \) and it
follows that \( g(x) = \frac{1}{x} \int_{0}^{x} f_s(z)dz = \frac{1}{x} \int_{-x}^{x} (f(z) - f(0))dz \). Equation (4.2.2) follows from the fact that \( g(x) \leq \frac{1}{x}g(2x) \) for \( x = 2^{-(j+1)} \) and equation (4.2.3) follow from the fact that \( g(2x) \leq 2/3g(x) + 1/3g(4x) \).

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]

\[ \bbox[1pt]{\begin{array}{c}
\end{array}} \]
Denote the bias of $\tilde{\delta}_j$ by $\bar{b}_j = E\tilde{\delta}_j - f(0)$. Then

$$\bar{b}_j = 2^{-(j-1)} \sum_{k=1}^{2^{j-1}} f_s\left(\frac{k}{n}\right) = 2^{-(j-1)} \left\{ \sum_{k=2^{j-2}+1}^{2^{j-1}} f_s\left(\frac{k}{n}\right) + \sum_{k=1}^{2^{j-2}} f_s\left(\frac{k}{n}\right) \right\}.$$  

It follows from (4.4.1) that for $k > 2^{j-2}$, $f_s\left(\frac{k}{n}\right) \geq \frac{k}{2^{j-2}} f_s\left(\frac{2^{j-2}}{n}\right)$, and for $k \leq 2^{j-2}$, $f_s\left(\frac{k}{n}\right) \leq \frac{k}{2^{j-2}} f_s\left(\frac{2^{j-2}}{n}\right)$. Hence

$$\sum_{k=2^{j-2}+1}^{2^{j-1}} f_s\left(\frac{k}{n}\right) \geq \sum_{k=2^{j-2}+1}^{2^{j-1}} \frac{k}{2^{j-2}} f_s\left(\frac{2^{j-2}}{n}\right) \geq \frac{\sum_{k=2^{j-2}+1}^{2^{j-1}} k}{\sum_{k=1}^{2^{j-2}} k} \sum_{k=1}^{2^{j-2}} f_s\left(\frac{k}{n}\right)$$

$$= \frac{3 \cdot 2^{j-2} + 1}{2^{j-2} + 1} \sum_{k=1}^{2^{j-2}} f_s\left(\frac{k}{n}\right).$$

Hence,

$$\bar{b}_j \geq 2^{-(j-1)} \cdot \left( \frac{3 \cdot 2^{j-2} + 1}{2^{j-2} + 1} + 1 \right) \sum_{k=1}^{2^{j-2}} f_s\left(\frac{k}{n}\right) = 2^{j-1} \frac{1}{2^{j-2} + 1} \bar{b}_{j-1}. \quad \Box$$

128
Chapter 5

Conclusion

The oracle test $\Phi_\alpha(\Omega)$ introduced in Chapter 2 performs well against sparse alternatives and the data-driven procedure $\Phi_\alpha(\hat{\Omega})$ requires a good estimate of the precision matrix $\Omega$. In this thesis we mainly focused on sparse precision matrices for which the CLIME estimator is known to perform well. The test $\Phi_\alpha(\hat{\Omega})$ can be used with a much wider range of covariance/precision matrices. As mentioned in Section 2.2.2, one only needs an estimate $\hat{\Omega}$ satisfying the $\ell_1$ condition (2.2.15) and then the result given in Theorem 4 extends directly. For example, when the covariance matrix $\Sigma$ is either sparse or bandable, Condition (2.2.15) can be achieved by inverting thresholding or tapering estimators of the covariance matrix $\Sigma$. The simulation results showed that the data-driven test $\Phi_\alpha(\hat{\Omega})$ performs well when $\Sigma$ is sparse. See Cai and Zhou (2011) for further details on estimating covariance matrices and their inverse under the matrix $\ell_1$ norm.
In Chapter 2 it is assumed that the two populations have the same covariance matrix. More generally, suppose we observe $X_k \overset{iid}{\sim} N_p(\mu_1, \Sigma_1)$, $k = 1, ..., n_1$, and $Y_k \overset{iid}{\sim} N_p(\mu_2, \Sigma_2)$, $k = 1, ..., n_2$ and wish to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. In order to apply the procedure proposed in this thesis, one needs to first test $H_0 : \Sigma_1 = \Sigma_2$ versus $H_1 : \Sigma_1 \neq \Sigma_2$. For this purpose, for example, the test introduced in Chapter 3 can be used. If the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$ is rejected, the test proposed in this paper is not directly applicable. However, a modified version of the procedure can still be used. Note that the covariance matrix of $\bar{X} - \bar{Y}$ is $\Sigma_1/n_1 + \Sigma_2/n_2$. To apply the test procedure in Section 2.1, one needs to estimate $(\Sigma_1 + \frac{n_1}{n_2} \Sigma_2)^{-1}$. When both $\Sigma_1$ and $\Sigma_2$ are sparse, the inverse can be estimated well by $(\hat{\Sigma}_{1,\text{thr}} + \frac{n_1}{n_2} \hat{\Sigma}_{2,\text{thr}})^{-1}$ using the adaptive thresholding estimators $\hat{\Sigma}_{1,\text{thr}}$ and $\hat{\Sigma}_{2,\text{thr}}$ introduced in Cai and Liu (2011). Similarly, when both $\Sigma_1$ and $\Sigma_2$ are bandable, $(\Sigma_1 + \frac{n_1}{n_2} \Sigma_2)^{-1}$ can also be estimated well. A more interesting problem is the estimation of $(\Sigma_1 + \frac{n_1}{n_2} \Sigma_2)^{-1}$ when the precision matrices $\Omega_1$ and $\Omega_2$ are sparse.

The asymptotic properties in Section 2.2.2 rely on the assumption that the locations of the nonzero entries of $\mu_1 - \mu_2$ are uniformly drawn from $\{1, ..., p\}$. When this assumption does not hold, the asymptotic power results may fail. A simple solution is to first apply a random permutation to the coordinates of $\mu_1 - \mu_2$ (and correspondingly the coordinates of $\bar{X} - \bar{Y}$) so that the nonzero locations are uniformly drawn from $\{1, ..., p\}$, and apply the testing procedures to the permuted
data and the results given in Section 2.2.2 then hold.

It is well known that the convergence rate in distribution of the extreme value type statistics is slow. There are several possible ways to improve the rate of convergence. See, for example, Hall (1991), Liu, Lin and Shao (2008) and Birnbaum and Nadler (2012). It is interesting to investigate whether these methods can be applied to improve the convergence rate of our test statistic. We leave this to future work.

In Chapter 3, we introduced a test for the equality of two covariance matrices which is proved to have the prespecified significance level asymptotically and to enjoy certain optimality in terms of its power. In particular, it was shown theoretically and numerically that the test is especially powerful against sparse alternatives. Support recovery and testing two covariance matrices row by row with applications to gene selection are also considered.

There are several possible extensions of our method. For example, an interesting direction is to generalize the procedure to testing the hypothesis of homogeneity of several covariance matrices, $H_0 : \Sigma_1 = \cdots = \Sigma_K$, where $K \geq 2$. A test statistic similar to $M_n$ can be constructed to test this hypothesis and analogous theoretical results can be developed. We shall report the details of the results elsewhere in the future as a significant amount of additional work is still needed.

Much recent attention has focused on the estimation of large covariance and precision matrices. The current work here is related to the estimation of covariance
matrices. An adaptive thresholding estimator of sparse covariance matrices was introduced in Cai and Liu (2011). The procedure is based on the standardized statistics $\hat{\sigma}_{ij}/\hat{\theta}_{ij}^{1/2}$, which is closely related to $M_{ij}$. In this thesis, the asymptotic distribution of $M_n = \max_{1 \leq i \leq j \leq p} M_{ij}$ is obtained. It gives an justification on the family-wise error of simultaneous tests $H_{0ij} : \sigma_{ij1} = \sigma_{ij2}$ for $1 \leq i \leq j \leq p$. For example, by thresholding $M_{ij}$ at level $4 \log p - \log \log p + q \alpha$, the family-wise error is controlled asymptotically at level $\alpha$, i.e.

$$\text{FWER} = \mathbb{P} \left( \max_{(i,j) \in G} M_{ij} \geq 4 \log p - \log \log p + q \alpha \right) \to \alpha,$$

where $G = \{(i, j) : \sigma_{ij1} = \sigma_{ij2}\}$ and $\text{Card}(G^c) = o(p^2)$. These tests are useful in the studies of differential coexpression in genetics; see de la Fuentea (2010).

The test introduced in Chapter 3 are based on the asymptotic results. When the sample size is small, say $n \leq 30$, the critical value derived from the asymptotic distribution is not sufficiently accurate and modification is thus needed. The following “normal cut off” method can be used instead. Let $(X_k^*, 1 \leq k \leq n_1)$ and $(Y_k^*, 1 \leq k \leq n_2)$ be generated from i.i.d $N(0, I_{p \times p})$. Let $M_n^*$ be the maximum test statistic constructed from $(X_k^*, 1 \leq k \leq n_1)$ and $(Y_k^*, 1 \leq k \leq n_2)$. It follows from Theorem 7 that under (C1)-(C3), we have

$$\sup_{y \in \mathbb{R}} |\mathbb{P}(M_n \geq y) - \mathbb{P}(M_n^* \geq y)| \to 0$$

under the null, where $M_n$ is the maximum test statistic computed from the original data. Let $y_n(\alpha)$ be the critical value such that $\mathbb{P}(M_n^* \geq y_n(\alpha)) = \alpha$. Then we have
\[ P(M_n \geq y_n(\alpha)) \rightarrow \alpha \text{ under } H_0. \] So \( y_n(\alpha) \) can be used as the critical value when the sample size is small and it can be obtained easily by simulation. Some additional simulation results illustrating the case with a small sample size (\( n = 30 \)) are given in Appendix B.

Besides testing the means and covariance matrices of two populations, another interesting and related problem is the testing of the equality of two distributions based on the two samples. That is, we wish to test \( H_0 : \mathbb{P}_1 = \mathbb{P}_2 \) versus \( H_1 : \mathbb{P}_1 \neq \mathbb{P}_2 \), where \( \mathbb{P}_i \) is the distribution of \( \mathcal{N}_p(\mu_i, \Sigma_i) \), \( i = 1, 2 \). We shall report the details of the results elsewhere in the future as a significant amount of additional work is still needed.

The major emphasis of Chapter 4 has been to show that with shape constraints it is possible to construct confidence intervals that have expected length that adapts to individual functions. We shall discuss briefly the maximum expected lengths of our procedures over Lipschitz classes that are either monotone or convex in a way that is similar to that provided in Dümbgen (1998) and Dümbgen (2003) for the maximum width of a confidence band. We shall also explain how our results can be extended to the problem of estimating the value of \( f \) at points other than 0.

Although the focus of Chapter 4 has been on the construction of a confidence interval with the expected length adaptive to each individual convex or monotone function, these results do yield immediately adaptive minimax results for the ex-
pected length in the conventional sense. Define

\[ F_c(\beta, M) = F_c \cap \Lambda(\beta, M) \quad \text{and} \quad F_m(\beta, M) = F_m \cap \Lambda(\beta, M). \]

The following results are direct consequence of Theorems 16 and 18.

**Corollary 1.** (i). The confidence interval \( CI^m_\ast \) defined in (4.2.6) satisfies

\[ \sup_{f \in F_m(\beta, M)} E_f L(CI^m_\ast) \leq C_1 M^{\frac{1}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}}. \]  

(5.0.1)

simultaneously for all \( 0 \leq \beta \leq 1 \) and \( 1 < M < \infty \), for some absolute constant \( C_1 > 0 \).

(ii). The confidence interval \( CI^c_\ast \) defined in (4.2.8) satisfies

\[ \sup_{f \in F_c(\beta, M)} E_f L(CI^c_\ast) \leq C_2 M^{\frac{1}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}}. \]  

(5.0.2)

simultaneously for all \( 1 \leq \beta \leq 2 \) and \( 1 < M < \infty \), for some absolute constant \( C_2 > 0 \).

We should note that these ranges of Lipschitz classes are the only ones of interest in these cases. In particular suppose that \( CI \) is a confidence interval with guaranteed coverage over the class of monotonically nondecreasing functions. Then for any \( \beta > 1 \) the class \( \Lambda(\beta, M) \) includes the linear function \( f_k(t) = kt \). As shown in Example 1 in Section 4.1.2

\[ L^*_\alpha(f_k, F_m) \geq (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(3k)^{\frac{1}{3}} z_\alpha^{\frac{2}{3}} n^{-\frac{1}{3}}. \]

Hence,

\[ \sup_{f \in F_m(\beta, M)} E_f L(CI) \geq \sup_k L^*_\alpha(f_k, F_m) = \sup_k (1 - \frac{1}{\sqrt{2\pi z_\alpha}})(3k)^{\frac{1}{3}} z_\alpha^{\frac{2}{3}} n^{-\frac{1}{3}} = \infty. \]  

(5.0.3)
A similar result holds for convex functions assumed to belong to \( \Lambda(\beta, M) \) with \( \beta > 2 \). On the other hand suppose \( f \) is convex and assumed to belong to \( \Lambda(\beta, M) \) with \( \beta < 1 \). Then from the assumption that \( f \) is in \( \Lambda(\beta, M) \) it follows that \( |f(1/2) - f(-1/2)| \leq M \). Convexity then shows that \( f \in \Lambda(1, M) \) and the maximum expected length over this class is given above.

The focus of Chapter 4 has been on the problem of estimating the value of \( f(0) \). The basic development is similar for any other point \( t \) in the interior of the interval \([-1/2, 1/2]\) unless \( t \) is near to the boundary. More specifically for any \( 0 \leq t < 1/2 \) we can consider estimators \( \delta^R_j(t) = 2^j(Y(t + 2^{-j}) - Y(t)) \) and \( \delta_j^L(t) = 2^j(Y(t) - Y(t - 2^{-j})) \) where \( j \geq -\log_2(\frac{1}{4} - \frac{t}{2}) \) for monotone functions and \( \delta_j(t) = 2^{j-1}(Y(t + 2^{-j}) - Y(t - 2^{-j})) \) where \( j \geq -\log_2(\frac{1}{2} - t) \) for convex functions.

The basic theory is the same as before.

For monotonically nondecreasing functions, the confidence interval \( CI^m_j(t) \) is replaced by

\[
CI^m_j(t) = [\delta^L_j(t) - z_{\alpha/2}\sqrt{2}\sigma_j, \delta^R_j(t) + z_{\alpha/2}\sqrt{2}\sigma_j]
\]

and the choice of \( \hat{j} \) is given by

\[
\hat{j}(t) = \inf_{j \geq -\log_2(\frac{1}{4} - \frac{t}{2})} \left\{ j : \xi_j(t) \leq \frac{3}{2}z_{\alpha}\sigma_j \right\},
\]

where \( \xi_j(t) = 2^{j-1}(Y(t + 2^{-j+1}) - Y(t + 2^{-j})) - 2^{j-1}(Y(t - 2^{-j}) - Y(t - 2^{-j+1})) \).

The final confidence interval is defined by

\[
CI^m_\ast = CI^m_{\hat{j}(t)},
\]

(5.0.4)
For convex functions, the confidence interval $CI_j^c$ is replaced by

$$CI_j^c(t) = [\delta_{j+1}(t) - (\delta_j(t) - \delta_{j+1}(t))_+ - z_{\alpha} \sqrt{5} \sigma_j, \delta_{j+1}(t) + z_{\alpha} \sigma_{j+1}]$$

and $\hat{j}$ is chosen to be

$$\hat{j}(t) = \inf_{j \geq -\log_2(\frac{1}{2} - t)} \{j : T_j(t) \leq z_{\alpha} \sigma_j \}.$$ 

Define the final confidence interval by

$$CI^* = CI_{\hat{j}(t)}^c.$$ 

The modulus of continuity defined in (5.0.5) is replaced by

$$\omega(E, f, t, F) = \sup\{|g(t) - f(t)| : \|g - f\|_2 \leq E, g \in F\}.$$  

(5.0.5)

The earlier analysis then yields

$$E_fL(CI^m_\alpha(t)) \leq c_1 L_\alpha^*(f, t, F_m),$$

and

$$E_fL(CI^c_\alpha(t)) \leq c_2 L_\alpha^*(f, t, F_c),$$

where we now have

$$L_\alpha^*(f, t, F) \geq (1 - \frac{1}{\sqrt{2\pi} z_{\alpha}}) \omega(z_{\alpha} \sqrt{n}, f, t, F).$$

Finally we should note that at the boundary the construction of a confidence interval must be unbounded. For example any honest confidence interval for $f(1/2)$ must be of the form $[\hat{f}(1/2), \infty)$, otherwise it cannot have guaranteed coverage probability.
My future research will in part build on the foundation of my current work. Within the general area of high dimensional statistics, I am particularly interested in understanding sparse discriminant analysis and testing precision matrices. In the area of shape constraint inference, I am mostly interested in the construction of adaptive and robust confidence bands and their extension to multi-dimensions. I do however have fairly broad statistical interests and in addition to continuing my present research agenda, I hope to expand into additional collaborative research.
6.1 High Dimensional Sparse Discriminant Analysis

Linear discriminant analysis relies on an assumption of equality of two covariance matrices. In this case, under a particular sparsity condition, Cai and Liu (2011) introduced an $\ell_1$ constrained minimization method, with a tuning parameter $\lambda_n$ chosen by cross-validation, that has a misclassification rate that converges to the Bayes misclassification rate. This approach did not exploit the fact that the components of the linear discriminant are usually heteroscedastic. It appears that an alternative approach based on first standardizing the linear discriminant, and then applying a universal threshold value $\lambda_n$ may lead to a similar result under weaker conditions. I have preliminary results that show that a universal threshold value $\lambda_n = \sqrt{\frac{c \log p}{n}}$, for some $c > 2$ works when the groups are sufficiently separated.

It would be interesting to further investigate the minimal separation between two groups allowing for the misclassification rate to converge to the Bayes misclassification rate. The approach with the universal threshold should also allow for the classification of multiple groups. Pairwise classification is not efficient in this case and I believe a maximum-type procedure would be useful to solve this problem. My expertise in constructing maximum-type tests may be of real help here.

In the case where the covariance matrices are not equal, a quadratic discriminant analysis method should be used. This is a very interesting problem because it
involves the estimation of the difference of two precision matrices, the determinants of the covariance matrices, as well as the functionals such as \((\mathbf{\mu}_1 - \mathbf{\mu}_2)^T(\Sigma_1^{-1} - \Sigma_2^{-1})(\mathbf{\mu}_1 - \mathbf{\mu}_2)\), where \(\mathbf{\mu}_1, \mathbf{\mu}_2\) are the means of two populations. I believe my studies on the multivariate methods can contribute to a detailed analysis of this problem.

6.2 High Dimensional Testing of Precision Matrices

The equality of two precision matrices is equivalent to the equality of two covariance matrices under the null hypothesis. However, under the alternative, it is sometimes more reasonable to assume the sparse structure of the difference of two precision matrices, instead of two covariance matrices, because zero partial correlations imply a graph structure and it is often the case that two graph structures are quite similar to each other. Application includes gene expression analysis, networks, etc.

Let \(X\) and \(Y\) be two \(p\) variate normal random vectors with covariance matrices \(\Sigma_1\) and \(\Sigma_2\) respectively. Let \(\{X_1, \ldots, X_{n_1}\}\) be i.i.d. random samples from \(X\) and let \(\{Y_1, \ldots, Y_{n_2}\}\) be i.i.d. random samples from \(Y\) that are independent of \(\{X_1, \ldots, X_{n_1}\}\). We wish to test the hypotheses

\[
H_0 : \Omega_1 = \Omega_2 \quad \text{versus} \quad H_1 : \Omega_1 \neq \Omega_2,
\]

where \(\Omega_1 = \Sigma_1^{-1}\) and \(\Omega_2 = \Sigma_2^{-1}\) are two corresponding precision matrices. It is well know that precision matrix is closely related to the regression model. That is, for
\( X = (X_1, ..., X_p)' \sim N(\mu, \Sigma) \), we can write

\[
X_i = \alpha_i + X'_i \beta_i + \epsilon_i,
\]

where \( \epsilon_i \sim N(0, \sigma_{ii} - \Sigma_{i, -i} \Sigma^{-1}_{-i, -i} \Sigma_{-i, i}) \) is independent of \( X_{-i} \),

\[
\alpha_i = \mu_i - \Sigma_{i, -i} \Sigma^{-1}_{-i, -i} \mu_{-i},
\]

and \((\sigma_{ij}) = \Sigma\). The regression coefficients vector \( \beta_i \) and the error terms \( \epsilon_i \) satisfy

\[
\beta_i = -\omega^{-1}_{ii} \Omega_{-i, i} \quad \text{and} \quad \text{Cov}(\epsilon_i, \epsilon_j) = \frac{\omega_{ij}}{\omega_{ii} \omega_{jj}}.
\]

Thus, a maximum statistic can be constructed based on the estimates of covariances among the residuals \( \hat{\epsilon}_i \) and \( \hat{\epsilon}_j \) for \( 1 \leq i \leq j \leq p \). Similarly, I will study the limiting distribution of the maximum test statistic under the null hypothesis and the asymptotic powers under the alternative.

### 6.3 Robust and Adaptive Confidence Bands Under Shape Constraints

My work on confidence sets under shape constraints has so far concentrated on confidence intervals. However as in the case for confidence intervals, adaptive confidence bands can also be constructed if the functions obey certain shape constraints. Based on Dümbgen (2003), I have recently been studying the optimal average expected width of confidence bands for both monotone and convex functions. In
these cases under a Gaussian regression model it is possible to construct confidence bands with maximum expected average width of order \((\log n/n)^{1/3}\) and \((\log n/n)^{2/5}\) respectively for these two functional classes. I am interested in extending this work to seek procedures that adapt to each monotone and convex function as there are some monotone and convex functions which should have bands with a much smaller average width. I am also interested in extending this work to models with heavy-tailed noise. Here, a local median transformation introduced in Brown, Cai and Zhou (2008) and Cai and Zhou (2009) can be used to transform the data so that it fits the normal model well. It is however unclear whether full adaptation to each monotone or convex function would be possible in such a set up.

Finally I am also interested in considering the constructions of adaptive confidence sets under shape constraints in higher dimensions. For example additive models for both monotone and convex functions can be considered. I have made some progress on confidence intervals in these problems but new ideas are still needed in order to construct fully adaptive procedures.
Appendix A

Testing High Dimensional Means

A.1 Power Comparison of the Oracle Tests

The test $\Phi_\alpha(\Omega)$ is shown in Section 3.2 to be minimax rate optimal for testing against sparse alternatives. We now compare the power of the test $\Phi_\alpha(\Omega)$ with that of $\Phi_\alpha(\Omega^{1/2})$ and $\Phi_\alpha(I)$ under the same alternative $H_1$ as in Section 3.2. Let

$$A = \{1 \leq i \leq p : (\Omega^{1/2})_{ij} = 0 \text{ for all } j \neq i\}.$$

That is, $i \in A$ if and only if all the entries in the $i$-th row of $\Omega^{1/2}$ are zero except for the diagonal entry.

Proposition 5. (i) Suppose (C1)-(C3) hold. Then under $H_1$ with $r < 1/6$, we have

$$\lim_{p \to \infty} \frac{P_{H_1}(\Phi_\alpha(\Omega) = 1)}{P_{H_1}(\Phi_\alpha(I) = 1)} \geq 1. \quad (A.1.1)$$

(ii) Suppose (C1)-(C3) hold. Assume there exists a constant $\epsilon_0 > 0$, such that for each $i \in A^c$ at least one non-diagonal element in the $i$-th row of $\Omega^{1/2}$ has a
magnitude larger than \( \epsilon_0 \). Then, under \( H_1 \) with \( k_p = O(p^{\tau}) \) for all \( 0 < \tau < 1 \), we have

\[
\lim_{p \to \infty} \frac{P_{H_1}(\Phi_\alpha(\Omega) = 1)}{P_{H_1}(\Phi_\alpha(\Omega^{1/2}) = 1)} \geq 1. \tag{A.1.2}
\]

The condition on \( \Omega^{1/2} \) is mild. In fact, by the definition of \( A \), there is at least one nonzero and non-diagonal element in each \( i \)-th row of \( \Omega^{1/2} \) with \( i \in \mathcal{A}^c \). In Proposition 5, we assume that these nonzero and non-diagonal elements have magnitudes larger than \( \epsilon_0 \).

Proposition 5 shows that, under some sparsity condition on \( \delta \), \( \Phi_\alpha(\Omega) \) is uniformly at least as powerful as both \( \Phi_\alpha(\Omega^{1/2}) \) and \( \Phi_\alpha(I) \).

We now briefly discuss the different conditions on \( r \) in the theoretical results. For the maximum type test statistics, the range \( r < 1/2 \) is nearly optimal. Indeed, in the mean testing problem, the case \( r > 1/2 \) is treated as the dense setting and \( r < 1/2 \) as the sparse setting, similar to other sequence estimation problems. In the dense setting, the sum square type test statistics may outperform the maximum type test statistics under certain conditions. The different conditions imposed on \( r \) in this section are due to the technical arguments used in the proofs. We believe these ranges for \( r \) can be improved to \( r < 1/2 \) but the proof would be difficult. When the assumption on \( r \) does not hold, the tests are still valid but the comparison results may fail.

The test \( \Phi_\alpha(\Omega) \) can be strictly more powerful than \( \Phi_\alpha(\Omega^{1/2}) \) and \( \Phi_\alpha(I) \). Assume
that

\[ H_1^\prime: \quad \delta \text{ has } m = p^r, \ r < 1/2 \text{ nonzero coordinates with} \]

\[
\frac{|\delta_i|}{\sqrt{\sigma_{ii}}} = \sqrt{\frac{2\beta \log p}{n}}, \text{ where } \beta \in (0, 1) \quad (A.1.3)
\]

if \( \delta_i \neq 0 \). The nonzero locations \( l_1 < l_2 < \cdots < l_m \) are randomly and uniformly drawn from \( \{1, 2, \ldots, p\} \).

**Proposition 6.** (i). Suppose that (C1) and (C2) hold. Then, under \( H_1^\prime \) with \( \beta \geq (1 - \sqrt{r})^2 + \varepsilon \) for some \( \varepsilon > 0 \), we have

\[
\lim_{p \to \infty} P_{H_1^\prime}(\Phi_0(I) = 1) = 1.
\]

If \( \beta < (1 - \sqrt{r})^2 \), then

\[
\lim_{p \to \infty} P_{H_1^\prime}(\Phi_0(I) = 1) \leq \alpha.
\]

(ii). Suppose that (C1) and (C3) hold and \( r < 1/4 \). Then, under \( H_1^\prime \) with

\[
\beta \geq (1 - \sqrt{r})^2 / \left( \min_{1 \leq i \leq p} \sigma_{ii} \omega_{ii} \right) + \varepsilon \quad \text{for some } \varepsilon > 0, \quad (A.1.4)
\]

we have

\[
\lim_{p \to \infty} P_{H_1^\prime}(\Phi_0(\Omega) = 1) = 1.
\]

The condition \( r < 1/4 \) can be weakened if we assume some stronger condition on \( \Omega \). In fact, based on the proof, we can see that it can be weakened to \( r < 1/2 \) if \( \Omega \) is \( s_p \)-sparse and \( s_p = O(p^\tau), \forall \tau > 0 \).
Note that $\sigma_{ii}\omega_{ii} \geq 1$ for $1 \leq i \leq p$. When the variables are correlated, $\omega_{ii}$ can be strictly larger than $1/\sigma_{ii}$. For example, let $\Sigma = (\phi^{|i-j|})$ with $|\phi| < 1$. Then $\min_{1 \leq i \leq p} \sigma_{ii}\omega_{ii} \geq (1 - \phi^2)^{-1} > 1$. That is, $M_\Omega$ is strictly more powerful than $M_I$ under $H'_1$.

In Proposition 6, the comparison between $\Phi_\alpha(\Omega)$ and $\Phi_\alpha(I)$ is restricted to $H'_1$, which is special. However, the proof of Proposition 6 in fact implies the following more general result. Suppose that $\min_{1 \leq i \leq p} \sigma_{ii}\omega_{ii} \geq 1 + \varepsilon_1$ for some $\varepsilon_1 > 0$. Let $\beta_0$ and $\beta_1$ be any constants satisfying

$$\frac{(1 - \sqrt{r})^2}{\min_{1 \leq i \leq p} \sigma_{ii}\omega_{ii}} + \varepsilon \leq \beta_0 < \beta_1 < (1 - \sqrt{r})^2$$

for some $\varepsilon > 0$. Replacing (A.1.3) by $\sqrt{\frac{2\beta_0 \log p}{n}} \leq \frac{|\delta_i|}{\sqrt{\sigma_{ii}}} \leq \sqrt{\frac{2\beta_1 \log p}{n}}$, we have

$$\lim_{p \to \infty} P_{H'_1}(\Phi_\alpha(\Omega) = 1) = 1,$$

and

$$\overline{\lim}_{p \to \infty} P_{H'_1}(\Phi_\alpha(I) = 1) \leq \alpha.$$

We now turn to the comparison of the power of $\Phi_\alpha(\Omega)$ with that of $\Phi_\alpha(\Omega^{1/2})$ under the alternative

$$H'_1'': \quad \delta \text{ has } m = p^r, \quad r < 1/7 \text{ nonzero coordinates with }$$

$$\max_j |a_{ji}\delta_i| = \sqrt{\frac{2\beta \log p}{n}}, \quad \text{where } \beta \in (0, 1) \quad (A.1.5)$$

if $\delta_i \neq 0$, where $\Omega^{1/2} = (a_{ij})$. The nonzero locations $l_1 < l_2 < \cdots < l_m$ are randomly and uniformly drawn from $\{1, 2, \ldots, p\}$.
Proposition 7. (i) Suppose (C1) holds. Then under $H''_1$ with $\beta < (1 - \sqrt{r})^2$, we have
\[
\lim_{p \to \infty} P_{H''_1} \left( \Phi_\alpha(\Omega^\frac{1}{2}) = 1 \right) \leq \alpha.
\]

(ii) Suppose that (C1) and (C3) hold. Then under $H''_1$ with
\[
\beta \geq (1 - \sqrt{r})^2 / \left( \min_{1 \leq i \leq p} (\omega_{ii} / \max_j a_{ji}^2) \right) + \varepsilon \quad \text{for some } \varepsilon > 0,
\]
we have
\[
\lim_{p \to \infty} P_{H''_1} \left( \Phi_\alpha(\Omega) = 1 \right) = 1.
\]

It is easy to check that $\omega_{ii} / (\max_j a_{ji}^2) \geq 1$ for all $1 \leq i \leq p$. When the variables are correlated, $\omega_{ii}$ can be much larger than $\max_j a_{ji}^2$. For example, if for every row of $\Omega^\frac{1}{2}$, there are at least 2 nonzero $a_{ij}$, then $\omega_{ii} = \sum_{j=1}^p a_{ij}^2 > \max_j a_{ji}^2$. In this case, $M_\Omega$ is strictly more powerful than $M_{\Omega^\frac{1}{2}}$.

As the discussion below Proposition 6, the condition (A.1.5) can be generalized. Suppose that $\min_{1 \leq i \leq p} (\omega_{ii} / \max_j a_{ji}^2) > 1 + \varepsilon_1$ for some $\varepsilon_1 > 0$. Let $\beta_0$ and $\beta_1$ be any constants satisfying
\[
\frac{(1 - \sqrt{r})^2}{\min_{1 \leq i \leq p} (\omega_{ii} / \max_j a_{ji}^2)} + \epsilon \leq \beta_0 < \beta_1 < (1 - \sqrt{r})^2
\]
for some constant $\epsilon > 0$. If (A.1.5) is replaced by $\sqrt{\frac{2 \beta_0 \log p}{n}} \leq \max_j |a_{ji} \delta_i| \leq \sqrt{\frac{2 \beta_1 \log p}{n}}$, then
\[
\lim_{p \to \infty} P_{H''_1} \left( \Phi_\alpha(\Omega) = 1 \right) = 1
\]
and
\[
\lim_{p \to \infty} P_{H''_1} \left( \Phi_\alpha(\Omega^\frac{1}{2}) = 1 \right) \leq \alpha.
\]
A.2 Additional Simulation Results

In this section we present additional simulation results comparing the numerical performance of the proposed test with that of other tests in the non-Gaussian setting.

For Model 6, two independent random samples \( \{X_k\}_{k=1}^{n_1} \) and \( \{Y_k\}_{k=1}^{n_2} \) are generated from multivariate models

\[
X_k = \Gamma Z_k^{(1)} + \mu_1 \quad \text{and} \quad Y_k = \Gamma Z_k^{(2)} + \mu_2,
\]

\( \Gamma \Gamma' = \Sigma \), where the components of \( Z_k^{(i)} = (Z_{k1}, ..., Z_{kp})' \) are i.i.d. standardized Gamma(10,1) random variables. We consider the case when \( \mu_1 - \mu_2 \) has \( m = \sqrt{p} \) nonzero elements with the same signal strength for each coordinate. The results are summarized in Table A.1.
<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>50</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Empirical Sizes</td>
<td>Empirical Powers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T^2$</td>
<td>0.041</td>
<td>0.051</td>
<td>–</td>
<td>0.716</td>
<td>0.456</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>BS</td>
<td>0.055</td>
<td>0.069</td>
<td>0.062</td>
<td>0.153</td>
<td>0.183</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.060</td>
<td>0.063</td>
<td>0.063</td>
<td>0.170</td>
<td>0.177</td>
<td>0.207</td>
<td></td>
</tr>
<tr>
<td>CQ</td>
<td>0.058</td>
<td>0.072</td>
<td>0.064</td>
<td>0.152</td>
<td>0.180</td>
<td>0.186</td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(I)$</td>
<td>0.032</td>
<td>0.034</td>
<td>0.033</td>
<td>0.114</td>
<td>0.116</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\Omega^\frac{1}{2})$</td>
<td>0.035</td>
<td>0.039</td>
<td>0.041</td>
<td>0.632</td>
<td>0.508</td>
<td>0.417</td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\Omega)$</td>
<td>0.044</td>
<td>0.036</td>
<td>0.043</td>
<td>0.873</td>
<td>0.702</td>
<td>0.565</td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha(\hat{\Omega})$</td>
<td>0.050</td>
<td>0.039</td>
<td>0.048</td>
<td>0.875</td>
<td>0.722</td>
<td>0.597</td>
<td></td>
</tr>
</tbody>
</table>

Table A.1: Sizes and Powers of tests with $\alpha = 0.05$ and $n = 100$ for Model 6. $\mu_1 - \mu_2$ has $m = \sqrt{p}$ nonzero elements. Signal strength keeps the same. Based on 1000 replications.

### A.3 Proof of Technical Lemmas

**Proof of Lemma 4.** We only need to prove (3.2.14) because the proof of (3.2.18) is similar. We re-order $\omega_1, \ldots, \omega_p$ as $|\omega_{i(1)}| \geq \ldots \geq |\omega_{i(p)}|$ for $i = 1, \ldots, p$. Let $a$
satisfy $2r < a < 1 - 2r$ with $r < 1/4$. Define $\mathcal{I} = \{1 \leq i_1 < \ldots < i_m \leq p\}$ and

$$\mathcal{I}_0 = \left\{1 \leq i_1 < \ldots < i_m \leq p : \text{there exist some } 1 \leq k \leq m \text{ and some } j \neq k \right. \left.\text{with } 1 \leq j \leq m, \right. \left.\text{such that } |\omega_{i_ki_j}| \geq |\omega_{i_k(p^r)}| \right\}.$$  

We can show that

$$|\mathcal{I}_0| = O\left(p \cdot p^a \left(\frac{p}{p^r - 2}\right)\right) \text{ and } |\mathcal{I}| = \left(\frac{p}{p^r}\right).$$

Therefore

$$|\mathcal{I}_0|/|\mathcal{I}| = O\left(p^{a+2r-1}\right) = o(1). \quad (A.3.1)$$

For $1 \leq t \leq m$, write

$$(\Omega\delta)_{lt} = \sum_{k=1}^{p} \omega_{l,k} \delta_k = \omega_{lt} \delta_t + \sum_{j=1, j \neq t}^{m} \omega_{lt_j} \delta_{l_j}.$$  

Note that for every $(l_1, \ldots, l_m) \in \mathcal{I}_0^c$,

$$\sum_{j=1, j \neq t}^{m} |\omega_{lt_j}| \leq p^r \sqrt{\frac{C_0}{p^a}}.$$  

It follows that for $H \in \mathcal{I}_0^c$ and $i \in H$,

$$\left| \frac{(\Omega\delta)}{\sqrt{\omega_{ii}}} - \sqrt{\omega_{ii}} \delta_i \right| = O\left(p^{r-a/2}\right) \max_{i \in H} |\delta_i|. \quad (A.3.2)$$

By (A.3.1) and (A.3.2), (3.2.14) is proved.

**Proof of Lemma 5.** We have

$$P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) = 1 - \prod_{i=1}^{k} \left(1 - P(Y_i \geq x + a_n)\right)$$
\[ = 1 - \exp \left( \sum_{i=1}^{k} \log \left( 1 - \Pr(Y_i \geq x + a_n) \right) \right). \]

Let \(0 < \varepsilon < 1/2\) be any small number and \(M\) be any large number. Define

\[ E = \{1 \leq i \leq k : \Pr(Y_i \geq x + a_n) \geq \varepsilon\}. \]

We first consider those \(n\) and \(k\) such that \(\text{Card}(E) \leq \varepsilon^{-2}\) and

\[ \sum_{i=1}^{k} \Pr(Y_i \geq x + a_n) \leq M. \]

For \(i \in E^c\), by the inequality \(|\log(1 - x) + x| \leq x^2\) with \(|x| < 1/2\), we have

\[ \left| \frac{\log \left( 1 - \Pr(Y_i \geq x + a_n) \right)}{-\Pr(Y_i \geq x + a_n)} - 1 \right| \leq \varepsilon. \]  

(A.3.3)

Write \(a_n = \varepsilon_n(\log n)^{-1/2}\), where \(\varepsilon_n \to 0\). Let \(b_n = |\varepsilon_n|^{-1/2}(\log n)^{1/2}\). We have for any large \(\alpha > 0\),

\[ \Pr(Y_i \geq x + a_n) = \int_{y \geq x - \mu_i + a_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy \]

\[ = \int_{y \geq x - \mu_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-a_n)^2}{2}} \, dy \]

\[ = \int_{y \geq x - \mu_i, |y| \leq b_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-a_n)^2}{2}} \, dy + \int_{y \geq x - \mu_i, |y| > b_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-a_n)^2}{2}} \, dy \]

\[ = (1 + o(1)) \Pr(Y_i \geq x) + O(n^{-\alpha}) \]

(A.3.4)

where \(O(1)\) and \(o(1)\) are uniformly in \(i\) and \(\mu_i\). Thus, we have

\[ \sum_{i \in E^c} \log \left( 1 - \Pr(Y_i \geq x + a_n) \right) \]

\[ \leq -(1 - 2\varepsilon) \sum_{i \in E^c} \Pr(Y_i \geq x) + O(n^{-\alpha+1}) \]

150
\[
\leq (1 - 2\varepsilon)(1 + 2\varepsilon)^{-1} \sum_{i \in E^c} \log \left(1 - P(Y_i \geq x)\right) + O(n^{-\alpha+1}), \tag{A.3.5}
\]

where in the last inequality we used (A.3.3) with \(a_n = 0\). By (A.3.4),

\[
\sum_{i \in E} \log \left(1 - P(Y_i \geq x + a_n)\right) = \sum_{i \in E} \log \left(1 - P(Y_i \geq x)\right) + o(1)\varepsilon^{-2}. \tag{A.3.6}
\]

Combining (A.3.5) and (A.3.6), we have

\[
\sum_{i=1}^{k} \log \left(1 - P(Y_i \geq x + a_n)\right) \leq \sum_{i=1}^{k} \log \left(1 - P(Y_i \geq x)\right) + 4\varepsilon M + o(1)\varepsilon^{-2}.
\]

Hence

\[
P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) \geq P\left(\max_{1 \leq i \leq k} Y_i \geq x\right) - |e^{4\varepsilon M + o(1)\varepsilon^{-2}} - 1|. \tag{A.3.7}
\]

Note that if \(\sum_{i=1}^{k} P(Y_i \geq x + a_n) > M\), then

\[
P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) \geq 1 - e^{-M}. \tag{A.3.8}
\]

If \(\text{Card}(E) > \varepsilon^{-2}\), then

\[
P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) \geq 1 - (1 - \varepsilon)^{\varepsilon^{-2}} \geq 1 - e^{-\varepsilon^{-1}}. \tag{A.3.9}
\]

By (A.3.7)-(A.3.9), we have

\[
P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) \geq P\left(\max_{1 \leq i \leq k} Y_i \geq x\right) - |e^{4\varepsilon M + o(1)\varepsilon^{-2}} - 1| - e^{-\varepsilon^{-1}} - e^{-M}.
\]

Similarly, we can prove

\[
P\left(\max_{1 \leq i \leq k} Y_i \geq x + a_n\right) \leq P\left(\max_{1 \leq i \leq k} Y_i \geq x\right) + |e^{4\varepsilon M + o(1)\varepsilon^{-2}} - 1| + e^{-\varepsilon^{-1}} + e^{-M}.
\]

By letting \(n \to \infty\) first, following by \(\varepsilon \to 0\) and then \(M \to \infty\), the lemma is proved. \(\Box\)
A.4 Proof of Propositions

Proof of Proposition 5 (i). Let $U$ be a multivariate normal random vector with zero mean and covariance matrix $\Sigma$. Let $Z = \delta + U$, where $\delta$ and $U$ are independent. Without loss of generality, we assume that $\sigma_{ii} = 1$ for $1 \leq i \leq p$. Then $\omega_{ii} \geq 1$ for $1 \leq i \leq p$. Set $A = \{\max_{1 \leq i \leq p} |\delta_i| \leq 6\sqrt{\log p}\}$. By Lemma 4, we have

$$P\left(\max_{1 \leq i \leq p} |(\Omega\delta)_i|/\sqrt{\omega_{ii}} \geq (1 - o(1)) \max_{1 \leq i \leq p} |\delta_i|\right) \to 1. \quad (A.4.1)$$

Thus we have

$$P\left(M_\Omega \in R_\alpha, A^c\right) \geq P\left(4\sqrt{\log p} - \max_{1 \leq i \leq p} |(\Omega U)_i|/\sqrt{\omega_{ii}} \geq \sqrt{2\log p}, A^c\right) + o(1) = P(A^c) + o(1). \quad (A.4.2)$$

Similarly, we have

$$P\left(M_I \in R_\alpha, A^c\right) = P(A^c) + o(1). \quad (A.4.3)$$

We next consider $P\left(M_\Omega \in R_\alpha, A\right)$ and $P\left(M_I \in R_\alpha, A\right)$. For notation briefness, we denote $P(BA|\delta)$ and $P(B|\delta)$ by $P_{\delta,A}(B)$ and $P_{\delta}(B)$ respectively for any event $B$. Let $H = \text{supp}(\delta) = \{l_1, ..., l_m\}$ with $m = p^\gamma$ and $H^c = \{1, ..., p\} \setminus H$. We have

$$P_{\delta,A}(M_I \in R_\alpha) = P_{\delta,A}(\max_{i \in H} |Z_i| \geq \sqrt{x_p}) + P_{\delta,A}(\max_{i \in H} |Z_i| < \sqrt{x_p}, \max_{j \in H^c} |Z_j| \geq \sqrt{x_p}), \quad (A.4.4)$$

where $x_p = 2\log p - \log \log p + x$. Define

$$H_1^c = \{j \in H^c : |\sigma_{ij}| \leq p^{-\xi} \text{ for any } i \in H\}, \quad H_1 = H^c - H_1^c$$
for \(2r < \xi < (1 - r)/2\). It is easy to see that \(\text{Card}(H_1) \leq Cp^{r+2\xi}\). It follows that
\[
P\left(\max_{j \in H_1} |Z_j| \geq \sqrt{x_p} \right) \leq p^{r+2\xi}P\left(|N(0, 1)| \geq \sqrt{x_p} \right) = O(p^{r+2\xi-1}) = o(1). \quad (A.4.5)
\]
We claim that
\[
P_{\delta, A}\left(\max_{i \in H} |Z_i| < \sqrt{x_p}, \max_{j \in H_1^c} |Z_j| \geq \sqrt{x_p} \right)
\leq P_{\delta, A}\left(\max_{i \in H} |Z_i| < \sqrt{x_p}\right)P_{\delta, A}\left(\max_{j \in H_1^c} |Z_j| \geq \sqrt{x_p} \right) + o(1). \quad (A.4.6)
\]
Throughout the proof, \(O(1)\) and \(o(1)\) are uniformly for \(\delta\). To prove (B.1.2), we set \(E = \{\max_{i \in H} |Z_i| < \sqrt{x_p}\}\), \(F_j = \{|Z_j| \geq \sqrt{x_p}\}, j \in H_1^c\). Then by Bonferroni inequality, we have for any fixed integer \(k > 0\),
\[
P_{\delta, A}\left(\bigcup_{j \in H_1^c} \{E \cap F_j\}\right) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} \sum_{i_1 < \cdots < i_t \in H_1^c} P_{\delta, A}(E \cap \bigcap_{i=1}^t F_{i_t}) \quad (A.4.7)
\]
Let \(W = (w_{ij})\) be the covariance matrix of the vector \((Z_i, i \in H, Z_{i_1}, \ldots, Z_{i_t})'\) given \(\delta\). Note that \(W\) satisfies \(|w_{ij}| \leq p^{-\xi}\) for \(i \in H\) and \(j = i_1, \ldots, i_t \in H_1^c\). Define the matrix \(\tilde{W} = (\tilde{w}_{ij})\) with \(\tilde{w}_{ij} = w_{ij}\) for \(i, j \in H, \tilde{w}_{ij} = w_{ij}\) for \(i = i_1, \ldots, i_t \in H_1^c\) and \(\tilde{w}_{ij} = \tilde{w}_{ji} = 0\) for \(i \in H\) and \(j = i_1, \ldots, i_t \in H_1^c\). Set \(z = (u_i, i \in H, z_{i_1}, \ldots, z_{i_t})'\) and
\[
\mathcal{R} = \{|u_i + \delta_i| \leq \sqrt{x_p}, i \in H, |z_{i_1}| \geq \sqrt{x_p}, \ldots, |z_{i_t}| \geq \sqrt{x_p}\},
\]
\[
\mathcal{R}_1 = \mathcal{R} \cap \{|z|_{\infty} \leq 8\sqrt{t \log p}\},
\]
\[
\mathcal{R}_2 = \mathcal{R} \cap \{|z|_{\infty} > 8\sqrt{t \log p}\}.
\]
We have
\[
P_{\delta, A}(E \cap F_{i_1} \cap \cdots \cap F_{i_t}) = \frac{I\{A\}}{(2\pi)^{p+t}|W|^\frac{t}{2}} \int_{\mathcal{R}} \exp\left(-\frac{1}{2} z' W^{-1} z\right) dz. \quad (A.4.8)
\]
By (C1) we have $C_0^{-1} \leq \lambda_{\min}(\tilde{W}) \leq \lambda_{\max}(\tilde{W}) \leq C_0$. Note that $\|W - \tilde{W}\|_2 = O(p^{-\xi})$ and $|W| = (1 + O(p^{-\xi}))e^{+t} |\tilde{W}| = (1 + O(p^{2r-\xi}))|\tilde{W}|$. This implies that

$$\frac{1}{(2\pi)^{p^r+t}|W|^\frac{3}{2}} \int_{\mathcal{R}_1} \exp \left( -\frac{1}{2} z' W^{-1} z \right) dz = (1 + O(p^{2r-\xi} \log p)) \frac{1}{(2\pi)^{p^r+t}|\tilde{W}|^\frac{3}{2}} \int_{\mathcal{R}_1} \exp \left( -\frac{1}{2} z' \tilde{W}^{-1} z \right) dz.$$ (A.4.9)

Furthermore, it is easy to see that

$$\frac{1}{(2\pi)^{p^r+t}|W|^\frac{3}{2}} \int_{\mathcal{R}_2} \exp \left( -\frac{1}{2} z' W^{-1} z \right) dz = O(p^{-32t}),$$

$$\frac{1}{(2\pi)^{p^r+t}|W|^\frac{3}{2}} \int_{\mathcal{R}_2} \exp \left( -\frac{1}{2} z' \tilde{W}^{-1} z \right) dz = O(p^{-32t}).$$ (A.4.10)

Thus, it follows from (B.1.3)-(A.4.10) that

$$P_{\delta, A}(\mathcal{E} \cap F_{i_1} \cap \cdots \cap F_{i_t}) = (1 + O(p^{2r-\xi} \log p)) P_{\delta, A}(\mathcal{E}) P_{\delta}(F_{i_1} \cap \cdots \cap F_{i_t}) + O(p^{-32t}).$$

As the proof of Lemma 1, we can show that

$$\sum_{i_1 < \cdots < i_t \in H^c_1} P_{\delta}(F_{i_1} \cap \cdots \cap F_{i_t}) = (1 + o(1)) \pi^{-\frac{t}{2}} \frac{1}{t!} \exp \left( -\frac{tq_a}{2} \right).$$

It follows from (A.4.7) that

$$P_{\delta, A} \left( \bigcup_{j \in H_1^c} \{E \cap F_j\} \right) \leq \alpha P_{\delta, A}(\mathcal{E}) + o(1).$$

This, together with (A.4.4) and (B.1.1), implies that

$$P_{\delta, A}(M_I \in R_\alpha) \leq \alpha P(A) + (1 - \alpha) P_{\delta, A}(\mathcal{E}^c) + o(1),$$

where $o(1)$ is uniformly for $\delta$. Hence, we have

$$P(M_I \in R_\alpha, A) \leq \alpha P(A) + (1 - \alpha) P(\mathcal{E}^c, A) + o(1).$$
and
\[ P(M_I \in R_\alpha) \leq \alpha P(A) + P(A^c) + (1 - \alpha)P(E^c, A) + o(1). \] (A.4.11)

We next prove that
\[ P(M_\Omega \in R_\alpha, A) \geq \alpha P(A) + (1 - \alpha)P(\tilde{E}^c, A) + o(1), \] (A.4.12)
and hence
\[ P(M_\Omega \in R_\alpha) \geq \alpha P(A) + P(A^c) + (1 - \alpha)P(\tilde{E}^c, A) + o(1), \] (A.4.13)
where \( \tilde{E} = \{\max_{i \in H} |Z_i^o| < \sqrt{xp} \} \), \( Z^o = (Z_1^o, \ldots, Z_p^o)' \), and \( Z_i^o = \frac{(\Omega Z_i)}{\sqrt{\omega_i}} \). It suffices to show that
\[ P_{\delta, A}(M_\Omega \in R_\alpha) \geq \alpha I\{A\} + (1 - \alpha)P_{\delta, A}(\tilde{E}^c) + o(1). \] (A.4.14)

Define \( \tilde{H}_1^c = \{j \in H^c : |\omega_{ij}| \leq p^{-\xi} \text{ for any } i \in H\} \) for \( 2r < \xi < (1 - r)/2 \). It is easy to see that \( \text{Card}(\tilde{H}_1^c) \geq p - O(p^{r+2\xi}) \). Then
\[
P_{\delta, A}(M_\Omega \in R_\alpha)
= P_{\delta, A}(\max_{i \in H} |Z_i^o| \geq \sqrt{xp}) + P_{\delta, A}(\max_{i \in H} |Z_i^o| < \sqrt{xp}, \max_{j \in H^c} |Z_j^o| \geq \sqrt{xp})
\geq P_{\delta, A}(\max_{i \in H} |Z_i^o| \geq \sqrt{xp}) + P_{\delta, A}(\max_{i \in H} |Z_i^o| < \sqrt{xp}, \max_{j \in H^c} |Z_j^o| \geq \sqrt{xp}).
\]

Note that on \( A \), \( \max_{j \in H^c} |(\Omega \delta)_j| = \max_{j \in \tilde{H}_1^c} \left| \sum_{i \in H} \omega_{ij} \delta_i \right| \leq 4p^{r-\xi} \sqrt{\log p} \). It follows from the same arguments as above and using the left hand side of Bonferroni inequality that
\[
P_{\delta, A}(\max_{i \in H} |Z_i^o| < \sqrt{xp}, \max_{j \in H^c} |Z_j^o| \geq \sqrt{xp})
\]
\[ \begin{align*}
\geq & \, \mathbb{P}_{\delta, A}(\max_{i \in H} |Z_i^0| < \sqrt{x_p}, \max_{j \in H_i^c} |Z_j^0 - (\Omega \delta)_j/\sqrt{\omega_{jj}}| \geq \sqrt{x_p} + C p^{-\xi} \sqrt{\log p}) \\
\geq & \, \alpha \mathbb{P}_{\delta, A}(\bar{E}) + o(1)
\end{align*} \]

Hence, (A.4.14) is proved.

We next compare \( \mathbb{P}(\bar{E}^c, A) \) with \( \mathbb{P}(E^c, A) \). Without loss of generality, we assume that for any \( i \in H \), \( \delta_i > 0 \). By Lemma 4 we also can assume that, on the event \( A \), \( \frac{\Omega \delta_i}{\sqrt{\omega_{ii}}} := \delta_i^o \geq \delta_i - O(p^{\frac{r}{2}}) \) for some \( 2r < a < 1 - 2r \). Note that

\[
\left| \mathbb{P}_{\delta, A}(\max_{i \in H} |Z_i| \geq \sqrt{x_p}) - \mathbb{P}_{\delta, A}(\max_{i \in H} Z_i^0 \geq \sqrt{x_p}) \right| \leq \mathbb{P}_{\delta, A}(\min_{i \in H} Z_i \geq -\sqrt{x_p}) = o(1)
\]

and

\[
\left| \mathbb{P}_{\delta, A}(\max_{i \in H} |Z_i^0| \geq \sqrt{x_p}) - \mathbb{P}_{\delta, A}(\max_{i \in H} Z_i^0 \geq \sqrt{x_p}) \right| \leq \mathbb{P}_{\delta, A}(\min_{i \in H} Z_i^0 \geq -\sqrt{x_p}) = o(1).
\]

It suffices to show that

\[
\mathbb{P}(\max_{i \in H} Z_i \geq \sqrt{x_p}, A) \leq \mathbb{P}(\max_{i \in H} Z_i^0 \geq \sqrt{x_p}, A) + o(1). \tag{A.4.15}
\]

Let \( \mathcal{I}_0 = \{ (i_1, ..., i_m) : \exists 1 \leq l < j \leq m, \text{ such that } |\sigma_{i_l, i_j}| \geq p^{-\xi} \} \) and let \( \mathcal{I} = \{ (i_1, ..., i_m) : 1 \leq i_1 < \cdots < i_m \leq p \} \). We can show that

\[
|\mathcal{I}_0| \leq O\left(p \cdot p^{2\xi} \left(p \atop p^{r-2} \right) \right).
\]

By some simple calculations, for \( \xi < \frac{1}{2}(1 - 2r) \), we have \( |\mathcal{I}_0|/|\mathcal{I}| = o(1) \). Thus, \( \mathbb{P}(\delta \in \mathcal{I}_0) = o(1) \). For \( \delta \in \mathcal{I}_0^c \) with \( 2r < \xi < \frac{1}{2}(1 - 2r) \), using the same arguments from (B.1.3) to (A.4.10), we obtain that

\[
\mathbb{P}_{\delta, A}(\max_{i \in H} Z_i \geq \sqrt{x_p}) = I\{A\} - I\{A\} \prod_{i \in H} \left(1 - \mathbb{P}_{\delta}(Z_i \geq \sqrt{x_p})\right) + o(1).
\]
Similarly, let $I_1 = \{(i_1, \ldots, i_m) : 1 \leq l < j \leq m, \text{ such that } |\omega_{i_l, i_j}| \geq p^{-\xi}\}$, then we can get $|I_1|/|I| = o(1)$, and for $\delta \in I_1^c$,

$$P_{\delta, \Omega} (\max_{i \in H} Z_i^\alpha \geq \sqrt{x_p}) = I\{\mathcal{A}\} - I\{\mathcal{A}\} \prod_{i \in H} \left(1 - P_{\delta}(Z_i^\alpha \geq \sqrt{x_p})\right) + o(1)$$

for any $a$ satisfying $2r < a < 1 - 2r$. By Lemma 5, we have for $\delta \in I_0^c \cap I_1^c$,

$$P_{\delta, \Omega} (\max_{i \in H} Z_i^\alpha \geq \sqrt{x_p}) \geq I\{\mathcal{A}\} - I\{\mathcal{A}\} \prod_{i \in H} \left(1 - P_{\delta}(Z_i^\alpha \geq \sqrt{x_p}) + O(p^{r-a/2})\right) + o(1),$$

which, together with the fact $P(\delta \in I_0) = o(1)$ and $P(\delta \in I_1) = o(1)$, proves (A.4.15). Proposition 5 (i) is proved by (A.4.11), (A.4.13) and (A.4.15).

**Proof of Proposition 5 (ii).**

Define $M'_{\Omega} = \max_{i \in A^c} |(\Omega Z)_i|/\sqrt{\omega_i}|$, $M''_{\Omega} = \max_{i \in A} |(\Omega Z)_i|/\sqrt{\omega_i}|$, $M'_0 = \max_{i \in A^c} |(\Omega^2 Z)_i|$, and $M''_0 = \max_{i \in A} |(\Omega^2 Z)_i|$. By the definition of $\mathcal{A}$, we see that $M'_{\Omega}$ and $M''_{\Omega}$ are independent. Hence we have

$$P\left(M_{\Omega} \geq \sqrt{x_p}\right) = P\left(M'_{\Omega} \geq \sqrt{x_p}\right) + P\left(M''_{\Omega} < \sqrt{x_p}\right) P\left(M'_{\Omega} \geq \sqrt{x_p}\right)$$

$$= P\left(M''_{\Omega} \geq \sqrt{x_p}\right) + P\left(M''_{\Omega} < \sqrt{x_p}\right) P\left(M''_{\Omega} \geq \sqrt{x_p}\right).$$

We next prove that

$$P\left(M'_{\Omega} \geq \sqrt{x_p}\right) \geq P\left(M'_{\Omega} \geq \sqrt{x_p}\right) + o(1). \quad \text{(A.4.17)}$$

From the proof of Proposition 5(i), we can assume that $\max_{1 \leq i \leq p} |\delta_i| \leq 6\sqrt{\log p}$. Set $A = \{\max_{i \in A^c} (\max_j |a_{ij}\delta_i|) < \sqrt{2\beta_0 \log p}\}$ for some $\beta_0 < 1$ being sufficiently
close to 1. Because \( \omega_{ii} = \sum_{j=1}^{p} a_{ij}^2 \) and \( \sum_{i=1}^{p} a_{ij}^2 \geq \max_j a_{ij}^2 + \epsilon_1^2 \) for \( i \in \mathcal{A}^c \) with some \( \epsilon_1 > 0 \), we have by Lemma 4,

\[
P(\max_{i \in \mathcal{A}^c} \left| \frac{(\Omega \delta)_i}{\sqrt{\omega_{ii}}} \right| \geq (1 + 2\epsilon_2) \max_{i \in \mathcal{A}^c} (\max_j |a_{ij}\delta_i|) + o(1)) \rightarrow 1,
\]

for some constant \( \epsilon_2 > 0 \). Thus we have

\[
P(M'_{\Omega} \in R_\alpha, \mathcal{A}^c) \geq P(\max_{i \in \mathcal{A}^c} Z_{i_0}^o \geq \sqrt{x_p}) \geq P(\max_{i \in \mathcal{A}^c} Z_{i_0}^o \geq \sqrt{x_p} - (1 + \epsilon_2)\sqrt{2\beta_0 \log p} + o(1))
\]

where \( i_0 = \arg \max_{i \in \mathcal{A}} (\max_j |a_{ij}\delta_i|) \). We next consider \( P(M'_{\Omega} \in R_\alpha, \mathcal{A}) \) and \( P(M'_{\Omega^c} \in R_\alpha, \mathcal{A}) \). Let \( Z_i^* = (\Omega^1_\delta Z)_i \) and \( \delta_i^* = (\Omega^1_\delta \delta)_i \). Then

\[
P(M'_{\Omega^c} \in R_\alpha)
\]

It is easy to see that \( \text{Card}(H_1) \leq p^{5r} \). We next show that \( |(\Omega^1_\delta \delta)_i| \leq \sqrt{2\beta_0 \log p} + o(1) \)

uniformly for \( i \in H_1 \) with probability tending to one. As in the proof of Lemma 4,
we modify the definition of $\mathcal{I}_0$ to

$$
\mathcal{I}_0 = \left\{ 1 \leq i_1 < \ldots < i_m \leq p : \text{there exist some } i_k \text{ and } i_j \neq i_k \text{ such that } |a_{i,j}| > |a_{i,(p^r)}| \text{ for some } i \in H_{1k} \right\}.
$$

It’s easy to show that $\text{Card}(\mathcal{I}_0) \leq p^{1+9r} C_p^{m-2} = o(1)C_p^m$ since $m = p^r$ and $r$ is arbitrarily small. Note that, for $i \in H_{1k}$ and $(i_1, \ldots, i_m) \in \mathcal{I}_0^c$,

$$(\Omega^1_{\frac{1}{2}} \delta)_i = \sum_{j=1}^m a_{i,i_j} \delta_{i_j} + \sum_{j \neq k} a_{i,i_j} \delta_{i_j} = a_{i,i_k} \delta_{i_k} + O(p^{-r} \sqrt{\log p}).$$

This implies that on $A$, $P_{\delta, A}(\max_{i \in H_{1} \cap A^c} |Z_i^*| \geq \sqrt{x_p}) \rightarrow 1$, which in turn yields that for $\text{supp}(\delta) \in \mathcal{I}_0^c$,

$$P_{\delta, A}(\max_{i \in H_{1} \cap A^c} |Z_i^*| \geq \sqrt{x_p}) \leq \text{Card}(H_1)P(|N(0,1)| \geq \sqrt{x_p} - \sqrt{2 \beta_0 \log p} + O(p^{-r} \sqrt{\log p}) + o(1) = o(1).$$

For $i \in H_2$, we have $(\Omega^1_{\frac{1}{2}} \delta)_i = \sum_{j=1}^m a_{i,i_j} \delta_{i_j} = O(p^{-r} \sqrt{\log p})$. Thus,

$$P_{\delta, A}(\max_{j \in H_{2} \cap A^c} |Z_j^*| \geq \sqrt{x_p}) = P(\max_{j \in H_{2} \cap A^c} |Y_j| \geq \sqrt{x_p}) I\{A\} + o(1) =: \alpha_p I\{A\} + o(1),$$

where $Y_1, \ldots, Y_p$ are i.i.d. $N(0,1)$ random variables. Let $E^* = \{\max_{i \in H \cap A^c} |Z_i^*| < \sqrt{x_p}\}$. We have

$$P_{\delta, A}(M_{\Omega \frac{1}{2}}^* \in R_\alpha) \leq \alpha_p + (1 - \alpha_p)P_{\delta, A}(E^{ec}) + o(1). \quad (A.4.19)$$

Without loss of generality, we assume that for any $i \in H$, $\delta_i > 0$. By Lemma 4, we have $\delta_i^p \geq \delta_i^* \sim O(p^{r-a/2})$ with some $2r < a < 1 - 2r$. Similarly as (A.4.14) and (A.4.15), it follows from Bonferroni’s inequality that

$$P_{\delta, A}(M_{\Omega} \in R_\alpha) \geq \alpha_p + (1 - \alpha_p)P_{\delta, A}(\max_{i \in H \cap A^c} Z_i^* \geq \sqrt{x_p}) + o(1). \quad (A.4.20)$$
By Lemma 4 and $\delta_i > 0$, we have $P\left( \min_{i \in H} \delta_i^* \geq 0 - O(p^{r-a/2}) \right) \to 1$. Hence

$$P_{\delta, A}\left( \min_{i \in H} Z_i^* \leq -\sqrt{x_p} \right) \leq P\left( \min_{1 \leq i \leq m} Y_i \leq -\sqrt{x_p} + O(p^{r-a/2}) \right) = o(1).$$

This implies that

$$\left| P_{\delta, A}\left( \max_{i \in H \cap A^c} |Z_i^*| \geq \sqrt{x_p} \right) - P_{\delta, A}\left( \max_{i \in H \cap A^c} Z_i^* \geq \sqrt{x_p} \right) \right| \leq P_{\delta, A}(\min_{i \in H} Z_i^* \leq -\sqrt{x_p}) = o(1). \tag{A.4.21}$$

By (A.4.18)-(A.4.21), (A.4.17) is proved. Hence we have

$$P\left( M_{\Omega} \geq \sqrt{x_p} \right) \geq P\left( M_{\Omega^p} \geq \sqrt{x_p} \right) + o(1)$$

and this proves Proposition 5(ii).

**Proof of Proposition 6 (i).** We first prove that for $\beta \geq (1 - \sqrt{r})^2 + \epsilon$,

$$P\left( M_I \in R_\alpha \right) \to 1. \tag{A.4.22}$$

Let $(Z_1, \ldots, Z_p)'$ be a multivariate normal random vector with $p^r$-sparse ($r < \frac{1}{2}$) mean $\sqrt{n} \delta = \sqrt{n}(\delta_1, \ldots, \delta_p)'$ and covariance matrix $\Sigma = (\sigma_{ij})$. We assume that the diagonal $\sigma_{ii} = 1$ for $1 \leq i \leq p$, and $\Sigma$ satisfies condition (C1) and (C2). Then it suffices to show $P\left( \max_{1 \leq i \leq p} |Z_i| \geq \sqrt{x_p} \right) \to 1$, where $x_p = 2 \log p - \log \log p + q_\alpha$, and $q_\alpha$ is the $1 - \alpha$ quantile of $\exp(-\frac{1}{\sqrt{\pi}} \exp(-x/2))$. Note that

$$P\left( \max_{1 \leq i \leq p} |Z_i| \geq \sqrt{x_p} \right) \geq P\left( \max_{i \in H}(\text{sign}(\delta_i)Z_i) \geq \sqrt{x_p} \right),$$

where $H = \{i: \delta_i \neq 0, 1 \leq i \leq p\}$. Thus,

$$P\left( \max_{1 \leq i \leq p} |Z_i| \geq \sqrt{x_p} \right) \geq P\left( \max_{i \in H} U_i \geq \sqrt{x_p} - a \right),$$

$$160$$
where $a = \sqrt{2\beta \log p}$ for $\beta \geq (1 - \sqrt{r})^2 + \epsilon$ and $U_i$, $1 \leq i \leq p$, are $N(0, 1)$ random variables with covariance matrix $\Sigma$. Because

$$\sqrt{x_p} - a = \sqrt{2\log p - \log \log p + q_\alpha} - \sqrt{2\beta \log p} \leq (\sqrt{2} - \sqrt{2\beta})\sqrt{\log p} < \sqrt{2\log p - \log \log p^r - M}$$

for any $M \in \mathbb{R}$, we have by Lemma 1

$$\mathbb{P}(\max_{i \in H} U_i \geq \sqrt{x_p} - a) \geq \mathbb{P}(\max_{i \in H} U_i \geq \sqrt{2\log p^r - \log \log p^r} - M) \to 1 - \exp\left(-\frac{1}{2\sqrt{\pi}}\exp(M/2)\right),$$

for arbitrary large $M$. By letting $M \to \infty$, we have $\mathbb{P}(\max_{i \in H} U_i \geq \sqrt{x_p} - a) \to 1$. Thus $\mathbb{P}(M_I \in R_\alpha) \to 1$ for any $\beta \geq (1 - \sqrt{r})^2 + \epsilon$. It remains to prove that for $\beta < (1 - \sqrt{r})^2$,

$$\lim_{p \to \infty} \mathbb{P}(M_I \in R_\alpha) \leq \alpha.$$ 

By noting that $\mathbb{P}(M_I \in R_\alpha) \leq \mathbb{P}(\max_{i \in H} Z_i^2 \geq x_p) + \mathbb{P}(\max_{i \in H^c} Z_i^2 \geq x_p)$, it suffices to show that for $\beta < (1 - \sqrt{r})^2$,

$$\lim_{p \to \infty} \mathbb{P}(\max_{i \in H^c} Z_i^2 \geq x_p) \leq \alpha, \tag{A.4.23}$$

and

$$\mathbb{P}(\max_{i \in H} Z_i^2 \geq x_p) \to 0. \tag{A.4.24}$$

Note that $\delta_i = 0$ for $i \in H^c$. It follows from Lemma 1 that (A.4.23) holds. For (A.4.24), we have

$$\mathbb{P}(\max_{i \in H} Z_i^2 \geq x_p) \leq p^r \mathbb{P}(|N(0, 1)| \geq \sqrt{x_p - \sqrt{2\beta \log p}}) \leq C p^{r-(1-\sqrt{3})^2}(\log p)^2,$$
where $C$ is a positive constant. Because $\beta < (1 - \sqrt{r})^2$, we have (A.4.24). Combing (A.4.23) and (A.4.24), Proposition 6 (i) is proved.

Proof of Proposition 6 (ii). To prove Proposition 6 (ii), we only need to prove the following lemma.

Lemma 14. Consider $H_1'$: $\delta$ has $m = p^r$, $r < 1/4$ nonzero coordinates with $\sqrt{\omega_{ii}} |\delta_i| \geq \sqrt{\frac{2\beta \log p}{n}}$, where $\beta_* > 0$ if $\delta_i \neq 0$. The nonzero locations $l_1 < l_2 < \cdots < l_m$ are randomly and uniformly drawn from $\{1, 2, \ldots, p\}$. If $\beta_* \geq (1 - \sqrt{r})^2 + \epsilon$ for some $\epsilon > 0$, then

$$\mathbb{P}\left( M_\Omega \in R_\alpha \right) \to 1.$$

Note that $\sqrt{\omega_{ii}} |\delta_i| = \frac{|\delta_i|}{\sqrt{\sigma_{ii}}} \cdot \sqrt{\sigma_{ii} \omega_{ii}} = \sqrt{\frac{2\beta \log p}{n}} \cdot \sqrt{\sigma_{ii} \omega_{ii}} = \sqrt{\frac{2\beta_* \log p}{n}}$, where $\beta_* = \beta \sigma_{ii} \omega_{ii}$ and $\beta \geq (1 - \sqrt{r})^2/(\min_{1 \leq i \leq p} \sigma_{ii} \omega_{ii}) + \epsilon$, we have $\beta_* \geq (1 - \sqrt{r})^2 + \epsilon$. Thus by Lemma 14, we have $\mathbb{P}\left( M_\Omega \in R_\alpha \right) \to 1$.

We next prove Lemma 14. As the proof of (A.4.22), it suffices to show that

$$\mathbb{P}\left( \min_{i \in H} \frac{|(\Omega \delta)_i|}{\sqrt{\omega_{ii}}} \geq \sqrt{\frac{2\beta \log p}{n}} \right) \to 1$$

for some $\beta \geq (1 - \sqrt{r})^2 + \epsilon$. This follows from Lemma 4 immediately.

Proof of Proposition 7 (i). Let $(Z_1, \ldots, Z_p)'$ be a multivariate normal random vector with mean $\delta^o = \sqrt{n} \Omega^\frac{1}{2} \delta$ and covariance matrix $I_p$. Let $H = \{i_1, \ldots, i_m\}$ be
the support of $\delta$. Define

$$H_1 = \bigcup_{k=1}^{m} \{ 1 \leq j \leq p : |a_{i,j}| > |a_{i,(p^r)}| \} =: \bigcup_{k=1}^{m} H_{1k},$$

where $r_1 > 0$ satisfies $\sqrt{r_1 + r} < 1 - \sqrt{\beta}$. We have

$$\mathbb{P}(\max_{1 \leq i \leq p} |Z_i| \geq \sqrt{xp}) \leq \mathbb{P}(\max_{i \in H_1} |Z_i| \geq \sqrt{xp}) + \mathbb{P}(\max_{i \in H_1^c} |Z_i| \geq \sqrt{xp}).$$

Thus it suffices to show

$$\mathbb{P}(\max_{i \in H_1} |Z_i| \geq \sqrt{xp}) \to 0, \quad (A.4.25)$$

and

$$\lim_{p \to \infty} \mathbb{P}(\max_{i \in H_1^c} |Z_i| \geq \sqrt{xp}) \leq \alpha. \quad (A.4.26)$$

Define

$$\mathcal{I}_1 = \left\{ 1 \leq i_1 < \ldots < i_m \leq p : \text{there exist some } i_k \text{ and } i_j \neq i_k \right\},$$

where $a > 2r$ satisfies $3r + r_1 + a < 1$. Then $\text{Card}(\mathcal{I}_1) \leq p^{1+r+r_1+a} C_p^{m-2} = o(1) C_p^m$.

It follows that for $i \in H_{1k}$ and $(i_1, \ldots, i_m) \notin \mathcal{I}_1$,

$$\sqrt{n}|(\Omega^* \delta)_{i}| = \sqrt{n} \left| \sum_{k=1}^{p} a_{i,k} \delta_k \right| = \sqrt{n} \left| a_{i,i_k} \delta_{i_k} + \sum_{j \in H,j \neq i_k} a_{i,j} \delta_j \right| \leq \sqrt{2\beta \log p} + O(p^{r-a/2} (\log p)^{1/2}).$$

Thus

$$\mathbb{P}\left( \max_{i \in H_1} |Z_i| \geq \sqrt{xp} \right) \leq C p^{r+r_1-(1-\sqrt{\beta})^2 (\log p)^2} + o(1) = o(1)$$

and this proves (A.4.25).
Set $H_{2k} = \{1 \leq j \leq p : |a_{ik(p^1)}| \geq |a_{ik,j}| \geq |a_{ik(p^2)}|\}$ and $H_2 = \cup_{k=1}^m H_{2k}$. Let

$$I_2 = \{1 \leq i_1 < \ldots < i_m \leq p : \text{there exist some } i_k \text{ and } i_j \neq i_k \text{ such that } |a_{i,i_j}| > |a_{i,(p^j)}| \text{ for some } i \in H_{2k}\},$$

where $a > 2r$ and $r_2 > 2r$ satisfy $3r + r_2 + a < 1$. We have $\text{Card}(I_2) \leq p^{1+r+r_2+a}C_p^{m-2} = o(1)C_p^m$. For $i \in H_{2k}$ and $(i_1, \ldots, i_m) \notin I_2$, we have

$$\sqrt{n_i}(\Omega^2 \delta_i) = \sqrt{n} \sum_{k=1}^p a_{ik} \delta_k = \sqrt{n_i} \sum_{j \in H,j \neq i_k} a_{ij} \delta_j = O(p^{-r_1/2}(\log p)^{1/2}) + O(p^{-a/2}(\log p)^{1/2}).$$

Hence

$$P\left(\max_{i \in H_2} |Z_i| \geq \sqrt{x_p}\right) \leq C\text{Card}(H_2)p^{-1} + o(1) = o(1). \quad (A.4.27)$$

For $i \in H_3 := (H_1 \cup H_2)^c$, we have $\sqrt{n_i}(\Omega^2 \delta_i) = \sqrt{n} \sum_{k=1}^p a_{ik} \delta_k = O(p^{-r_2/2}(\log p)^{1/2})$. This, together with Lemma 1, implies that

$$\lim_{p \to \infty} P\left(\max_{i \in H_3} |Z_i| \geq \sqrt{x_p}\right) \leq \alpha. \quad (A.4.28)$$

By (A.4.27) and (A.4.28), we prove (A.4.26) and complete the proof of Proposition 7 (i). \hfill \Box

**Proof of Proposition 7 (ii).** It suffices to verify the condition in Lemma 14. Note that for $i \in H$,

$$\sqrt{\omega_i || \delta_i ||} = \max_{1 \leq j \leq p} |a_{ij}| || \delta_i || \cdot \sqrt{\omega_i / \max_{1 \leq j \leq p} a^2_{ij}} = \sqrt{2/3} \log p n \cdot \sqrt{\omega_i} / \max_{1 \leq j \leq p} a^2_{ij}$$

164
\[ = \sqrt{\frac{2\beta^* \log p}{n}}, \]

where \( \beta^* = \beta \omega_{ii} / \max_{1 \leq j \leq p} a_{ij}^2 \). If \( \beta \geq (1 - \sqrt{r})^2 / (\min_{1 \leq i \leq p} (\omega_{ii} / \max_{1 \leq j \leq p} a_{ij}^2)) + \varepsilon \), then we have \( \beta^* \geq (1 - \sqrt{r})^2 + \varepsilon \). Thus, by Lemma 14, we have \( \mathbb{P}(M_\Omega \in R_\alpha) \to 1 \).

and the proof of Proposition 7 (ii) is complete. \( \square \)
Appendix B

Testing High Dimensional Covariance Matrices

B.1 Proof of Lemmas 10-11

Proof of Lemma 10. We only prove (3.5.1) in Lemma 10 because the proof of (3.5.2) is similar. Without loss of generality, we assume that \( \mathbf{E} \mathbf{X} = 0 \) and \( \text{Var}(X_i) = 1 \) for \( 1 \leq i \leq p \). Let

\[
\tilde{\theta}_{ij} = \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ X_{ki} X_{kj} - \tilde{\sigma}_{ij1} \right] \]

with \( \tilde{\sigma}_{ij1} = \frac{1}{n_1} \sum_{k=1}^{n_1} X_{ki} X_{kj} \).

By the proof of Lemma 2 in Cai and Liu (2011), we have for any \( M > 0 \), there exists a constant \( C \) such that

\[
P \left( \max_{i,j} |\hat{\theta}_{ij1} - \tilde{\theta}_{ij1}| \geq C \sqrt{\log p/n} \right) = O(p^{-M} + n^{-\epsilon/8}). 
\]  

(B.1.1)
Write
\[ \tilde{\theta}_{ij} - \theta_{ij} = \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ (X_{ki}X_{kj})^2 - \mathbb{E}(X_{ki}X_{kj})^2 \right] - \tilde{\sigma}_{ij}^2 + \sigma_{ij}^2. \]

We can see that
\[ P\left( \max_{i,j} |\tilde{\sigma}_{ij} - \sigma_{ij}| \geq C \sqrt{\log p/n} \right) = O(p^{-M} + n^{-\epsilon/8}). \]  \hfill (B.1.2)

We first assume that (C2) holds. It suffices to show that
\[ P\left( \max_{i,j} \left| \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ (X_{ki}X_{kj})^2 - \mathbb{E}(X_{ki}X_{kj})^2 \right] \right| \geq C \sqrt{\frac{\log p}{n}} \right) = O(p^{-M}). \]  \hfill (B.1.3)

Define
\[ \hat{X}_{kj} = X_{kj} I \{|X_{kj}| \leq \tau \sqrt{\log (p + n)} \}, \]

where \( \tau \) is sufficiently large. We have
\[ |\mathbb{E}(X_{ki}X_{kj})^2 - \mathbb{E}(X_{ki}\hat{X}_{kj})^2| \leq C \left( \mathbb{E}X_{kj}^4 I \{|X_{kj}| \geq \tau \sqrt{\log (p + n)} \} \right)^{1/2} \]
\[ \leq C(n + p)^{-\tau^2 \eta/2} \left( \mathbb{E}X_{kj}^4 \exp \left( 2^{-1} \eta X_{kj}^2 \right) \right)^{1/2} \]
\[ \leq C(n + p)^{-\tau^2 \eta/2}, \]  \hfill (B.1.4)

where \( C \) does not depend on \( n, p \). Thus it follows that
\[ P\left( \max_{i,j} \left| \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ (X_{ki}X_{kj})^2 - \mathbb{E}(X_{ki}X_{kj})^2 \right] \right| \geq C \sqrt{\frac{\log p}{n}} \right) \]
\[ \leq P\left( \max_{i,j} \left| \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ (X_{ki}\hat{X}_{kj})^2 - \mathbb{E}(X_{ki}\hat{X}_{kj})^2 \right] \right| \geq 2^{-1} C \sqrt{\frac{\log p}{n}} \right) \]
\[ + np P\left( |X_{11}| \geq \tau \sqrt{\log (p + n)} \right). \]

Note that
\[ np P\left( |X_{11}| \geq \tau \sqrt{\log (p + n)} \right) \leq np(n + p)^{-\tau^2 \eta} \mathbb{E} \exp \left( \eta X_{11}^2 \right) = O(p^{-M}). \]
Let \( t = \eta (8\tau^2)^{-1} \sqrt{\log p/n_1} \) and \( \hat{Z}_{kij} = (X_{ki}\hat{X}_{kj})^2 - E(X_{ki}\hat{X}_{kj})^2 \). Then we have

\[
P\left( \frac{1}{n_1} \sum_{k=1}^{n_1} \left[ (X_{ki}\hat{X}_{kj})^2 - E(X_{ki}\hat{X}_{kj})^2 \right] \geq C \sqrt{\frac{\log p}{n}} \right)
\leq \exp \left( -Ct \sqrt{n_1 \log p} \prod_{k=1}^{n_1} E \exp \left( t \hat{Z}_{kij} \right) \right)
\leq \exp \left( -Ct \sqrt{n_1 \log p} \prod_{k=1}^{n_1} \left( 1 + E t^2 \hat{Z}_{kij}^2 \exp \left( t |\hat{Z}_{kij}| \right) \right) \right)
\leq \exp \left( -C \eta (8\tau^2)^{-1} \log p + c_{\tau, \eta} \log p \right)
\leq Cp^{-M},
\]

where \( c_{\tau, \eta} \) is a positive constant depending only on \( \tau \) and \( \eta \). Similarly, we can show that

\[
P\left( \frac{1}{n_1} \sum_{k=1}^{n_1} \left| (X_{ki}\hat{X}_{kj})^2 - E(X_{ki}\hat{X}_{kj})^2 \right| \right) \leq -C \sqrt{\frac{\log p}{n}} \leq Cp^{-M}.
\]

Thus (B.1.3) is proved.

It remains to prove the lemma under (C2*). Define

\[ Y_{ij,k} = (X_{ki}X_{kj})^2, \quad \hat{Y}_{ij,k} = Y_{ij,k} I\{|Y_{ij,k}| \leq n/(\log p)^8\}. \]

Then as in (B.1.4), we can show that \(|EY_{ij,k} - E\hat{Y}_{ij,k}| \leq Cn^{-\gamma_0/4}\). Note that \( \varepsilon_n = (\log p)^{-1} \). It follows that

\[
P\left( \max_{i,j} \left| \sum_{k=1}^{n_1} (Y_{ij,k} - EY_{ij,k}) \right| \geq \frac{n\varepsilon_n}{\log p} \right)
\leq P\left( \max_{i,j} \left| \sum_{k=1}^{n_1} (\hat{Y}_{ij,k} - E\hat{Y}_{ij,k}) \right| \geq 2^{-1} \frac{n\varepsilon_n}{\log p} \right) + P\left( \max_{i,j,k} |Y_{ij,k}| \geq \frac{n}{(\log p)^8} \right)
\leq Cp^2 \exp(-C(\log p)^4) + Cn^{-\epsilon_0/8}, \quad (B.1.5)
\]
where the last inequality follows from Bernstein’s inequality and (C2∗). The lemma is proved.  

\textbf{Proof of Lemma 11.} Without loss of generality, we assume that \( \mu_1 = 0 \) and \( \mu_2 = 0 \). Set

\[
Z_{ij,k} = \frac{n_2}{n_1} (X_{ki}X_{kj} - \sigma_{ij1}), \quad 1 \leq k \leq n_1,
\]

\[
Z_{ij,k} = -(Y_{ki}Y_{kj} - \sigma_{ij2}), \quad n_1 + 1 \leq k \leq n_1 + n_2.
\]

By (B.1.2), (B.1.3) and (B.1.5), we obtain for any \( M > 0 \),

\[
P\left( \max_{ij} \left| \frac{1}{n_1} \sum_{k=1}^{n_1} Z_{ij,k}^2 - \frac{n_2^2}{n_1^2} \theta_{ij1} \right| \geq C \frac{\varepsilon n}{\log p} \right) = O(p^{-M} + n^{-\epsilon/8}),
\]

\[
P\left( \max_{ij} \left| \frac{1}{n_2} \sum_{k=n_1+1}^{n_2} Z_{ij,k}^2 - \theta_{ij2} \right| \geq C \frac{\varepsilon n}{\log p} \right) = O(p^{-M} + n^{-\epsilon/8}). \quad (B.1.6)
\]

We can write

\[
\frac{(\bar{\sigma}_{ij1} - \sigma_{ij2} - \sigma_{ij1} + \sigma_{ij2})^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} = \left( \frac{\sum_{k=1}^{n_1+n_2} Z_{ij,k}}{\sum_{k=1}^{n_1+n_2} Z_{ij,k}^2} \right)^2 \times \frac{\sum_{k=1}^{n_1+n_2} Z_{ij,k}^2}{n_2^2 \theta_{ij1}/n_1 + n_2 \theta_{ij2}}.
\]

By the self-normalized large deviation theorem for independent random variables (Theorem 1, Jing, Shao and Wang (2003)), we can get

\[
\max_{1 \leq i \leq j \leq p} P\left( \frac{\left( \sum_{k=1}^{n_1+n_2} Z_{ij,k}^2 \right)^2}{\sum_{k=1}^{n_1+n_2} Z_{ij,k}^2} \geq x^2 \right) \leq C(1 - \Phi(x))
\]

uniformly for \( 0 \leq x \leq (8 \log p)^{1/2} \). This, together with (B.1.6), proves the lemma.  

\[\qed\]
B.2 Additional simulation results

In this section we present additional simulation results comparing the numerical performance of the proposed test with that of other tests, particularly in the non-Gaussian setting and the small sample size setting.

B.2.1 Simulation results for non-Gaussian distributions

For each model, we generate two independent random samples \( \{ X_k \}_{k=1}^{n_1} \) and \( \{ Y_l \}_{l=1}^{n_2} \) from multivariate models \( X_k = \Gamma_1 Z_k^{(1)} \) and \( Y_k = \Gamma_2 Z_k^{(2)} \), with \( \Gamma_1 \Gamma_1' = \Sigma_1, \Gamma_2 \Gamma_2' = \Sigma_2 \). We consider the following four types of distribution of \( Z_k^{(i)}, i = 1, 2 \).

(I) The components of \( Z_k^{(i)} = (Z_{k1}, ..., Z_{kp})' \) are i.i.d. Gamma(10,1) random variables.

(II) \( Z_k^{(i)} = (Z_{k1}, ..., Z_{kp})' \) are generated by letting each component be an standard normal \( (N(0,1)) \) variable with probability 0.9 and an standard exponential random variable \( (\exp(1)) \) with probability 0.1, where the components are independent.

(III) The components of \( Z_k^{(i)} = (Z_{k1}, ..., Z_{kp})' \) are i.i.d. \( t \) distributed random variables with 12 degrees of freedom.

The simulation results are summarized in Tables B.1-B.3.
<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Empirical size</td>
<td>Empirical size</td>
<td>Empirical size</td>
<td>Empirical size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.048</td>
<td>0.048</td>
<td>0.049</td>
<td>0.049</td>
<td>0.051</td>
<td>0.043</td>
<td>0.050</td>
<td>0.046</td>
<td>0.046</td>
<td>0.043</td>
<td>0.048</td>
<td>0.039</td>
<td>0.037</td>
<td>0.044</td>
<td>0.036</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>likelihood</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Schott</td>
<td>0.086</td>
<td>0.073</td>
<td>0.072</td>
<td>0.065</td>
<td>0.063</td>
<td>0.090</td>
<td>0.085</td>
<td>0.076</td>
<td>0.067</td>
<td>0.068</td>
<td>0.095</td>
<td>0.089</td>
<td>0.082</td>
<td>0.079</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>Chen</td>
<td>0.070</td>
<td>0.056</td>
<td>0.054</td>
<td>0.049</td>
<td>0.049</td>
<td>0.067</td>
<td>0.058</td>
<td>0.055</td>
<td>0.047</td>
<td>0.048</td>
<td>0.064</td>
<td>0.058</td>
<td>0.055</td>
<td>0.051</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>Srivastava</td>
<td>0.062</td>
<td>0.051</td>
<td>0.053</td>
<td>0.046</td>
<td>0.044</td>
<td>0.057</td>
<td>0.049</td>
<td>0.046</td>
<td>0.047</td>
<td>0.043</td>
<td>0.051</td>
<td>0.054</td>
<td>0.048</td>
<td>0.047</td>
<td>0.042</td>
</tr>
<tr>
<td>100</td>
<td>0.042</td>
<td>0.043</td>
<td>0.043</td>
<td>0.040</td>
<td>0.035</td>
<td>0.040</td>
<td>0.044</td>
<td>0.042</td>
<td>0.044</td>
<td>0.042</td>
<td>0.039</td>
<td>0.034</td>
<td>0.042</td>
<td>0.035</td>
<td>0.034</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>likelihood</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Schott</td>
<td>0.082</td>
<td>0.073</td>
<td>0.077</td>
<td>0.073</td>
<td>0.074</td>
<td>0.084</td>
<td>0.079</td>
<td>0.076</td>
<td>0.076</td>
<td>0.075</td>
<td>0.092</td>
<td>0.087</td>
<td>0.086</td>
<td>0.083</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Chen</td>
<td>0.067</td>
<td>0.056</td>
<td>0.058</td>
<td>0.055</td>
<td>0.053</td>
<td>0.059</td>
<td>0.055</td>
<td>0.054</td>
<td>0.053</td>
<td>0.053</td>
<td>0.059</td>
<td>0.059</td>
<td>0.058</td>
<td>0.052</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>Srivastava</td>
<td>0.056</td>
<td>0.057</td>
<td>0.056</td>
<td>0.048</td>
<td>0.048</td>
<td>0.060</td>
<td>0.053</td>
<td>0.058</td>
<td>0.050</td>
<td>0.047</td>
<td>0.058</td>
<td>0.059</td>
<td>0.056</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>Empirical power</td>
<td>0.836</td>
<td>0.687</td>
<td>0.428</td>
<td>0.353</td>
<td>0.335</td>
<td>0.831</td>
<td>0.522</td>
<td>0.558</td>
<td>0.433</td>
<td>0.223</td>
<td>0.867</td>
<td>0.720</td>
<td>0.445</td>
<td>0.368</td>
<td>0.346</td>
</tr>
<tr>
<td></td>
<td>Schott</td>
<td>0.389</td>
<td>0.214</td>
<td>0.145</td>
<td>0.111</td>
<td>0.101</td>
<td>0.437</td>
<td>0.149</td>
<td>0.151</td>
<td>0.113</td>
<td>0.091</td>
<td>0.442</td>
<td>0.233</td>
<td>0.145</td>
<td>0.117</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>Chen</td>
<td>0.300</td>
<td>0.144</td>
<td>0.091</td>
<td>0.064</td>
<td>0.060</td>
<td>0.346</td>
<td>0.090</td>
<td>0.093</td>
<td>0.068</td>
<td>0.053</td>
<td>0.345</td>
<td>0.160</td>
<td>0.089</td>
<td>0.064</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>Srivastava</td>
<td>0.286</td>
<td>0.112</td>
<td>0.064</td>
<td>0.053</td>
<td>0.040</td>
<td>0.374</td>
<td>0.080</td>
<td>0.066</td>
<td>0.053</td>
<td>0.040</td>
<td>0.369</td>
<td>0.152</td>
<td>0.062</td>
<td>0.050</td>
<td>0.042</td>
</tr>
<tr>
<td>100</td>
<td>0.099</td>
<td>0.985</td>
<td>0.899</td>
<td>0.865</td>
<td>0.887</td>
<td>0.996</td>
<td>0.915</td>
<td>0.967</td>
<td>0.930</td>
<td>0.707</td>
<td>0.999</td>
<td>0.989</td>
<td>0.912</td>
<td>0.877</td>
<td>0.898</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>Schott</td>
<td>0.658</td>
<td>0.330</td>
<td>0.175</td>
<td>0.135</td>
<td>0.118</td>
<td>0.708</td>
<td>0.214</td>
<td>0.195</td>
<td>0.144</td>
<td>0.105</td>
<td>0.728</td>
<td>0.366</td>
<td>0.180</td>
<td>0.137</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>Chen</td>
<td>0.566</td>
<td>0.236</td>
<td>0.106</td>
<td>0.080</td>
<td>0.070</td>
<td>0.614</td>
<td>0.141</td>
<td>0.127</td>
<td>0.086</td>
<td>0.057</td>
<td>0.630</td>
<td>0.266</td>
<td>0.112</td>
<td>0.083</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>Srivastava</td>
<td>0.576</td>
<td>0.237</td>
<td>0.082</td>
<td>0.059</td>
<td>0.057</td>
<td>0.723</td>
<td>0.127</td>
<td>0.103</td>
<td>0.065</td>
<td>0.043</td>
<td>0.740</td>
<td>0.316</td>
<td>0.085</td>
<td>0.061</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Table B.1: $\Gamma(10,1)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications.
<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha$</td>
<td>Empirical size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.046</td>
<td>0.045</td>
<td>0.047</td>
<td>0.041</td>
<td>0.044</td>
<td>0.044</td>
<td>0.042</td>
<td>0.039</td>
<td>0.044</td>
<td>0.048</td>
<td>0.045</td>
<td>0.045</td>
<td>0.040</td>
<td>0.038</td>
<td>0.039</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>0.101</td>
<td>101</td>
<td>0.097</td>
<td>0.094</td>
<td>0.089</td>
<td>0.115</td>
<td>0.106</td>
<td>0.102</td>
<td>0.107</td>
<td>0.095</td>
<td>0.131</td>
<td>0.127</td>
<td>0.129</td>
<td>0.119</td>
<td>0.115</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>0.066</td>
<td>0.059</td>
<td>0.055</td>
<td>0.054</td>
<td>0.054</td>
<td>0.061</td>
<td>0.056</td>
<td>0.053</td>
<td>0.058</td>
<td>0.051</td>
<td>0.057</td>
<td>0.056</td>
<td>0.057</td>
<td>0.053</td>
<td>0.050</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>0.061</td>
<td>0.055</td>
<td>0.055</td>
<td>0.051</td>
<td>0.041</td>
<td>0.058</td>
<td>0.053</td>
<td>0.053</td>
<td>0.052</td>
<td>0.041</td>
<td>0.075</td>
<td>0.057</td>
<td>0.058</td>
<td>0.050</td>
<td>0.045</td>
<td>0.046</td>
</tr>
<tr>
<td>100</td>
<td>0.041</td>
<td>0.045</td>
<td>0.040</td>
<td>0.036</td>
<td>0.037</td>
<td>0.038</td>
<td>0.037</td>
<td>0.042</td>
<td>0.033</td>
<td>0.036</td>
<td>0.043</td>
<td>0.034</td>
<td>0.038</td>
<td>0.037</td>
<td>0.034</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>0.111</td>
<td>101</td>
<td>0.099</td>
<td>0.096</td>
<td>0.095</td>
<td>0.117</td>
<td>0.102</td>
<td>0.103</td>
<td>0.107</td>
<td>0.104</td>
<td>0.144</td>
<td>0.126</td>
<td>0.121</td>
<td>0.120</td>
<td>0.115</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>0.060</td>
<td>0.061</td>
<td>0.056</td>
<td>0.054</td>
<td>0.053</td>
<td>0.056</td>
<td>0.058</td>
<td>0.059</td>
<td>0.057</td>
<td>0.046</td>
<td>0.057</td>
<td>0.061</td>
<td>0.053</td>
<td>0.053</td>
<td>0.046</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td>0.053</td>
<td>0.053</td>
<td>0.051</td>
<td>0.055</td>
<td>0.052</td>
<td>0.059</td>
<td>0.060</td>
<td>0.052</td>
<td>0.056</td>
<td>0.052</td>
<td>0.065</td>
<td>0.066</td>
<td>0.053</td>
<td>0.050</td>
<td>0.047</td>
<td>0.054</td>
</tr>
<tr>
<td>$\Phi_\beta$</td>
<td>Empirical power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.817</td>
<td>0.661</td>
<td>0.418</td>
<td>0.308</td>
<td>0.301</td>
<td>0.799</td>
<td>0.478</td>
<td>0.496</td>
<td>0.360</td>
<td>0.198</td>
<td>0.848</td>
<td>0.690</td>
<td>0.434</td>
<td>0.327</td>
<td>0.315</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>0.474</td>
<td>0.283</td>
<td>0.197</td>
<td>0.170</td>
<td>0.153</td>
<td>0.510</td>
<td>0.215</td>
<td>0.215</td>
<td>0.178</td>
<td>0.136</td>
<td>0.522</td>
<td>0.304</td>
<td>0.209</td>
<td>0.171</td>
<td>0.154</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.140</td>
<td>0.089</td>
<td>0.070</td>
<td>0.056</td>
<td>0.333</td>
<td>0.100</td>
<td>0.092</td>
<td>0.072</td>
<td>0.047</td>
<td>0.338</td>
<td>0.149</td>
<td>0.091</td>
<td>0.071</td>
<td>0.059</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>0.299</td>
<td>0.126</td>
<td>0.066</td>
<td>0.056</td>
<td>0.040</td>
<td>0.355</td>
<td>0.083</td>
<td>0.076</td>
<td>0.055</td>
<td>0.040</td>
<td>0.374</td>
<td>0.154</td>
<td>0.073</td>
<td>0.058</td>
<td>0.040</td>
<td>0.171</td>
</tr>
<tr>
<td>100</td>
<td>0.993</td>
<td>0.980</td>
<td>0.874</td>
<td>0.833</td>
<td>0.861</td>
<td>0.994</td>
<td>0.892</td>
<td>0.933</td>
<td>0.906</td>
<td>0.680</td>
<td>0.995</td>
<td>0.988</td>
<td>0.889</td>
<td>0.850</td>
<td>0.872</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>0.731</td>
<td>0.419</td>
<td>0.249</td>
<td>0.195</td>
<td>0.166</td>
<td>0.782</td>
<td>0.273</td>
<td>0.273</td>
<td>0.205</td>
<td>0.150</td>
<td>0.785</td>
<td>0.465</td>
<td>0.254</td>
<td>0.204</td>
<td>0.171</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>0.561</td>
<td>0.236</td>
<td>0.114</td>
<td>0.085</td>
<td>0.067</td>
<td>0.620</td>
<td>0.129</td>
<td>0.127</td>
<td>0.091</td>
<td>0.063</td>
<td>0.633</td>
<td>0.269</td>
<td>0.113</td>
<td>0.086</td>
<td>0.069</td>
<td>0.498</td>
</tr>
<tr>
<td></td>
<td>0.570</td>
<td>0.231</td>
<td>0.088</td>
<td>0.063</td>
<td>0.050</td>
<td>0.709</td>
<td>0.122</td>
<td>0.110</td>
<td>0.073</td>
<td>0.051</td>
<td>0.730</td>
<td>0.317</td>
<td>0.092</td>
<td>0.061</td>
<td>0.050</td>
<td>0.253</td>
</tr>
</tbody>
</table>

Table B.2: Mixture random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 60$ and 100. 5000 replications.
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
\textbf{\(n\)} & \textbf{\(p\)} & \multicolumn{4}{c|}{\textbf{Model 1}} & \multicolumn{4}{c|}{\textbf{Model 2}} & \multicolumn{4}{c|}{\textbf{Model 3}} & \multicolumn{4}{c|}{\textbf{Model 4}} \\
\hline
\multirow{2}{*}{60} & \multirow{2}{*}{\textbf{\(\Phi_0\)}} & 0.048 & 0.037 & 0.045 & 0.043 & 0.034 & 0.045 & 0.041 & 0.044 & 0.043 & 0.035 & 0.035 & 0.038 & 0.039 & 0.040 & 0.037 & 0.038 & 0.035 & 0.036 & 0.039 \\
& likelihood & 1.000 & 1.000 & NA & NA & NA & 1.000 & 1.000 & NA & NA & NA & 1.000 & 1.000 & NA & NA & NA & 1.000 & 1.000 & NA & NA & NA \\
& Schott & 0.092 & 0.079 & 0.078 & 0.071 & 0.079 & 0.091 & 0.083 & 0.078 & 0.073 & 0.080 & 1.020 & 0.971 & 0.933 & 0.990 & 0.084 & 1.015 & 0.998 & 0.996 & 0.103 & 0.099 \\
& Chen & 0.069 & 0.056 & 0.057 & 0.050 & 0.051 & 0.063 & 0.057 & 0.048 & 0.043 & 0.050 & 0.057 & 0.052 & 0.047 & 0.050 & 0.048 & 0.096 & 0.093 & 0.092 & 0.101 & 0.098 \\
& Srivastava & 0.055 & 0.058 & 0.055 & 0.048 & 0.044 & 0.056 & 0.056 & 0.053 & 0.048 & 0.041 & 0.059 & 0.056 & 0.048 & 0.048 & 0.049 & 0.051 & 0.049 & 0.046 & 0.050 & 0.049 \\
\hline
\multirow{2}{*}{100} & \multirow{2}{*}{\textbf{\(\Phi_0\)}} & 0.043 & 0.038 & 0.038 & 0.034 & 0.038 & 0.041 & 0.039 & 0.039 & 0.038 & 0.030 & 0.042 & 0.036 & 0.037 & 0.031 & 0.032 & 0.035 & 0.032 & 0.030 & 0.033 & 0.033 \\
& likelihood & 1.000 & 1.000 & 1.000 & NA & NA & NA & 1.000 & 1.000 & 1.000 & NA & NA & 1.000 & 1.000 & 1.000 & NA & NA & 1.000 & 1.000 & 1.000 & NA & NA \\
& Schott & 0.089 & 0.075 & 0.079 & 0.080 & 0.078 & 0.093 & 0.081 & 0.080 & 0.085 & 0.079 & 0.095 & 0.089 & 0.089 & 0.099 & 0.088 & 0.100 & 0.094 & 0.103 & 0.094 & 0.106 \\
& Chen & 0.066 & 0.056 & 0.054 & 0.053 & 0.051 & 0.064 & 0.053 & 0.052 & 0.053 & 0.051 & 0.057 & 0.050 & 0.050 & 0.058 & 0.049 & 0.093 & 0.091 & 0.101 & 0.093 & 0.105 \\
& Srivastava & 0.052 & 0.049 & 0.053 & 0.049 & 0.049 & 0.055 & 0.052 & 0.055 & 0.051 & 0.052 & 0.062 & 0.056 & 0.054 & 0.051 & 0.048 & 0.046 & 0.048 & 0.052 & 0.047 & 0.052 \\
\hline
\end{tabular}
\caption{Empirical sizes and powers. \(\alpha = 0.05. \ n = 60 \text{ and } 100. \ 5000 \text{ replications.} \)}
\end{table}
B.2.2 Simulation results for small samples

When the sample size is small, say $n = 30$, as is explained in the discussion section in Chapter 3, the critical value derived from the asymptotic distribution is not sufficiently accurate and modification is thus needed. It is shown in this section that the proposed test with the modified critical value also performs well. As we can see from Section 3.2.3 in Chapter 3 that Chen and Li (2011)’s test performs similarly as Schott (2007)’s test. Hence to save computation cost, we considered three test statistics in this section, the proposed test, Schott (2007)’s test and Srivastava and Yanagihara (2010)’s test. We consider the same models as in Chapter 3 but with $D = I$ for the first three models and $O = I$ in the fourth model under the normal distribution. To evaluate the power of the tests, let $U = (u_{kl})$ be a matrix with $2\lfloor K/2 \rfloor$ random nonzero entries. The locations of the $\lfloor K/2 \rfloor$ nonzero entries are selected randomly from the upper triangle of $U$, each with magnitude 0.9. The other $\lfloor K/2 \rfloor$ nonzero entries in the lower triangle are determined by symmetry.

We consider two settings here. In the first case, the number of nonzero entries of the difference of two covariance matrices stays the same and we choose $K = 32$. In the other case, the number of nonzero entries increases as $p$ grows and we choose $K = \lfloor p/4 \rfloor$. We use the following four pairs of covariance matrices $(\Sigma_1^{(i)}, \Sigma_2^{(i)})$, $i = 1, 2, 3$ and 4, to compare the power of the tests, where $\Sigma_1^{(i)} = \Sigma^{(i)}$ and $\Sigma_2^{(i)} = \Sigma^{(i)} + U$.

The dimension $p$ still varies over the values 50, 100, 200, 400 and 800. Because
the sample size is small, we use the “normal cut off” method discussed in Section ?? in Chapter 3 to choose the critical value through simulation (20000 replications are used). The nominal significant level for all the tests is set at $\alpha = 0.05$. The actual sizes and powers for the four models, reported in Table B.4, are estimated from 5000 replications.

It can be seen from Table B.4 that the estimated sizes of our test $\Phi_\alpha$ and Srivastava and Yanagihara (2010)’s test are close to the nominal level 0.05 in all the cases. Schott (2007)’s test has nominal level close to 0.05 in the first three models, while having size distortion in the fourth model.

The power results in Table B.4 show that the proposed test has much higher power than the other tests in all settings. In the first case, the number of nonzero off-diagonal entries of $\Sigma_1 - \Sigma_2$ does not change when $p$ grows, so the estimated powers of all tests tend to decrease when the dimension $p$ increases. It can be seen in Table B.4 that the powers of Schott (2007) and Srivastava and Yanagihara (2010)’s tests decrease extremely fast as $p$ grows. In the second case, the number of nonzero entries increases as $p$ grows. The proposed test has much higher power than the other tests no matter how $p$ varies.

Simulations for Gamma distribution $\Gamma(10, 1)$ are also carried in this small sample case. Similar phenomena as those in the Gaussian case are observed. The results are summarized in Table B.5.
<table>
<thead>
<tr>
<th>$\Phi_\alpha$</th>
<th>$\Phi_{\alpha}$</th>
<th>$\Phi_{\alpha}$</th>
<th>$\Phi_{\alpha}$</th>
<th>$\Phi_{\alpha}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schott</td>
<td>Schott</td>
<td>Schott</td>
<td>Schott</td>
<td>Schott</td>
</tr>
<tr>
<td>Srivastava</td>
<td>Srivastava</td>
<td>Srivastava</td>
<td>Srivastava</td>
<td>Srivastava</td>
</tr>
<tr>
<td>Empirical power ($K = 32$)</td>
<td>Empirical power ($K = 32$)</td>
<td>Empirical power ($K = 32$)</td>
<td>Empirical power ($K = 32$)</td>
<td>Empirical power ($K = 32$)</td>
</tr>
<tr>
<td>Empirical power ($K = \lfloor p/4 \rfloor$)</td>
<td>Empirical power ($K = \lfloor p/4 \rfloor$)</td>
<td>Empirical power ($K = \lfloor p/4 \rfloor$)</td>
<td>Empirical power ($K = \lfloor p/4 \rfloor$)</td>
<td>Empirical power ($K = \lfloor p/4 \rfloor$)</td>
</tr>
</tbody>
</table>

Table B.4: $N(0, 1)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 30$. 5000 replications.
<table>
<thead>
<tr>
<th>$p$</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_\alpha$</td>
<td>0.042</td>
<td>0.043</td>
<td>0.045</td>
<td>0.048</td>
</tr>
<tr>
<td>Schott</td>
<td>0.073</td>
<td>0.068</td>
<td>0.061</td>
<td>0.067</td>
</tr>
<tr>
<td>Srivastava</td>
<td>0.060</td>
<td>0.049</td>
<td>0.042</td>
<td>0.039</td>
</tr>
</tbody>
</table>

|               |         |         |         |         |
| Empirical power ($K = 32$) |         |         |         |         |
| $\Phi_\alpha$ | 0.496   | 0.329   | 0.218   | 0.154   | 0.122 | 0.506 | 0.336   | 0.220   | 0.151   | 0.109 | 0.623 | 0.400 | 0.250 | 0.134 | 0.091 | 0.479 | 0.397 | 0.234 | 0.157 | 0.141 |
| Schott        | 0.355   | 0.181   | 0.114   | 0.087   | 0.080 | 0.494 | 0.247   | 0.130   | 0.102   | 0.086 | 0.745 | 0.409 | 0.198 | 0.118 | 0.100 | 0.149 | 0.113 | 0.104 | 0.094 | 0.093 |
| Srivastava    | 0.119   | 0.068   | 0.050   | 0.040   | 0.024 | 0.063 | 0.051   | 0.048   | 0.037   | 0.025 | 0.588 | 0.159 | 0.084 | 0.037 | 0.023 | 0.079 | 0.065 | 0.056 | 0.043 | 0.049 |

|               |         |         |         |         |
| Empirical power ($K = \lfloor p/4 \rfloor$) |         |         |         |         |
| $\Phi_\alpha$ | 0.260   | 0.262   | 0.313   | 0.357   | 0.417 | 0.280 | 0.287   | 0.306   | 0.339   | 0.389 | 0.342 | 0.315 | 0.328 | 0.311 | 0.303 | 0.305 | 0.264 | 0.304 | 0.432 | 0.492 |
| Schott        | 0.148   | 0.146   | 0.146   | 0.150   | 0.154 | 0.192 | 0.195   | 0.188   | 0.194   | 0.192 | 0.304 | 0.300 | 0.294 | 0.298 | 0.299 | 0.116 | 0.107 | 0.106 | 0.097 | 0.096 |
| Srivastava    | 0.065   | 0.061   | 0.053   | 0.049   | 0.032 | 0.060 | 0.052   | 0.046   | 0.039   | 0.027 | 0.138 | 0.117 | 0.138 | 0.106 | 0.076 | 0.063 | 0.056 | 0.060 | 0.053 | 0.054 |

Table B.5: $\Gamma(10,1)$ random variables. Model 1-4 Empirical sizes and powers. $\alpha = 0.05$. $n = 30$. 5000 replications.
Bibliography


