硕士学位论文

两样本高维稀疏相关性矩阵的检验

TEST ON TWO-SAMPLE HIGH-DIMENSIONAL CORRELATION MATRICES WITH SPARSE SETTINGS

杨栩智

哈尔滨工业大学

2020 年 6 月

国内图书分类号：O212 学校代码：10213
国际图书分类号：519.2 密级：公开

理学硕士学位论文

两样本高维稀疏相关性矩阵的检验

硕士研究生：杨栩智
导　　　师：田国梁教授
申请学位：理学硕士
学　　　科：概率论与数理统计
所在单位：南方科技大学
答辩日期：2020 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: O212

U.D.C: 519.2

Dissertation for the Master's Degree in Science

TEST ON TWO-SAMPLE HIGH-DIMENSIONAL CORRELATION MATRICES WITH SPARSE SETTINGS

Candidate:	Yang Xuzhi
Supervisor:	Prof. Tian Guoliang
Academic Degree Applied for:	Master of Science
Specialty:	Probability Theory and Statistics
Affiliation:	Southern University of Science and Technology
Date of Defence:	June, 2020
Degree-Conferring-Institution:	Harbin Institute of Technology

摘 要

随着计算机科学的发展，我们获取数据的能力越来越强，数据的获取场景越来越多样化，导致了数据的维度和数据量不断大幅度增加，产生了大量的高维数据问题。这样的高维数据在基因，金融，互联网领域出现得越来越多。例如：在蛋白质的分类问题中，我们往往是通过对蛋白质的基因对进行测序，从而根据不同蛋白质所蕴含的不同的基因对来区分不同种类的蛋白质。但是在实际操作过程中，由于基因测序的成本非常高，导致我们的样本量 (n) 非常少，但是每个样本所蕴含的基因对 (p) 却是成千上万，这就产生了一个“小 n 大 p”的问题。

对于这样的“小 n 大 p”问题，经典的统计方法往往会失效或者犯第一类错误（原假设为真的情况下拒绝原假设）的概率很大。产生这一现象的原因可以从随机矩阵领域中的 Marchenko-Pastur 分布的分布行为中看出：在高维数据的情形下，样本协方差矩阵所对应的特征值的波动开始和总体协方差矩阵所对应的特征值的波动发生显著性的偏差，这使得样本协方差矩阵不再是总体协方差矩阵的有效估计，自然的，高维情形下，样本相关性矩阵也不再是总体相关性矩阵的可靠估计了。这一事实导致许多经典的统计方法在高维数据的情形下表现非常糟糕，在高维统计检验的问题中，如果依然对高维数据应用低维的经典方法则往往会以较大的概率发生第一类错误。因此，在过去一段时间的研究工作中，提出新的统计方法以应对高维数据成为了现代统计学的主要挑战。

由于协方差矩阵和相关性矩阵在许多统计方法中扮演着十分重要的角色，因此在高维统计分析中，关于高维协方差矩阵和相关性矩阵的的问题又是其中至关重要的核心问题。皮尔逊协方差（下称“协方差”）往往被用来刻画不同的变量之间是否有线性关系，而由于现实数据中往往存在着不同的量纲，例如：身高数据和年龄数据，体重数据和摄入食物质量的数据，这样量纲的不同将会使得不同性质的变量之间的协方差无法比较，因此协方差往往需要被标准化来去量纲，从而可以达到比较不同性质的变量之间的相关性大小的目的。这样标准化之后的协方差我们就称之为皮尔逊相关系数（下称“相关系数”）。由不同变量之间的协方差构成的矩阵被称为协方差矩阵，由不同变量之间的相关系数构成的矩阵被称为相关性矩阵。

在统计分析中，协方差矩阵和相关性矩阵的相等性的问题往往受到格外的关注，因为有很多的统计方法是建立在协方差矩阵或相关性矩阵相等的假设之下的。

例如 Fisher 的线性判别分析就建立在两个样本的协方差矩阵相等的假设之下。因此，在高维数据分析中我们经常有必要事先检验两个样本的总体协方差矩阵或相关性矩阵是否相等，否则我们的统计方法可能会难以实施。

在检验协方差矩阵和相关性矩阵的相等性的问题上，传统的方法往往是采用由 Kullback 在 1969 年提出的似然比统计量（likelihood ratio test statistic）来处理两样本或多样本的相等性的检验问题。这种方法在传统的低维统计中有着优异且高效的表现，但是正如我们之前所提到的那样，在数据维度 p 相对于样本量 n 来说十分大的情况下，样本协方差矩阵已经不再是总体协方差矩阵的可靠的估计，因此在用似然比统计量去处理高维的协方差矩阵或相关矩阵的检验中往往会有较大的概率发生第一类错误。

基于这一事实，我们需要发展出更多的针对高维协方差矩阵或相关性矩阵的相等性检验的方法来替代传统的似然比统计量的方法。在众多的高维统计方法中，其中一种有力工具是随机矩阵理论，随机矩阵理论是基于样本协方差矩阵的特征值来构造所谓的“线性谱统计量”，通过推断线性谱统计量的极限分布来得到各种基于特征值的统计量的极限分布。这种方法的好处在于，线性谱统计量往往能包含一类基于协方差矩阵的特征值的统计量，一旦得到了线性谱统计量的极限分布，则很多基于协方差矩阵的统计量的极限分布也随之自然得到。但是缺点在于很多情形下，线性谱统计量的极限分布非常难以推算，因此对于一些情形，随机矩阵的理论复杂性较高。而在本文中我们将主要利用统计渐进理论的方法针对两样本高维相关性矩阵的问题进行研究。

在两样本高维相关性矩阵的检验的问题中，另一种有力的工具就是通过构造极值统计量来判断两个相关性矩阵是否来自同一总体。Jiang 在其 2004 年的工作中首次构造了针对单样本的高维相关性矩阵的检验问题的极值统计量，Jiang 利用 Stein 的方法证明了这种极值统计量是依分布收敛到某个 I-型极值分布的。Jiang 的工作给相关性矩阵的检验问题带来了新的思路，我们可以将极值理论引入到统计检验中来，将极值统计量的极限分布问题转化为独立随机变量或几乎独立的随机变量的和的极限分布问题，从而再利用 Stein 方法得到最终的极值统计量的渐近分布。

而 Cai, Liu, Xia 在 2013 年同样利用构造极值统计量的方法提出了针对两样本高维协方差矩阵的检验的极值统计量 M_n，并在总体相关性矩阵满足一定的稀疏性条件的情形下证明了 M_n 的极限分布也是 I-型极值分布。受到 Cai 等人的启发，Cai, Zhang 在 2016 年对于两样本高维相关性矩阵的检验提出了类似的基于最大值范数的检验统计量 T_n ，并且他们断言该统计量的极限分布将与 M_n 的极限分布完全相

同。

但是，由于 Cai, Zhang 并没有就 T_n 的极限分布给出严格的理论证明，并且考虑到协方差矩阵和相关性矩阵的是具有内在的不同的，例如：Kullback 所提出的针对协方差矩阵的似然比统计量是渐近收敛到卡方分布的，而他同时提出的针对相关性矩阵的似然比统计量是渐近收敛到卡方分布的线性组合的，二者并不具有相同的渐近行为。因此尽管 T_n 和 M_n 具有类似的构造方法，但是我们仍然有理由怀疑 T_n 的极限分布和 M_n 不完全一致，所以我们认为对这一断言给出严格的数学证明是有必要的。

在这一问题的推动下，我们的严格验证了 Cai 和 Zhang 的猜想：我们严格证明了 T_n 的极限分布的确是一种 $I-$ 型的极值分布并且形式与 M_n 的极限分布完全相同。在这一问题中我们所采用的方法类似于 Cai, Liu, Xia 在 2013 年的文章中所采用的证明技巧，而没有利用 Stein 的方法，这是因为在我们的假定中，并没有要求随机变量的独立性，而允许两个相关性矩阵具有某种稀疏性条件，因此 Jiang 中所采用的方法难以实行。

在我们的方法中我们首先证明了 T_n 的标准化部分的相合性，所以我们可以用总体的标准化部分的来代替样本的标准化部分，从而使统计量的标准化分母被其总体形式所代替。接着我们利用“截断法”证明了 T_n 可以被其“非中心化”的形式所代替，也就是证明了我们可以假定所有的总体均值和总体方差是已知，从而可以用已知的总体均值和总体方差去代替样本均值和样本方差。最后我们利用稀疏性假定和 Zaitev(1987) 中的一种推广的 Bernstein 不等式可以证明 T_n 的极限分布的确是一种 $I-$ 型的极值分布。

在 Cai, Liu, Xia 的文章中包含一个有关于随机变量四阶矩的假定，这个假设对于一切椭圆分布是成立的，但是对于更加一般的情形却不一定成立，所以我们试图去掉这一分布假设。因此我们的另一个贡献就在于：我们在另一种稀疏性假定的条件下，再次证明了 T_n 的极限分布，同时我们并不要求在 Cai, Liu, Xia 文章中的关于随机变量的四阶矩的假定成立。

这一工作的展开是基于 Xiao Han 和 Wei Biao Wu 在 2013 年发表的一项工作中所提出的有关于多元正态尾概率的估计的不等式。在该工作中，他们对于相关性矩阵满足某一特定条件的正态随机向量的尾概率给出了上界，从而我们基于这一不等式和对应的所需要的稀疏性假定，就能够在一定程度上去掉随机变量四阶矩的条件，进一步证明 T_n 的极限分布对于一般的随机变量也是 I-型极值分布。这一推广使得定理的应用范围得到了扩展。

因此，本文的主要贡献在于对两样本高维稀疏相关性矩阵的相等性检验的检

验统计量 T_n 的极限分布给出了严格的理论证明，补充了之前 Cai 和 Zhang 的研究中的遗漏问题。进一步的，我们引入了新的稀疏性条件，将证明建立在不需要分布假定的情形下，从而扩大了这一极限分布的适用范围。在此之后，我们分别对轻尾的正态分布总体的情形和重尾的伽马分布总体的情形做了统计模拟实验，模拟的结果证实了我们的结论：I-型极值分布对 T_n 的渐近行为有着比较好的拟合作用。最后，我们对全文进行了总结，并提出了展望：希望通过利用 Stein的方法能够得到这一渐近过程的收敛速度，从而通过收敛速度的研究，希望进一步提出新的极限分布来加快 T_n 的收敛速度，以求增强检验的有效性。

关键词：高维统计；相关性矩阵；极值理论

Abstract

With the development of computer science, people's ability to collect data is becoming stronger and stronger, and the forms of data are becoming more and more diversified, which lead to an incridiable increasement of the dimension and size of the data. Thus, a large number of high-dimensional statistical problems have come into being. Such high-dimensional data is increasingly available in the fields of genetics, finance and the Internet. For example, in the classification of proteins, we often sequence the gene pairs of proteins to distinguish different types of proteins. However, in practice, due to the high cost of gene sequencing, our sample size (n) is very small, while the gene pairs that contained in each sample (p) are tens of thousands, which caused a "small n large p" problem.

For such "small n large p" problems, classical methods tend to be fail or perform a high statistical size. From the Marchenko - Pastur distribution we can see why this kind of failure often happens in using of traditional statistical methods: in the case of the high-dimensional data, the flunctuation of the eigenvalues of the sample covariance matrix will significantly deviate from the flunctuation of the eigenvalues of population covariance matrix. Therefore, the sample covariance matrix is no longer a reliable estimate for the population covariance matrix and naturally, the sample correlation matrix will also no longer a reliable estimate for the population correlation matrix. This fact leads to the poor performance of many classical statistical methods in the case of high-dimensional data. So, if the traditional methods are still applied to the high-dimensional data, then it is probably for us to make a Type I error. Recently, it has become the main objective for the modern statistics to propose new statistical methods to deal with the high-dimensional data.

Because covariance matrix and correlation matrix both play important roles in many statistical methods, the problem about the covariance matrix and correlation matrix in high-dimensional statistical analysis always draw many attentions. Pearson covariance (hereinafter referred to as "covariance") is often used to describe whether there exist a linear relationship between two variables. Note that the real data often exist different scales, such as: height and age, weight and the quantity of the food intake. The data on different scales will make the covariance between two variables becomes uncomparable.

Thus, the covariance between two variables often needs to be standardized, which can unify the scales of different kinds of data. This normalized covariance is called the Pearson correlation coefficient (hereinafter referred to as the "correlation coefficient"). The matrix that composed of the covariances between different variables is called the covariance matrix, and the matrix composed of the correlation coefficients between different variables is called the correlation matrix.

In statistical analysis, the problem of the equality of two covariance matrices and correlation matrices is often paid great attention because many statistical methods are based on the assumption that the covariance matrices or correlation matrices are equal. For example, Fisher's linear discriminant analysis is based on the assumption that the covariance matrices of the two samples are equal. Therefore, in the analysis of high-dimensional data, it is often necessary to check whether the population covariance matrices or the population correlation matrices are equal, or our statistical method may be difficult to be implemented.

On the problem of testing the equality of two covariance matrices and correlation matrices, the traditional methods often apply the likelihood ratio test statistic, which was proposed by Kullback in 1969. This method performs perfectly in the low-dimensional case, but as we mentioned before, when dimension p relative to sample size n is very large, the sample covariance matrix is no longer a reliable estimate for the population one. Thus the likelihood ratio test statistic will lead to a high statistical size.

Based on this fact, we need to develop some alternative methods to replace the likelihood ratio statistic method. Among many high-dimensional statistical methods, one powerful tool is the random matrix theory. In this theory, we would like to get the asymptotic distribution of a class of test statistics through establish the limiting distribution of the "linear spectral statistic". The advantage of this method is that many test statistics can be seen as a special case of the linear spectral statistic so that once the central limit theorem of the linear spectral statistic is obtained, the limiting distribution of other specific statistic is also obtained naturally. However, the disadvantage of this method is that the inference of the aymptotic behavior of linear spectral statistic often involves complexity, sometimes it is even impossible to get the central limit theorem of the linear spectral statstic. In this paper, we will focus on using the statistical asymptotic method to study the test for two-sample high-dimensional correlation matrices.

Another powerful tool to handle on this problem is to construct the extreme value statistic. In Jiang(2004)'s work, he constructed an extreme value statistic for the one-sample test of the high-dimensional correlation matrix for the first time and he proved that this extreme value statistic wiil tend to a Type-I extreme value distribution. Jiang's work brought us some new insights on the test of correlation matrix. We can introduce the extreme theory into statistical test problem and convert the problem of an extreme value statistic into a problem of the sum of independent random variable. From here, the Stein's method can be applied to develop the central limit theorem of the extreme value statistic.

In 2013, Cai, Liu and Xia also proposed an extreme value test statistic M_n for the two-sample high-dimensional covariance matrices test, and proved that the limiting distribution of M_n is also a Type-I extreme value distribution under some sparse settings. Inspired by Cai ,Liu and Xia's work, Cai and Zhang proposed a similar test statistic T_n based on the supreme norm for the two-sample high-dimensional correlation matrices test. And they assertted that the limiting distribution of this statistic will be exactly the same as the limiting distribution of M_n.

But Cai and Zhang did not offer a strict proof for their assertion. Considering the covariance matrix and correlation matrix are intrinsically different, for example: Kullback proposed a likelihood ratio statistic based on covariance matrix and proved that this test statistic converges to chi-square distribution, meanwhile, he also constructed a likelihood ratio statistic based on correlation matrix but the latter test statistic is asymptotically distributed as a linear combination of some chi-square distribution. Thus two kinds of likelihood ratio statistic do not share the same distribution. So, although T_n and M_n are both constructed based on supreme norm, we still think it is necessary for us to give a mathematical proof for their assertion.

Under this motivation, we strictly proved Cai and Zhang's conjecture: we strictly proved that the limiting distribution of T_n is indeed an Type-I extreme value distribution with exactly the same form as it is in the case of M_n. The method of proof is developed similar to Cai, Liu, Xia's way. Due to the independent assumption is not required in our problem, the Stein's method can not be applied directly.

In our method we first prove the consistency of the normalized part of T_n, so we can replace the sample normalized part with the population one. Through this method, we simplify the denominator of T_n to be its population form, which can be seen as a constant directly. Then we use "truncation method" to prove that T_n could be approximated by its "noncentralized" form, which means that we can assume the population means and

the population variances are known. Thus we can use the known population mean and population variance to replace the sample mean and variance, and further, we assume the population mean is 0 and population variance is 1. Finally, by using the sparse assumptions and Bernstein's inequality we can obtain the limit distribution of T_n.

In Cai, Liu, Xia's work, a fourth moment condition is needed. This assumption is true for all elliptic distributions, but we don't know wheather it is true for more general case. Therefore, we try to remove this distribution assumption. Hence, another contribution of our work is that we deduced the limiting distribution of T_n without any distribution assumption but under some alternative sparse settings. This part is based on Xiao Han and Wei Biao Wu's work. They proposed an inequality to estimate the tail probability of the multivariate normal distribution. Based on this inequality, we can further prove that the asymptotic distribution of T_n is truely a Type-I extreme value distribution under some sparse assumptions but without the forth moment condition. This generalization extends the application of the theorem.

Therefore, the main contribution of our work is to give a theoratical proof for the limiting distribution of T_n. Further, we introduce a new sparse conditions, and establish the central limit theorem in the distribution free case, which help to expand the range of the application of the result. After that, we carry out some statistical simulations for the normal distribution case and gamma distribution case, respectively. The simulation results confirmed our conclusion that the Type-I extreme value distribution is a good approximation of the distribution of T_n. At last, we summarize the whole paper and give some prospects: we want to to obtain the convergence rate of the asymptotic behavior of T_n by Stein's method, so that we can further propose a new limiting distribution of T_n with a higher rate of convergence.

Keywords: high-dimensional statistics, correlation matrix, extreme value theory

目　录

Contents

Chapter 1 Introduction

1.1 Background and Significance

With the development of computer science, modern statistical data are increasingly high dimensional but relative small sample size. Genetic data is a typical example with thousands of dimensions but limited observations due to the high cost of collecting data. The same situation is also commonly found in the fields of portfolio and risk managerment. An involved statistical problem is to know whether two populations share the same distribution or same key parameters, for instance, mean vector and covariance matrix. More specifically, we denote $\mathbf{X} = (X_{ij})_{1\le i\le n_1, 1\le j\le p}$ to be the data matrix whose n_1 rows are i.i.d. p-dimensional random vectors with mean $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})$ and covariance matrix $\Sigma_1 = (\sigma_{ij1})_{p\times p}$ and $\mathbf{Y} = (Y_{ij})_{1\le i\le n_2, 1\le j\le p}$ to be another data matrix whose n_2 rows are i.i.d. p-dimesional random vectors with mean $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})$ and covariance matrix $\Sigma_2 = (\sigma_{ij2})_{p\times p}$. Let $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ be the n_1 rows of $\mathbf{X}$ and let $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ be the n_2 rows of $\mathbf{Y}$. Define the sample covariance matrices $\mathbf{S}_1 = (\hat{s}_{ij1})_{p\times p}$ and $\mathbf{S}_2 = (\hat{s}_{ij2})_{p\times p}$ as

$$\mathbf{S}_1 = (n_1 - 1)^{-1} \sum_{i=1}^{n_1} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^T,$$

$$\mathbf{S}_2 = (n_2 - 1)^{-1} \sum_{i=1}^{n_2} \left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right)^T,$$

where $\bar{\mathbf{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{X}_i$, $\bar{\mathbf{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{Y}_i$, and sample correlation matrices $\mathbf{R}_1 = (\rho_{ij1})_{p\times p}$ and $\mathbf{R}_2 = (\rho_{ij2})_{p\times p}$

$$\mathbf{R}_1 = [\operatorname{diag}(\mathbf{S}_1)]^{-1/2}\,\mathbf{S}_1\,[\operatorname{diag}(\mathbf{S}_1)]^{-1/2} \text{ and } \mathbf{R}_2 = [\operatorname{diag}(\mathbf{S}_2)]^{-1/2}\,\mathbf{S}_2\,[\operatorname{diag}(\mathbf{S}_2)]^{-1/2}$$

here $\operatorname{diag}(\mathbf{S}_\ell)$ is a diagonal matrix that consist of diagonal elements of $\mathbf{S}_\ell$. Testing the equality of two covariance matrices Σ_1 and Σ_2 is an important problem in multivariate analysis. Under the normal assumption a well-known likelihood ratio test (LRT) statistic is

$$T_1 = -2\log L_1,$$

where

$$L_1 = \frac{|\mathbf{S}_1|^{N_1/2} \cdot |\mathbf{S}_2|^{N_2/2}}{|c_1\mathbf{S}_1 + c_2\mathbf{S}_2|^{N/2}}, \tag{1-1}$$

here $N_i = n_i - 1$, $N = N_1 + N_2$ and $c_i = N_i/N$, $i = 1, 2$. The LRT statistic is commonly enjoyed in low-dimensional setting because the limiting distribution of T_1 can be dipicted by $\chi^2_{1/2p(p+1)}$ (See [1]). However, as the dimension of the data grows dramatically, the fluctuation of the eigenvalues of the sample covariance matrix be pretty different from that of population covariance matrix. Thus the size of the test could be pretty high (see [2]).

Several alternative methods have been introduced to overcome this difficulty. The first persepective is through random matrix theory. [3] firstly proposed a well-known CLT for the linear spectral statistic of large dimensional covariance matrix with the Gussian-like moment condition. Later, [4] removed the Guassian-like moment condition by alternative assumptions that are easy to verify. Besides, for two-sample covariance matrices test, [5] deduced a CLT for large dimensional F-matrix under the null hypothesis ($\Sigma_1 = \Sigma_2$) and extended this CLT to general F-matrix in [6].

The second persepective is by using extreme value theory. For example, for $\ell = 1, 2$, we write $\hat{\sigma}_{ij\ell} = \frac{n_\ell - 1}{n_\ell}\hat{s}_{ij\ell}$ and define

$$M_n =: \max_{1 \le i \le j \le p} \frac{\left(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}\right)^2}{\hat{\theta}_{ij1}/n_1 + \hat{\theta}_{ij2}/n_2}, \tag{1-2}$$

where

$$\hat{\theta}_{ij1} = \frac{1}{n_1}\sum_{k=1}^{n_1}\left[\left(X_{ki} - \bar{X}_i\right)\left(X_{kj} - \bar{X}_j\right) - \hat{\sigma}_{ij1}\right]^2, \tag{1-3}$$

and

$$\hat{\theta}_{ij2} = \frac{1}{n_2}\sum_{k=1}^{n_2}\left[\left(Y_{ki} - \bar{Y}_i\right)\left(Y_{kj} - \bar{Y}_j\right) - \hat{\sigma}_{ij1}\right]^2,$$
$$\bar{Y}_i = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_{ki},$$

then under null hypothesis and some conditions, [7] proved that M_n tends to a Type-I extreme value distribution when both populations follow elliptical distribution. In our paper, we proved that the correlation matrix version of M_n also follows the Type-I extreme value distribution.

Another interesting problem is the test

$$H_0 : \mathbf{P}_1 = \mathbf{P}_2, \tag{1-4}$$

where $\mathbf{P}_1 = (r_{ij1})_{p\times p}$ and $\mathbf{P}_2 = (r_{ij2})_{p\times p}$ denote the population correlation matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Similar to (1-1), [8] also considered the following statistic

$$L_2 = \frac{|\widehat{\mathbf{R}}_1|^{N_1/2} \cdot |\widehat{\mathbf{R}}_2|^{N_2/2}}{\left|c_1\widehat{\mathbf{R}}_1 + c_2\widehat{\mathbf{R}}_2\right|^{N/2}},$$

and claimed that $T_2 = -2\log L_2$ was still asymptotically distributed as $\chi^2_{1/2p(p+1)}$. However, [9] gave a counterexample to state that T_2 does not, in general, have an asymptotic χ^2 distribution under the null hypothesis (1-4), and the true limiting distribution, which was found by [10], is in fact a linear form in $\frac{1}{2}p(p-1)$ independent χ^2_1 variables. The different limiting distribution between T_1 and T_2 indicates that the asymptotic properties about covariance matrix and correlation matrix may vary from each other.

Inspired by this insight and [7]'s work, we consider test (1-4). In fact, a test statistic that similar to (1-2), but with the sample covariances replaced by sample correlations, has been introduced in [11] and they claimed that this newly proposed statistic shares the same asymptotic distribution as M_n. However, as we have discussed before, the inconsistency might exist between covariance matrix and correlation matrix, hence our first contribution is to provide a theoretical proof for Cai and Zhang's assertion. Moreover, many results about the test for high-dimensional covariance matrix calls for distribution assumption. For instance, the Gaussian assumption is required in [12] and [13]. In [14] 's work, they also propose condition **(C3)** (see [14]), which restrict the distribution in somewhat elliptical distribution. Thus our second contribution is to adjust the sparse settings such that the proof could be given in distribution free case.

1.2 Literature Review

Testing the equality of two correlation matrices is an important problem in multivariate analysis. Many statistical procedures including the classical Fisher's linear discriminant analysis rely on the assumption of equal correlation matrices. The test problem has been well-developed in the low-dimensional case. As we have mentioned, [8] proposed test statistic T_2 for correlation matrices based on the principle of minimum discrimination information. However, the claimed asymptotic properties in [8] was disproved by [9] through a counterexample, and [10] proved that the asymptotic distribution of T_2 is to be a linear form in $\frac{1}{2}p(p-1)$ independent χ_1^2 variables and not in general $\chi^2_{\frac{1}{2}p(p-1)}$. For further results on this topic the readers could refer to [15], [16] and [17].

With the advent of modern computer science technology, the "small n large p" problem has arouse more and more interest. But the traditional statistical tools perform poorly or are not even well defined. One way to understand this point is by looking at the flunctuation of the density of MP-Law (For more details about MP-Law, see [18]). For example, if we draw $n = 320$ i.i.d. random vectors $\{\mathbf{x}_i\}$, each with $p = 40$ i.i.d standard Gaussian components. Then from the flunctuation of the eigenvalue of the sample covariance matrix and the flunctuation of the population covariance matrix (See [19]) we can see that the sample eigenvalues of $\mathbf{S}_n$ range on a wide dispersion from unit value 1. However, the classical large-sample asymptotic property of covariance matrix indicates that the sample covariance matrix should be closed to the population covariance matrix $\mathbf{I}_p$. Thus, this contradiction implies that the sample covariance matrix is no longer a reliable estimator for population counterpart Σ. Therefoere, it is natural to understand that the sample correlation matrix may also have significant biasedness from the population one and this nature will lead to the failure of classial methods.

Therefore, in the high-dimensional setting, several test statistics for one- or two-sample correlation matrices test have been proposed. [20] firstly derived

$$\Pr\left(nW_n^2 - 4\log p + \log\log p \le x\right) - \exp\left[-(8\pi)^{-1/2}\exp(-0.5x)\right] \to 0,$$

uniformly for $x \in \mathbb{R}$ as $p/n \to y \in (0, +\infty)$ under the existence of $(30+\epsilon)$-th moment of $x_{ij}^{(\ell)}$, where $W_n = \max\limits_{1\le i<j\le p} |\hat{r}_{ij\ell}|$. However, the studies of [21] showed that the Type-I extreme value distribution still have significant biasedness from the true distribution of $nW_n^2 - 4\log p + \log\log p$. To fix this biasedness [22] involved the skewness of the population into the asymptotic distribution. Then, [23] extended [20]'s result in two directions: (i) the dimenision p of the data could grow exponentially as the sample size n for the light-tail distribution; (ii) a kind of weak dependency among variables is allowed for the first time. However, in order to compensate the weaker condition on dependency, the normal assumption is needed in the latter extension. Later on, inspired by the one-sample test problem, [14] introduced a self-normalized extreme value statistic M_n for the two-sample covariance matrices test under some sparse settings and elliptical distribution assumption. At the same time, [24] showed that a self-normalized version of W_n converges

to the Type-I extreme distribution under mild dependence conditions on the sample vector, and they offered a technique to remove distribution assumptions on the population. As for the testing of two-sample correlation matrices, [11] has proposed an extreme value statistic, say T_n, and they claimed that T_n follows the same distribution as M_n without a theoretical proof. In our work we will provide a strict proof for this conjecture, moreover, the distribution assumption in earlier ones will be removed by using some techiniques in [24].

Besides establishing test statistic through supreme norm, [25] and [26] also established statistics based on Frobenius norm to carry out test for covariance matrix and change point detection problem. Then [27] combined the supreme norm and Frobenius norm to construct new test statistics for one-, two- and three-sample correlation matrices test. Another newly proposed method by [28] is that they used random matrix theory to establish a CLT for linear spectral statstic of correlation matrix. Through this CLT, the statistics in [29] and [30] can be covered natrually.

1.3 Our Work

In this thesis, we provide an mathematical proof for the claim that the self-normalized extreme value statistic for two-sample correlation matrices test will asymptotically follows the type-I extreme value distribution. The proof will be divided into three steps. And due to we do not assume the random variables are independent, the method that used in [20] will fail in our case. Thus, we follow the idea in [14], which constructs the proof through Bonferroni's inequality. By this kind of method, the sparse settings in the correlation matrix are allowed. Moreover, motivated by [24]'s work, our work will extend the result to the case without any distribution assumptions under some alternative sparse settings.

Chapter 2 Asymptotic Distribution of T_n

2.1 Sparse Settings and Moment Conditions

We consider test statistic T_n that has been proposed in [11] in a distribution free situation where the mild dependence between variables is also allowed. Let $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ are two p-dimensional random vector with mean vectors $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})$ and covariance matrices $\Sigma_1 = (\sigma_{ij1})_{p\times p}$ and $\Sigma_2 = (\sigma_{ij2})_{p\times p}$, respectively. Furthermore, the two-sample sizes are assumed to be comparable, that is, there exist constants $C_1 > C_2 > 0$ such that $C_2 n_2 \leq n_1 \leq C_1 n_2$. Let $n = \max(n_1, n_2)$.

To test 1-4, it is unnatural to study the maximum of a collection of random variables which are on different scales, so [11] consider the normalized verson of $\max_{1\leq i\leq j\leq p} |\hat{r}_{ij1} - \hat{r}_{ij2}|$, that is,

$$T_n =: \max_{1\leq i\leq j\leq p} \frac{(\hat{r}_{ij1} - \hat{r}_{ij2})^2}{\hat{\eta}_{ij1}/n_1 + \hat{\eta}_{ij2}/n_2},$$

where

$$\hat{\eta}_{ij1} = \frac{1}{n_1}\sum_{k=1}^{n_1}\left\{\frac{\left(X_{ki} - \bar{X}_i\right)\left(X_{kj} - \bar{X}_j\right)}{\left(\hat{\sigma}_{ii1}\hat{\sigma}_{jj1}\right)^{1/2}} - \frac{\hat{r}_{ij1}}{2}\left[\frac{\left(X_{ki} - \bar{X}_i\right)^2}{\hat{\sigma}_{ii1}} + \frac{\left(X_{kj} - \bar{X}_j\right)^2}{\hat{\sigma}_{jj1}}\right]\right\}^2,$$

$$\hat{\eta}_{ij2} = \frac{1}{n_2}\sum_{k=1}^{n_2}\left\{\frac{\left(Y_{ki} - \bar{Y}_i\right)\left(Y_{kj} - \bar{Y}_j\right)}{\left(\hat{\sigma}_{ii2}\hat{\sigma}_{jj2}\right)^{1/2}} - \frac{\hat{r}_{ij2}}{2}\left[\frac{\left(Y_{ki} - \bar{Y}_i\right)^2}{\hat{\sigma}_{ii2}} + \frac{\left(Y_{kj} - \bar{Y}_j\right)^2}{\hat{\sigma}_{jj2}}\right]\right\}^2 \tag{2-1}$$

and $\hat{r}_{ij\ell} = \frac{\hat{\sigma}_{ij\ell}}{\hat{\sigma}_{ii\ell}^{1/2}\hat{\sigma}_{jj\ell}^{1/2}}$, $\ell = 1, 2$. As the normalized part, $\hat{\eta}_{ij1}$ and $\hat{\eta}_{ij2}$ can be seen as estimators for

$$\eta_{ij1} = \mathrm{Var}\left[\frac{(X_{1i} - \mu_{1i})(X_{1j} - \mu_{1j})}{(\sigma_{ii1}\sigma_{jj1})^{1/2}} - \frac{r_{ij1}}{2}\left(\frac{(X_{1i} - \mu_{1i})^2}{\sigma_{ii2}} + \frac{(X_{1j} - \mu_{1j})^2}{\sigma_{jj2}}\right)\right],$$

and

$$\eta_{ij2} = \mathrm{Var}\left[\frac{(Y_{1i} - \mu_{2i})(Y_{1j} - \mu_{2j})}{(\sigma_{ii2}\sigma_{jj2})^{1/2}} - \frac{r_{ij2}}{2}\left(\frac{(Y_{1i} - \mu_{2i})^2}{\sigma_{ii2}} + \frac{(Y_{1j} - \mu_{2j})^2}{\sigma_{jj2}}\right)\right],$$

Define $\mathcal{I}_n = \{m : 1 \leq m \leq \frac{p^2+p}{2}\}$ and $q = \mathrm{Card}(\mathcal{I}_n) = \frac{p^2+p}{2}$, j we can arrange the two-dimensional indices $\{(i,j) : 1 \leq i \leq j \leq p\}$ in any ordering and set them as $\{(i_m, j_m) : m \in \mathcal{I}_n\}$. Let

$$W_{kij}^{(\ell)} = \begin{cases} \dfrac{n_2}{n_1}(X_{ki}X_{kj} - \sigma_{ij1}), \text{ if } \ell = 1, \\ -(Y_{ki}Y_{kj} - \sigma_{ij2}), \text{ if } \ell = 2, \end{cases}$$

and

$$V_{km}^{(\ell)} = W_{ki_mj_m}^{(\ell)} - \frac{\sigma_{i_mj_m\ell}}{2}(W_{ki_mi_m}^{(\ell)} + W_{kj_mj_m}^{(\ell)}),\ \ell = 1,\ 2. \tag{2-2}$$

Define

$$\gamma_\ell(n,b) = \sup_{t\in\mathcal{I}_n} \sup_{\mathcal{A}\subset\mathcal{I}_n,|\mathcal{A}|=b} \inf_{s\in\mathcal{A}} |\mathrm{Cov}(V_{1t}^{(\ell)}, V_{1s}^{(\ell)})|,\ \ell = 1,\ 2,$$
$$\gamma_{n\ell} = \sup_{s\neq t} |\mathrm{Cov}(V_{1t}^{(\ell)}, V_{1s}^{(\ell)})|,\ \ell = 1,\ 2. \tag{2-3}$$

We consider the following sparse settings and moment conditions.

(C1) For any sequence $\{b_n\}$ such that $b_n \to \infty$, we have

- $\gamma_\ell(n, b_n) = o(1/\log b_n)$, for $\ell = 1, 2$;
- $\limsup_n \gamma_{n\ell} < 1$, for $\ell = 1, 2$.

(C1*) For any sequence $\{b_n\}$ such that $b_n \to \infty$, we have

- $\gamma_\ell(n, b_n) = o(1)$, for $\ell = 1, 2$;
- $\limsup_n \gamma_{n\ell} < 1$, for $\ell = 1, 2$;
- $\sum_{s\neq t} |\mathrm{Cov}(V_{1s}^{(\ell)}, V_{1t}^{(\ell)})|^2 = O(p^{4-\delta_\ell})$, for some $\delta_\ell > 0$, $\ell = 1, 2$.

(C2) Suppose that $\log p = o(n^{1/5})$. There exist some constant $\eta > 0$ and $K > 0$ such that

$$\mathrm{E}\exp\left[\eta(X_{1i} - \mu_{1i})^2/\sigma_{ii1}\right] \leq K,$$
$$\mathrm{E}\exp\left[\eta(Y_{1i} - \mu_{2i})^2/\sigma_{ii2}\right] \leq K,$$

for $1 \leq i \leq p$.

(C2*) Suppose that for some γ_0, $c_1 > 0$, $p \leq c_1 n^{\gamma_0}$ and for some $\epsilon > 0$

$$\mathrm{E}|(X_{1i} - \mu_{1i})/\sigma_{ii1}^{1/2}|^{4\gamma_0+4+\epsilon} \leq K,$$
$$\mathrm{E}|(Y_{1i} - \mu_{2i})/\sigma_{ii2}^{1/2}|^{4\gamma_0+4+\epsilon} \leq K,$$

for $1 \leq i \leq p$.

(C3) Suppose that for some $\tau_1 > 0$, $\tau_2 > 0$,

$$\min_{1\leq i\leq j\leq p} \frac{\eta_{ij1}}{\sigma_{ii1}\sigma_{jj1}} > \tau_1 \text{ and } \min_{1\leq i\leq j\leq p} \frac{\eta_{ij2}}{\sigma_{ii2}\sigma_{jj2}} > \tau_2.$$

Condition **(C1)** and Condition **(C1*)** imply that both the dependence between $x_{1i_t}x_{1j_t}$

and $x_{1i_s}x_{1j_s}$ and the dependence between $y_{1i_t}y_{1j_t}$ and $y_{1i_t}y_{1j_t}$ are not too strong. Condition **(C2)** and Condition **(C2*)** were considered in [14]. They indicate that for sub-guassian-type distribution, the growth speed of dimension could be an exponential function to the size of the sample while only polymonial growth rate is allowed for the distribution with polynomial tail. Condition **(C3)** will be satisfied with $\tau_1 = \tau_2 = (1-r)^2$, if the populations are normal distributions all the correlations are bounded away from ±1, that is, for some $0 < r < 1$

$$\max_{1\le i<j\le p} |r_{ij1}| \le r < 1 \text{ and } \max_{1\le i<j\le p} |r_{ij2}| \le r < 1.$$

2.2 Main Result

Now we are ready to present the asymptotic distribution of T_n under the null hypothesis. The following theorem states that $T_n - 4\log p + \log\log p$ converges to a type-I extreme value distribution with distribution function $F(x) = \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right]$.

Theorem 2.1 Suppose that **(C1)**(or **(C1*)**), **(C2)**(or **(C2)**) and **(C3)** hold. Then under H_0, for any $x \in \mathbb{R}$,

$$\Pr\left(T_n - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right]. \tag{2-4}$$

By Theorem 2.1, the asymptotic distribution of this two-sample self-normalized extreme value statistic for correlation matrices test are similar to that of the extreme value statistic for two-sample covariance matrices test in [7]. Moreover, to get rid of the elliptical assumption, we use the technique that is imposed by [24], thus an alternative sparse setting **(C1)** (**(C1*)**) is needed.

This outcome extends the result of [14] in two ways: (i) test statistic T_n, the correlation matrices version of M_n, is also asymptotically distributed as Type-I extreme value distribution; (ii) under some alternative spares assumptions, the asymptotic property in (i) will still hold without any distribution assumptions. The fundemental technique in the proof of Theorem 2.1 comes from truncation method and multivariates taylor expansion and the consistency of $\hat{\eta}_{ij1}$ and $\hat{\eta}_{ij2}$ will also be constructed. The details of the proof can be found in Chapter 4 of this dissertation.

2.3 Methodology

Before we give the specific details of the proof, the frame of our proof is decribed as follow.

2.3.1 Limiting null distribution of T_n

The idea of the first problem mainly comes from Bernstein's and Bonferroni's inequalities. The proof is divided into four steps:

Step 1. Firstly we will prove

$$\hat{\eta}_{ij\ell} \xrightarrow{p} \eta_{ij\ell}, \text{ as } n,\ p \to \infty.$$

This approximation indicates that the T_n in (2-4) can be substituted by $\hat{T}_n = \max\limits_{1 \le i \le j \le p} \frac{(\hat{r}_{ij1}-\hat{r}_{ij2})^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}$. Further, we can prove $\hat{T}_n$ can be replaced by the non-centralized version, that is

$$|\hat{T}_n - \tilde{T}_n| \xrightarrow{p} 0, \text{ as } n,\ p \to \infty,$$

where

$$\tilde{T}_n =: \max_{1 \le i \le j \le p} \frac{(\tilde{r}_{ij1} - \tilde{r}_{ij2})^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2}$$

and $\tilde{r}_{ij\ell}$ is obtained by substituting the covariances in $\hat{r}_{ij\ell}$ for non-centralized version covariances. Therefore, we can prove (2-4) under the assumption that all involved mean and variance parameters are known and the plugging in estimated mean and variance parameters does not change the limiting distribution. Therefore, (2-4) is equivalent to

$$\Pr\left(\tilde{T}_n - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

Step 2. Based the sparse assumptions in [14], we will consider a subset of $\{(i,j) : 1 \le i \le j \le p\}$, say $\mathcal{I}_0$. Then we prove the following approximation

$$\left|\max_{1 \le i \le j \le p} \frac{(\tilde{r}_{ij1} - \tilde{r}_{ij2})^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} - \max_{(i,j)\in\mathcal{I}_0} \frac{(\tilde{r}_{ij1} - \tilde{r}_{ij2})^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2}\right| \xrightarrow{p} 0, \text{ as } n,\ p \to \infty.$$

If we define $\tilde{T}_{\mathcal{I}_0} =: \max\limits_{(i,j)\in\mathcal{I}_0} \frac{(\tilde{r}_{ij1}-\tilde{r}_{ij2})^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}$, then we only need to prove

$$\Pr\left(\tilde{T}_{\mathcal{I}_0} - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

Step 3. We can combine the two samples into one sample and rewrite $\tilde{T}_{\mathcal{I}_0}$ as a sum of independent random variables. Then we apply a truncation step, which will make the Guassian approximation (see Theorem 1.1 in [31]) applicable. We write the truncated version of $\tilde{T}_{\mathcal{I}_0}$ as $\hat{T}_{\mathcal{I}_0} =: \max\limits_{(i,j)\in\mathcal{I}_0} \hat{T}^0_{ij}$. Finally, our objective is to verify

$$\Pr\left(\hat{T}_{\mathcal{I}_0} - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right]. \tag{2-5}$$

Step 4. If we assume $card(\mathcal{I}_0) = q$ and let $y_p = x + 4\log p - \log\log p$, we apply Bonferroni's inequality on $\Pr\left(\hat{T}_{\mathcal{I}_0} \ge y_p\right)$ and it follows that for any $0 < s < q/2$,

$$\begin{aligned}\sum_{d=1}^{2s}(-1)^{d-1}\sum_{1\le l_1<\cdots<l_d\le q}\Pr\left(\bigcap_{k=1}^{d}E_{l_k}\right) &\le \Pr\left(\tilde{T}_{\mathcal{I}_0} \ge y_p\right)\\ &\le \sum_{d=1}^{2s-1}(-1)^{d-1}\sum_{1\le l_1<\cdots<l_d\le q}\Pr\left(\bigcap_{k=1}^{d}E_{l_k}\right),\end{aligned} \tag{2-6}$$

where $E_{l_k} = \{\hat{T}^0_{i_{l_k}j_{l_k}} > y_p\}$. Finally, we apply Theorem 1.1 in [31] and Lemma 5 in [14] onto the part $\Pr\left(\bigcap_{k=1}^{d}E_{l_k}\right)$ and by elementary calculation we obtain

$$\Pr\left(\hat{T}_{\mathcal{I}_0} - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right],$$

which lead to

$$\Pr\left(\tilde{T}_{\mathcal{I}_0} - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right],$$

and

$$\Pr\left(\tilde{T}_n - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

2.3.2 Remove of the distribution assumption

In previous section, we need to estimate $\Pr\left(\bigcap_{k=1}^{d}E_{l_k}\right)$ in **Step 4**. To do this, we firstly define $x_p = 4\log p - \log\log p + x$ and $\theta_n = \epsilon_n(\log p)^{-1/2}$, where $\epsilon_n = o(1)$. By Theorem 1.1 in [31], we have following normal approximation:

$$\begin{aligned}\Pr\left(|\mathbf{N}_d|_{\min} \ge x_p^{1/2} + \theta_n\right) + O(p^{-M}) &\le \Pr\left(\bigcap_{k=1}^{d}E_{l_k}\right)\\ &\le \Pr\left(|\mathbf{N}_d|_{\min} \ge x_p^{1/2} - \theta_n\right) + O(p^{-M}), \text{ for any } M > 0,\end{aligned}$$

where $\mathbf{N}_d = (N_{l_1}, \ldots, N_{l_d})^T$ is a d-dimensional normal variable with zero mean vector. Thus, the problem becomes to find an estimation for the tail probabilities of a d-dimensional normal distribution. In earlier work, the following lemma is needed to carry out the estimation.

Lemma 2.1 Suppose that there exist $\kappa_1, \kappa_2 \ge \frac{1}{3}$ such that for any $i,\ j,\ k,\ l \in \{1, 2, \ldots, p\}$,

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

and (2-7)

$$\mathrm{E}(Y_i - \mu_{i2})(Y_j - \mu_{j2})(Y_k - \mu_{k2})(Y_l - \mu_{l2}) = \kappa_2 \left(\sigma_{ij2}\sigma_{kl2} + \sigma_{ik2}\sigma_{jl2} + \sigma_{il2}\sigma_{jk2}\right).$$

It can be easily seen that if **X** and **Y** have elliptically contoured distributions, then Lemma 2.1 will hold natrually. However, for more general case, wheather this condition is still satisfied is unknown. This forth moment condition will restrict the range of application of the result. Therefore, the second goal in our work is to remove this technical condition and the main idea is motivated by [24]'s work.

In fact, the Lemma 7 of [24] has offered an appropprite tail probability estimation for $\sum_{1 \le l_1 < \cdots < l_d \le q} \Pr\left(|\mathbf{N}_d|_{\min} \ge x_p^{1/2} \pm \theta_n\right)$ under some mild dependency conditions but without any distribution assumptions. Thus, to apply this lemma to our problem, the methodology will be divided into two parts:

- in [24]'s work, they require each coordinate of $\mathbf{N}_d$ has unit variance, but in our case, the variance of the coordinates are not necessary exactly unit but asymptotic unit. Thus the first step is to prove that the result in Lemma 7 of [24] will still hold for the normal variable with asymptotic unit variance;
- the sparse settings in their work are also different with our settings. Thus the second part of our work is to adjust our sparse settings such that the (B1) and (B2) conditions in [24] are satistified.

The proof of this theorem is different from the techniques that used in [20] or [22]. In their work, they assume that the component in each random vector is independent and identically distributed, while in our case, the correlation matrix is not assumed to be an identical matrix and some kind of sparse structure is allowed. Thus, the Stein's method cannot be used directly. The proof will be developed similar to the method that was established in [14]. This kind of different technique can be used to handle on the case that the random variables with weak correlations.

Chapter 3 Simulation Studies

In this section, we carry out a simulation study to assess the performance of the proposed method. We consider the two-sample correlation matrices test problem, and set the dimension p to be $50, 100, 200, 500, 1000$ and the sample size $n_1 = n_2 = 100, 150, 200$. The data were generated according to $\mathbf{x_i} = \mathbf{R}_1^{1/2}\mathbf{w}_{1i}$ and $\mathbf{y_i} = \mathbf{R}_2^{1/2}\mathbf{w}_{2i}$, where each component of $\mathbf{w}_{\ell i}$ independently follows Gaussian population $N(0,1)$ or Gamma$-(4,2)$, for $\ell = 1, 2$. For sake of space we selectively present some results in Table 3-1 ~ Table 3-4 and include the additional results in the appendix pages. Model 1 is designed for check the empirical size of the test, and Model 2 ~ Model 4 are designed for calculate the empirical size when we take $\mathbf{R}_1$ as the correlaion matrices for both $\mathbf{x}_i$ and $\mathbf{y}_i$, and they are designed to claculate the empirical power when we assume $\mathbf{R}_1$ is the population correlation matrix of $\mathbf{x}_i$ and $\mathbf{R}_2$ is the population correlation matrix of $\mathbf{y}_i$. Four different models of population correlation matrixces are summarized as follows.

- Model 1: Let $\mathbf{R}_1 = \mathbf{R}_2 = (r^{|i-j|})_{i,j=1}^p$, where r = 0.25, 0.5, 0.75, 1.0.
- Model 2: Let $\mathbf{R}_1 = (0.5^{|i-j|})_{i,j=1}^p$ and $\mathbf{R}_2 = \mathbf{R}_1 + \epsilon(\mathbf{1}_p\mathbf{1}_p^T)$, where ϵ = 0.25, 0.3, 0.35, 0.4.
- Model 3: Let $\mathbf{R}_1 = \mathbf{I}_p$ and $\mathbf{R}_2 = \mathbf{R}_1 + \mathbf{D}$, where $\mathbf{D} = (d_{ij})_{i,j=1}^p$ and $d_{ij} = \epsilon$, if $|i-j| = 1$, for ϵ = 0.05, 0.08, 0.10, 0.12.
- Model 4: Let $\mathbf{R}_1 = \mathbf{I}_p$ and $\mathbf{R}_2 = (r^{|i-j|})_{i,j=1}^p$, for r = 0.5, 0.525, 0.55, 0.575.

We set that nominal size to be 0.05, run 1000 replications for empirical size and 1000 replications for empirical power.

It can be seen from Table 3-1 that the empirical sizes of the test are close to 0.05 as the dimension p and sample size n tend to become larger and larger which reflects the fact the null limit distribution of T_n is well approximated by Type-I extreme value distribution. Moreover, from Table 3-2 ~ 3-4 we can see that the power of the test is pretty high while p, n get larger. However, when p/n is small, the power of the test will become relatively small, this fact may indicate that the convergence rate of the approximation is not quick enough.Thus, we hope to obtain the rate of convergence of the asymptotic behavior so that we can modify the limiting distribution and get a better approximation.

Table 3-1 Empirical sizes for Model 1 under Normal population

			Empirical size of Model 1				
r	n	p	50	100	200	500	1000
0.25	100		0.051	0.058	0.062	0.075	0.086
	150		0.047	0.048	0.061	0.066	0.073
	200		0.04	0.059	0.058	0.066	0.074
0.5	100		0.057	0.06	0.069	0.067	0.073
	150		0.05	0.054	0.063	0.063	0.07
	200		0.04	0.055	0.06	0.064	0.073
0.75	100		0.051	0.061	0.067	0.078	0.089
	150		0.045	0.052	0.056	0.065	0.073
	200		0.029	0.042	0.052	0.053	0.071
1.0	100		0.0	0.0	0.0	0.0	0.0
	150		0.0	0.0	0.0	0.0	0.0
	200		0.0	0.0	0.0	0.0	0.0

Table 3-2 Empirical size and empirical power for Model 2 under Normal population

			Empirical size of Model 2					Empirical power of Model 2				
ϵ	n	p	50	100	200	500	1000	50	100	200	500	1000
0.25	100		0.051	0.064	0.069	0.073	0.083	0.861	0.949	0.986	0.995	1.0
	150		0.049	0.058	0.068	0.07	0.084	0.989	0.998	1.0	1.0	1.0
	200		0.045	0.052	0.059	0.067	0.076	1.0	1.0	1.0	1.0	1.0
0.3	100		0.058	0.073	0.075	0.077	0.089	0.963	0.991	0.998	1.0	1.0
	150		0.051	0.063	0.073	0.076	0.085	1.0	1.0	1.0	1.0	1.0
	200		0.038	0.057	0.064	0.077	0.084	1.0	1.0	1.0	1.0	1.0
0.35	100		0.056	0.062	0.077	0.083	0.093	0.995	0.999	1.0	1.0	1.0
	150		0.049	0.056	0.067	0.082	0.096	1.0	1.0	1.0	1.0	1.0
	200		0.041	0.047	0.064	0.068	0.078	1.0	1.0	1.0	1.0	1.0
0.4	100		0.057	0.069	0.073	0.091	0.101	0.997	0.998	0.999	1.0	1.0
	150		0.046	0.048	0.062	0.061	0.076	0.997	1.0	1.0	1.0	1.0
	200		0.041	0.055	0.051	0.056	0.065	0.992	1.0	1.0	1.0	1.0

Table 3-3 Empirical size and empirical power for Model 3 under Normal population

			Empirical size of Model 3					Empirical power of Model 3				
r	*n*	*p*	50	100	200	500	1000	50	100	200	500	1000
	100		0.05	0.052	0.066	0.079	0.086	0.61	0.464,	0.464	0.392	0.354
0.05	150		0.048	0.07	0.074	0.07	0.089	0.98	0.964	0.95	0.881	0.831
	200		0.036	0.053	0.067	0.07	0.074	1.0	1.0	1.0	1.0	1.0
	100		0.064	0.074	0.082	0.086	0.108	0.626	0.544	0.489	0.41	0.345
0.08	150		0.046	0.057	0.071	0.08	0.086	0.98	0.965	0.947	0.899	0.818
	200		0.05	0.056	0.056	0.063	0.077	1.0	1.0	1.0	0.999	1.0
	100		0.053	0.067	0.066	0.081	0.099	0.604	0.548	0.511	0.409	0.363
0.10	150		0.044	0.05	0.063	0.079	0.082	0.986	0.957	0.952	0.884	0.833
	200		0.046	0.052	0.061	0.061	0.074	1.0	1.0	1.0	1.0	0.999
	100		0.056	0.077	0.061	0.084	0.108	0.647	0.533	0.462	0.39	0.389
0.12	150		0.053	0.067	0.067	0.074	0.093	0.981	0.973	0.942	0.887	0.818
	200		0.049	0.050	0.065	0.071	0.077	1.0	1.0	1.0	1.0	1.0

Table 3-4 Empirical size and empirical power for Model 4 under Normal population

			Empirical size of Model 4					Empirical power of Model 4				
r	*n*	*p*	50	100	200	500	1000	50	100	200	500	1000
	100		0.049	0.053	0.062	0.073	0.077	0.787	0.858	0.796	0.701	0.658
0.5	150		0.052	0.052	0.062	0.066	0.073	1.0	1.0	1.0	1.0	0.999
	200		0.042	0.043	0.058	0.056	0.063	1.0	1.0	1.0	1.0	1.0
	100		0.051	0.053	0.081	0.079	0.091	0.937	0.912	0.852	0.823	0.713
0.525	150		0.045	0.044	0.055	0.069	0.070	1.0	1.0	1.0	1.0	1.0
	200		0.044	0.05	0.062	0.055	0.061	1.0	1.0	1.0	1.0	1.0
	100		0.048	0.064	0.078	0.071	0.092	0.974	0.949	0.917	0.806	0.746
0.55	150		0.041	0.049	0.049	0.075	0.085	1.0	1.0	1.0	1.0	1.0
	200		0.041	0.052	0.053	0.065	0.08	1.0	1.0	1.0	1.0	1.0
	100		0.051	0.066	0.062	0.076	0.098	0.985	0.974	0.941	0.882	0.807
0.575	150		0.044	0.051	0.063	0.065	0.076	1.0	1.0	1.0	1.0	1.0
	200		0.031	0.045	0.05	0.071	0.073	1.0	1.0	1.0	1.0	1.0

Chapter 4 Lemmas and Proofs

The proof of this theorem is different from the techniques that used in [20] or [22]. In their work, they assume that the component in each random vector is independent and identically distributed, while in our case, the correlation matrix is not assumed to be an identical matrix and some kind of sparse structure is allowed. Thus, the Stein's method cannot be used directly. The proof will be developed similar to the method that was established in [14]. This kind of different technique can be used to handle on the case that the random variables with weak correlations.

The proof is based on an estimation of the tail probability of the multivariate normal distribution. And we first prove the consistency of the normalized part of T_n so that the denominator of T_n can be considered as a constant. This step will simplify the problem and a theorem in [31] is needed to construct the limiting distribution of T_n.

4.1 Guassian Approximation

Suppose that $\mathbf{N}_q = (N_1, \cdots, N_q)$ is a q-dimensional normal random vector with zero mean and unit variance. Let $\Sigma_q = (\sigma_{ij})_{q\times q}$ be the covariance matrix of $\mathbf{N}_q$. For any sequence $\{b_n\}$ such that $b_n \to \infty$, define

$$\gamma(n, b_n) \triangleq \sup_{t\in\mathcal{I}_n} \sup_{\mathcal{A}\subset\mathcal{I}_n, |\mathcal{A}|=b_n} \inf_{s\in\mathcal{A}} |\sigma_{ts}|,$$

$$\gamma_n \triangleq \sup_{s\neq t} |\sigma_{ts}|.$$

Consider the following conditions.

(B1) $\gamma(n, b_n) = o(1/\log b_n)$ and $\limsup_n \gamma_n < 1$.

(B2) $\gamma(n, b_n) = o(1)$, $\limsup_n \gamma_n < 1$ and $\sum_{s\neq t} |\sigma_{st}|^2 = O(q^{2-\delta})$, for some $\delta > 0$. Then the following lemma, taken from [24], provides a guassian approximation.

Lemma 4.1 Assume **(B1)** either **(B2)**. For a fixed $x \in \mathbb{R}$ and x_q satisfying $x_q = 2\log q - \log\log q - \log(4\pi) + x + o(1)$, then we have

$$\lim_{n\to\infty} \sum_{\mathcal{A}\subset\mathcal{I}_n, |\mathcal{A}|=d} \Pr\left(\bigcap_{i\in\mathcal{A}} \{|N_i| > x_q^{1/2}\}\right) = \frac{e^{-dx/2}}{d!}.$$

4.2 Consistency of $\hat{\eta}_{ij1}$ and $\hat{\eta}_{ij2}$

By following two lemmas we can prove that T_n under the assumption that all involved mean and variance parameters are known and the plugging in estimated mean and variance parameters does not change the limiting distribution.

Lemma 4.2 Under condition **(C2)** or **(C2*****)**, there exists some constanc $C > 0$ such that for any $M > 0$ and $\varepsilon > 0$,

$$\Pr\left(\max_{i,j}\left|\hat{\eta}_{ij1} - \tilde{\eta}_{ij1}\right| \geq \frac{\varepsilon_n}{\log p}\right) = O\left(p^{-M} + n_1^{-\varepsilon/8}\right), \tag{4-1}$$

$$\Pr\left(\max_{i,j}\left|\hat{\eta}_{ij2} - \tilde{\eta}_{ij2}\right| \geq \frac{\varepsilon_n}{\log p}\right) = O\left(p^{-M} + n_2^{-\varepsilon/8}\right), \tag{4-2}$$

where

$$\tilde{\eta}_{ij1} = \frac{1}{n_1}\sum_{k=1}^{n_1}\left\{\frac{X_{ki}X_{kj}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}} - \frac{\tilde{r}_{ij1}}{2}\left[\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}} + \frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right]\right\}^2,$$

$$\tilde{\eta}_{ij2} = \frac{1}{n_2}\sum_{k=1}^{n_2}\left\{\frac{Y_{ki}Y_{kj}}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}} - \frac{\tilde{r}_{ij2}}{2}\left[\frac{Y_{ki}^2}{\tilde{\sigma}_{ii2}} + \frac{Y_{kj}^2}{\tilde{\sigma}_{jj2}}\right]\right\}^2,$$

and $\varepsilon_n = \frac{(\log p)^{3/2}}{n^{1/2}}$ if **(C2)** holds, $\varepsilon_n = \frac{1}{\log p}$ if **(C2*****)** holds.

Lemma 4.3 Under condition **(C2)** or **(C2*****)**, there exists some constanc $C > 0$ such that for any $M > 0$ and $\varepsilon > 0$,

$$\Pr\left(\max_{i,j}\left|\hat{\eta}_{ij1} - \eta_{ij1}\right|/\sigma_{ii1}\sigma_{jj1} \geq \frac{\varepsilon_n}{\log p}\right) = O\left(p^{-M} + n^{-\varepsilon/8}\right), \tag{4-3}$$

$$\Pr\left(\max_{i,j}\left|\hat{\eta}_{ij2} - \eta_{ij2}\right|/\sigma_{ii2}\sigma_{jj2} \geq \frac{\varepsilon_n}{\log p}\right) = O\left(p^{-M} + n^{-\varepsilon/8}\right), \tag{4-4}$$

here ε_n is defined in Lemma 4.2.

4.3 Proof of Theorem 2.1

Without loss of generality, we assume that $\mu_1 = \mu_2 = 0$, $\sigma_{ii1} = \sigma_{ii1} = 1$ for $1 \leq i \leq p$. Let

$$\hat{T}_n = \max_{1\leq i\leq j\leq p}\frac{\left(\hat{r}_{ij1} - \hat{r}_{ij2}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2}$$

We fristly prove that $\left|T_n - \hat{T}_n\right| \rightarrow 0$ in probability. Note that under the event $\{\left|\hat{\eta}_{ij1}/\eta_{ij1} - 1\right| \leq C\varepsilon_n/\log p,\ \left|\hat{\eta}_{ij2}/\eta_{ij2} - 1\right| \leq C\varepsilon_n/\log p\}$ we have

$$\left|T_n - \hat{T}_n\right| \leq C\hat{T}_n\frac{\varepsilon_n}{\log p}. \tag{4-5}$$

By using the taylor expansion

$$\hat{r}_{ij1}=\frac{\hat{\sigma}_{ij1}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}-\frac{\tilde{\sigma}_{ij1}}{2\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}\left(\frac{\hat{\sigma}_{ii1}-\tilde{\sigma}_{ii1}}{\tilde{\sigma}_{ii1}}+\frac{\hat{\sigma}_{jj1}-\tilde{\sigma}_{jj1}}{\tilde{\sigma}_{jj1}}\right)+o_p(n^{-1/2}),$$

$$\hat{r}_{ij2}=\frac{\hat{\sigma}_{ij2}}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{\sigma}_{ij2}}{2\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}\left(\frac{\hat{\sigma}_{ii2}-\tilde{\sigma}_{ii2}}{\tilde{\sigma}_{ii2}}+\frac{\hat{\sigma}_{jj2}-\tilde{\sigma}_{jj2}}{\tilde{\sigma}_{jj2}}\right)+o_p(n^{-1/2}),$$

where residue $o_p(n^{-1/2})$ is because $\max_i \left|\bar{X}_i\right| = O_p(\sqrt{\frac{\log p}{n}})$ (see Cai, Liu, Xia 2013), thus we can estimate $\hat{T}_n$

$$\hat{T}_n \le 2\max_{i,j}\frac{\left[\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{X_{ki}X_{kj}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}-\frac{\tilde{r}_{ij1}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)}{2}\right)-\frac{1}{n_2}\sum_{k=1}^{n_2}\left(\frac{Y_{ki}Y_{kj}}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{r}_{ij2}\left(\frac{Y_{ki}^2}{\tilde{\sigma}_{ii2}}+\frac{Y_{kj}^2}{\tilde{\sigma}_{jj2}}\right)}{2}\right)\right]^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2} \tag{4-6}$$

$$+C(n_1+n_2)\max_{i,j}\tilde{r}_{ij2}^2\left(\frac{\bar{Y}_i^2}{\tilde{\sigma}_{ii2}}+\frac{\bar{Y}_j^2}{\tilde{\sigma}_{jj2}}\right)^2+C(n_1+n_2)\max_{i,j}\tilde{r}_{ij1}^2\left(\frac{\bar{X}_i^2}{\tilde{\sigma}_{ii1}}+\frac{\bar{X}_j^2}{\tilde{\sigma}_{jj1}}\right)^2$$

$$+C(n_1+n_2)\max_i\frac{\bar{X}_i^4}{\tilde{\sigma}_{ii1}^2}+C(n_1+n_2)\max_i\frac{\bar{Y}_i^4}{\tilde{\sigma}_{ii2}^2}+\max_{i,j}\frac{\left(\tilde{r}_{ij1}-\tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}+o_p(1).$$

Since $\max_i \left|\bar{X}_i\right| = O_p(\sqrt{\frac{\log p}{n}})$, there exist a constant C such that

$$\Pr\left(\max_i\left|\bar{X}_i\right|\ge C\sqrt{\frac{\log p}{n}}\right)=O\left(p^{-1}+n^{-\varepsilon/8}\right).$$

We consider

$$\begin{aligned}\Pr\left(\max_i\frac{\left|\bar{X}_i\right|}{\tilde{\sigma}_{ii1}^{1/2}}\ge C\sqrt{\frac{\log p}{n}}\right)&=\Pr\left(\max_i\left|\bar{X}_i\right|\left(-\frac{1}{2}\tilde{\sigma}_{ii1}+\frac{3}{2}+O_p((\tilde{\sigma}_{ii1}-1)^2)\right)\ge C\sqrt{\frac{\log p}{n}}\right)\\&\le\Pr\left(\max_i\left|\bar{X}_i\right|\ge C\sqrt{\frac{\log p}{n}}\right)\\&+\Pr\left(\max_i\left|\bar{X}_i\right|\max_i\left|\tilde{\sigma}_{ii1}-1\right|\ge C\sqrt{\frac{\log p}{n}}\right)\\&+\Pr\left(\max_i\left|\bar{X}_i\right|o_p\left(\frac{\log p}{n}\right)\ge C\sqrt{\frac{\log p}{n}}\right)=o_p(1).\end{aligned}$$

Therefore, we have

$$\max_i\frac{|\bar{X}_i|}{\tilde{\sigma}_{ii1}^{1/2}}=O_p\left(\sqrt{\frac{\log p}{n}}\right). \tag{4-7}$$

Hence, we have

$$\begin{aligned}(n_1+n_2)\max_{i,j}\tilde{r}_{ij1}^2\left(\frac{\bar{X}_i^2}{\tilde{\sigma}_{ii1}}+\frac{\bar{X}_j^2}{\tilde{\sigma}_{jj1}}\right)^2 &\le (n_1+n_2)\max_{i,j}\tilde{r}_{ij1}^2\max_{i,j}\left(\frac{\bar{X}_i^2}{\tilde{\sigma}_{ii1}}+\frac{\bar{X}_j^2}{\tilde{\sigma}_{jj1}}\right)^2\\ &\le o_p(1)\max_{i,j}\tilde{r}_{ij1}^2 = o_p(1).\end{aligned}$$

Set

$$Z_{kij}=\frac{n_2}{n_1}\left[\frac{X_{ki}X_{kj}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}-\frac{\tilde{r}_{ij1}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)}{2}\right],\quad 1\le k\le n_1,$$

$$Z_{kij}=-\left[\frac{Y_{ki}Y_{kj}}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{r}_{ij2}\left(\frac{Y_{ki}^2}{\tilde{\sigma}_{ii2}}+\frac{Y_{kj}^2}{\tilde{\sigma}_{jj2}}\right)}{2}\right],\quad n_1+1\le k\le n_1+n_2.$$

We can write

$$\begin{aligned}&\max_{i,j}\frac{\left[\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{X_{ki}X_{kj}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}-\frac{\tilde{r}_{ij1}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)}{2}\right)-\frac{1}{n_2}\sum_{k=1}^{n_2}\left(\frac{Y_{ki}Y_{kj}}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{r}_{ij2}\left(\frac{Y_{ki}^2}{\tilde{\sigma}_{ii2}}+\frac{Y_{kj}^2}{\tilde{\sigma}_{jj2}}\right)}{2}\right)\right]^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}\\ &=\max_{i,j}\frac{\left(\sum_{k=1}^{n_1+n_2}Z_{kij}\right)^2}{n_2^2\eta_{ij1}/n_1+n_2\eta_{ij2}}\\ &\le 2\max_{i,j}\frac{\sum_{k=1}^{n_1+n_2}Z_{kij}^2}{n_2^2\eta_{ij1}/n_1+n_2\eta_{ij2}}\end{aligned}$$

In the proof of Lemma 4.3 we have obtained for any $M>0$,

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}Z_{kij}^2-\frac{n_2^2}{n_1^2}\eta_{ij1}\right|\ge C\frac{\varepsilon_n}{\log p}\right)=O\left(p^{-M}+n^{-\varepsilon/8}\right).$$

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_2}\sum_{k=1}^{n_2}Z_{kij}^2-\eta_{ij2}\right|\ge C\frac{\varepsilon_n}{\log p}\right)=O\left(p^{-M}+n^{-\varepsilon/8}\right).$$

Therefore, we have

$$\left|\max_{i,j}\frac{\sum_{k=1}^{n_1+n_2}Z_{kij}^2}{n_2^2\eta_{ij1}/n_1+n_2\eta_{ij2}}-1\right|\le\max_{i,j}\left|\frac{\sum_{k=1}^{n_1+n_2}Z_{kij}^2-n_2^2\eta_{ij1}/n_1-n_2\eta_{ij2}}{n_2^2\eta_{ij1}/n_1+n_2\eta_{ij2}}\right|$$

$$
\begin{aligned}
&= \max_{i,j}\left|\frac{(\sum\limits_{k=1}^{n_1} Z_{kij}^2 - n_2^2\eta_{ij1}/n_1) + (\sum\limits_{k=1}^{n_2} Z_{kij}^2 - n_2\eta_{ij2})}{n_2^2\eta_{ij1}/n_1 + n_2\eta_{ij2}}\right| \\
&= \max_{i,j}\left|\frac{(\frac{1}{n_1}\sum\limits_{k=1}^{n_1} Z_{kij}^2 - n_2^2\eta_{ij1}/n_1^2)}{n_2^2\eta_{ij1}/n_1^2 + \frac{n_2}{n_1}\eta_{ij2}} + \frac{(\frac{1}{n_2}\sum\limits_{k=1}^{n_2} Z_{kij}^2 - \eta_{ij2})}{n_2\eta_{ij1}/n_1 + \eta_{ij2}}\right| \\
&\leq \max_{i,j}\left|\frac{(\frac{1}{n_1}\sum\limits_{k=1}^{n_1} Z_{kij}^2 - n_2^2\eta_{ij1}/n_1^2)}{n_2^2\eta_{ij1}/n_1^2 + \frac{n_2}{n_1}\eta_{ij2}}\right| + \max_{i,j}\left|\frac{(\frac{1}{n_2}\sum\limits_{k=1}^{n_2} Z_{kij}^2 - \eta_{ij2})}{n_2\eta_{ij1}/n_1 + \eta_{ij2}}\right| \\
&\leq C_1 \max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1} Z_{kij}^2 - n_2^2\eta_{ij1}/n_1^2\right| + C_2\left|\frac{1}{n_2}\sum_{k=1}^{n_2} Z_{kij}^2 - \eta_{ij2}\right| \\
&= o_p(1).
\end{aligned}
$$

Here we proved that (4-6) is $O_p(1)$.

In order to estimate $\max\limits_{i,j}\frac{\left(\tilde{r}_{ij1}-\tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}$, we use the following taylor expansion

$$
\tilde{r}_{ij1} = \tilde{\sigma}_{ij1} - \frac{\sigma_{ij1}}{2}\left(\tilde{\sigma}_{ii1} + \tilde{\sigma}_{jj1} - 2\right) + O_p(\left|\tilde{\sigma}_{ij1} - \sigma_{ij1}\right|^2 + \left|\tilde{\sigma}_{ii1} - \sigma_{ii1}\right|^2 + \left|\tilde{\sigma}_{jj1} - \sigma_{jj1}\right|^2),
$$

$$
\tilde{r}_{ij2} = \tilde{\sigma}_{ij2} - \frac{\sigma_{ij2}}{2}\left(\tilde{\sigma}_{ii2} + \tilde{\sigma}_{jj2} - 2\right) + O_p(\left|\tilde{\sigma}_{ij2} - \sigma_{ij2}\right|^2 + \left|\tilde{\sigma}_{ii2} - \sigma_{ii2}\right|^2 + \left|\tilde{\sigma}_{jj2} - \sigma_{jj2}\right|^2),
$$

Thus

$$
\begin{aligned}
&\max_{i,j}\frac{\left(\tilde{r}_{ij1} - \tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} \\
&\leq 2\max_{i,j}\frac{\left(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} + \frac{\sigma_{ij1}^2}{2}\max_{i,j}\frac{\left(\tilde{\sigma}_{ii2} - \tilde{\sigma}_{ii1}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} \\
&+ \frac{\sigma_{ij1}^2}{2}\max_{i,j}\frac{\left(\tilde{\sigma}_{jj2} - \tilde{\sigma}_{jj1}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} + o_p(1),
\end{aligned}
\tag{4-8}
$$

the $o_p(1)$ here is because $\max\limits_{i,j}|\tilde{\sigma}_{ijl} - \sigma_{ijl}| = O_p(\sqrt{\frac{\log p}{n}})$. Set

$$
\begin{aligned}
W_{kij} &= \frac{n_2}{n_1}\left(X_{ki}X_{kj} - \sigma_{ij1}\right), \quad 1 \leq k \leq n_1 \\
W_{kij} &= -\left(Y_{ki}Y_{kj} - \sigma_{ij2}\right), \quad n_1 + 1 \leq k \leq n_1 + n_2.
\end{aligned}
$$

Then we can write

$$
\max_{i,j}\frac{\left(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} = \max_{i,j}\frac{\left(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}\right)^2}{\theta_{ij1}/n_1 + \theta_{ij2}/n_2} \times \frac{\theta_{ij1}/n_1 + \theta_{ij2}/n_2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2}
$$

$$\leq C \max_{i,j} \frac{\left(\tilde{\sigma}_{ij1}-\tilde{\sigma}_{ij2}\right)^2}{\theta_{ij1}/n_1+\theta_{ij2}/n_2}$$

$$\leq 2C \max_{i,j} \frac{\sum_{k=1}^{n_1+n_2} W_{kij}^2}{n_2^2\theta_{ij1}/n_1+n_2\theta_{ij2}}$$

By the proof of Lemma 3 in Cai, Liu, Xia(2013), we obtain for any $M>0$,

$$\Pr\left(\max_{ij}\left|\frac{1}{n_1}\sum_{k=1}^{n_1} W_{kij}^2-\frac{n_2^2}{n_1^2}\theta_{ij1}\right|\geq C\frac{\varepsilon_n}{\log p}\right)=O\left(p^{-M}+n^{-\varepsilon/8}\right)$$

$$\Pr\left(\max_{ij}\left|\frac{1}{n_2}\sum_{k=n_1+1}^{n_2} W_{kij}^2-\theta_{ij2}\right|\geq C\frac{\varepsilon_n}{\log p}\right)=O\left(p^{-M}+n^{-\varepsilon/8}\right)$$

Same as previous case, we can prove $\left|\max_{i,j}\frac{\sum_{k=1}^{n_1+n_2} W_{kij}^2}{n_2^2\theta_{ij1}/n_1+n_2\theta_{ij2}}-1\right| = o_p(1)$, and $\max_{i,j}\frac{\sum_{k=1}^{n_1+n_2} W_{kij}^2}{n_2^2\theta_{ij1}/n_1+n_2\theta_{ij2}} = O_p(1)$. Here we proved the first term in (4.3) can be bounded in the sense of probability, and other terms in (4.3) can be proved similarly. Here we proved that $\max_{i,j}\frac{\left(\tilde{r}_{ij1}-\tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}$ can be bounded by a constant in the sense of probability. Thus combine with (4-5) and Lemma 4.3, it sufficies to show that for any $x\in\mathbb{R}$

$$\Pr\left(\hat{T}_n-4\log p+\log\log p\leq x\right)\to\exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

Let

$$\tilde{T}_n=\max_{1\leq i\leq j\leq p}\frac{\left(\tilde{r}_{ij1}-\tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2},$$

we consider

$$\left|\hat{T}_n-\tilde{T}_n\right|\leq\max_{i,j}\left|\frac{\left(\hat{r}_{ij1}-\hat{r}_{ij2}\right)^2-\left(\tilde{r}_{ij1}-\tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1+\eta_{ij2}/n_2}\right|$$

$$\leq C(n_1+n_2)\times$$

$$\max_{i,j}\left[\frac{-\bar{X}_i\bar{X}_j}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}+\frac{\bar{Y}_i\bar{Y}_j}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{r}_{ij2}}{2}\left(\frac{\bar{Y}_i^2}{\tilde{\sigma}_{ii2}}+\frac{\bar{Y}_j^2}{\tilde{\sigma}_{jj2}}\right)+\frac{\tilde{r}_{ij1}}{2}\left(\frac{\bar{X}_i^2}{\tilde{\sigma}_{ii1}}+\frac{\bar{X}_j^2}{\tilde{\sigma}_{jj1}}\right)\right]^2$$

$$+C(n_1+n_2)^{1/2}\tilde{T}_n^{1/2}\times$$

$$\left\{\max_{i,j}\left[\frac{-\bar{X}_i\bar{X}_j}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}+\frac{\bar{Y}_i\bar{Y}_j}{\left(\tilde{\sigma}_{ii2}\tilde{\sigma}_{jj2}\right)^{1/2}}-\frac{\tilde{r}_{ij2}}{2}\left(\frac{\bar{Y}_i^2}{\tilde{\sigma}_{ii2}}+\frac{\bar{Y}_j^2}{\tilde{\sigma}_{jj2}}\right)+\frac{\tilde{r}_{ij1}}{2}\left(\frac{\bar{X}_i^2}{\tilde{\sigma}_{ii1}}+\frac{\bar{X}_j^2}{\tilde{\sigma}_{jj1}}\right)\right]^2\right\}^{1/2}$$

$$+o_p(1) \tag{4-9}$$

Combine (4-7), (4-9) and (4.3) we can see that in order to prove Theorem 2.1 we only need to prove

$$\Pr\left(\tilde{T}_n - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

Note that we can rewrite $\tilde{T}_n$ as

$$\begin{aligned}
\tilde{T}_n &= \max_{i,j} \frac{\left(\tilde{r}_{ij1} - \tilde{r}_{ij2}\right)^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} \\
&= \max_{i,j} \frac{\left[\left(\tilde{\sigma}_{ij1} - \tilde{\sigma}_{ij2}\right) + \frac{\sigma_{ij1}}{2}\left(\left(\tilde{\sigma}_{ii1} - \tilde{\sigma}_{ii2}\right) + \left(\tilde{\sigma}_{jj1} - \tilde{\sigma}_{jj2}\right)\right) + o_p(n^{-1/2})\right]^2}{\eta_{ij1}/n_1 + \eta_{ij2}/n_2} \\
&= \max_{i,j} \frac{\left[\left(\sum_{k=1}^{n_1+n_2} W_{kij}\right) - \frac{\sigma_{ij1}}{2}\left(\sum_{k=1}^{n_1+n_2} W_{kii} + \sum_{k=1}^{n_1+n_2} W_{kjj}\right) + o_p(n^{1/2})\right]^2}{n_2^2\eta_{ij1}/n_1 + n_2\eta_{ij2}} \\
&= \max_{m\in\mathcal{I}_n} \frac{\left[\sum_{k=1}^{n_1+n_2}\left(W_{ki_mj_m} - \frac{\sigma_{ij1}}{2}\left(W_{ki_mi_m} + W_{kj_mj_m}\right)\right)\right]^2}{n_2^2\eta_{ij1}/n_1 + n_2\eta_{ij2}} \\
&= \max_{m\in\mathcal{I}_n} \frac{\left(\sum_{k=1}^{n_1+n_2} V_{km}\right)^2}{n_2^2\eta_{ij1}/n_1 + n_2\eta_{ij2}} + o_p(1),
\end{aligned} \tag{4-10}$$

where

$$V_{km} = W_{ki_mj_m} - \frac{\sigma_{ij1}}{2}\left(W_{ki_mi_m} + W_{kj_mj_m}\right), 1 \le k \le n_1 + n_2.$$

We write

$$Q_m = \frac{\sum_{k=1}^{n_1+n_2} V_{km}}{\sqrt{n_2^2\eta_{m1}/n_1 + n_2\eta_{m2}}},$$

$$\hat{Q}_m = \frac{\sum_{k=1}^{n_1+n_2} \hat{V}_{km}}{\sqrt{n_2^2\eta_{m1}/n_1 + n_2\eta_{m2}}},$$

where

$$\hat{V}_{km} = V_{km}I\{|V_{km}| \le \tau_n\} - \mathrm{E}V_{km}I\{|V_{km}| \le \tau_n\}$$

and $\eta_{m1} = \eta_{i_mj_m1}$, $\eta_{m2} = \eta_{i_mj_m2}$. Let $\tau_n = \eta^{-1}K_1\log(p+n)$ if **(C2)** holds and $\tau_n = n^{\frac{1}{2}}/(\log p)^{5/2}$ if **(C2*)** holds. If **(C2)** hold, then

$$\begin{aligned}
&\max_{m\in\mathcal{I}_n}\frac{1}{\sqrt{n_2^2\eta_{m1}/n_1+n_2\eta_{m2}}}\sum_{k=1}^{n_1+n_2}\mathrm{E}\,|V_{km}|\,I\{|V_{km}|\geq\tau_n\}\\
&\leq C\sqrt{n}\max_{m\in\mathcal{I}_n}\max_{1\leq k\leq n_1+n_2}\mathrm{E}\,|V_{km}|\exp(\eta|V_{km}|/K_1)/\exp(\eta\tau_n/K_1)\\
&\leq C/\sqrt{n}, \qquad\qquad (4\text{-}11)
\end{aligned}$$

and if **(C2*****)** holds, we have

$$\begin{aligned}
&\max_{m\in\mathcal{I}_n}\frac{1}{\sqrt{n_2^2\eta_{m1}/n_1+n_2\eta_{m2}}}\sum_{k=1}^{n_1+n_2}\mathrm{E}\,|V_{km}|\,I\{|V_{km}|\geq\tau_n\}\\
&\leq C\sqrt{n}\max_{m\in\mathcal{I}_n}\max_{1\leq k\leq n_1+n_2}\mathrm{E}|V_{km}|^{2+2\gamma+\epsilon/2}/\tau_n^{1+2\gamma+\epsilon}\\
&\leq C/n^{\epsilon/2}. \qquad\qquad (4\text{-}12)
\end{aligned}$$

Combine (4-11) and (4-12) we have

$$\begin{aligned}
&\Pr\left(\max_{m\in\mathcal{I}_n}\left|Q_m-\hat{Q}_m\right|\geq(\log p)^{-1}\right)\\
&\leq\Pr\left(\max_{m\in\mathcal{I}_n}\frac{\sum_{k=1}^{n_1+n_2}|V_{km}I\{|V_{km}|\geq\tau_n\}-\mathrm{E}V_{km}I\{|V_{km}|\geq\tau_n\}|}{\sqrt{n_2^2\eta_{m1}/n_1+n_2\eta_{m2}}}\geq(\log p)^{-1}\right)\\
&\leq\Pr\left(\max_{m\in\mathcal{I}_n}\max_{1\leq k\leq n_1+n_2}|V_{km}|\geq\tau_n\right)\\
&\leq\sum_{k=1}^{n_1}\Pr\left(\max_{m\in\mathcal{I}_n}|V_{km}|\geq\tau_n\right)+\sum_{k=n_1+1}^{n_1+n_2}\Pr\left(\max_{m\in\mathcal{I}_n}|V_{km}|\geq\tau_n\right)\\
&\leq n_1\Pr\left(\max_{m\in\mathcal{I}_n}\left|X_{ki_m}X_{kj_m}-\frac{\sigma_{i_mj_m1}}{2}\left(X_{ki_m}^2+X_{kj_m}^2\right)\right|\geq\tau_n/K_1\right)\\
&+n_2\Pr\left(\max_{m\in\mathcal{I}_n}\left|Y_{ki_m}Y_{kj_m}-\frac{\sigma_{i_mj_m1}}{2}\left(Y_{ki_m}^2+Y_{kj_m}^2\right)\right|\geq\tau_n/K_1\right)\\
&\leq n\left[\Pr\left(\max_{m\in\mathcal{I}_n}X_{ki_m}^2\geq\tau_n/2\right)+\Pr\left(\max_{m\in\mathcal{I}_n}Y_{ki_m}^2\geq\tau_n/2\right)\right]\\
&\leq np\max_{m\in\mathcal{I}_n}\left[\Pr\left(X_i^2\geq\tau_n/2\right)+\Pr\left(Y_i^2\geq\tau_n/2\right)\right]=o(1)
\end{aligned}$$

Note that

$$\left|\max_{m\in\mathcal{I}_n}Q_m^2-\max_{m\in\mathcal{I}_n}\hat{Q}_m^2\right|\leq3\max_{m\in\mathcal{I}_n}\left|\hat{Q}_m-Q_m\right|^2+2\max_{m\in\mathcal{I}_n}|Q_m|\max_{m\in\mathcal{I}_n}\left|Q_m-\hat{Q}_m\right|,$$

hence we only need to prove that for any $x\in\mathbb{R}$ we have

$$\Pr\left(\max_{m\in\mathcal{I}_n}\hat{Q}_m^2-4\log p+\log\log p\leq x\right)\to\exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right]. \qquad (4\text{-}13)$$

According to Bonferroni inequality and let $q = \mathrm{Card}(\mathcal{I}_n)$, then for any integer $0 < s < q/2$, we have

$$\sum_{d=1}^{2s}(-1)^{d-1}\sum_{1\le m_1<\cdots<m_d\le q}\Pr\left(\bigcap_{j=1}^{d}E_{m_j}\right) \le \Pr\left(\max_{m\in\mathcal{I}_n}\hat{Q}_m^2 \ge y_p\right)$$
$$\le \sum_{d=1}^{2s-1}(-1)^{d-1}\sum_{1\le m_1<\cdots<m_d\le q}\Pr\left(\bigcap_{j=1}^{d}E_{m_j}\right), \tag{4-14}$$

where $E_{m_j} = \{\hat{Q}_{m_j}^2 \ge y_p\}$. Let

$$\tilde{V}_{km} = \hat{V}_{km}/(n_2\eta_{m1}/n_1 + \eta_{m2})^{1/2}, \text{for } m \in \mathcal{I}_n$$

and $\mathbf{M}_k = \left(\tilde{V}_{km_1}, \tilde{V}_{km_2}, \ldots, \tilde{V}_{km_d}\right)$, $1 \le k \le n_1 + n_2$. Thus, we can see

$$\Pr\left(\bigcap_{j=1}^{d}E_{m_j}\right) = \Pr\left(\left|n_2^{-1/2}\sum_{k=1}^{n_1+n_2}\mathbf{M}_k\right|_{\min} \ge y_p^{1/2}\right). \tag{4-15}$$

By Theorem 1 in Zaitsev, A. Yu (1987) we have

$$\Pr\left(\left|n_2^{-1/2}\sum_{k=1}^{n_1+n_2}\mathbf{M}_k\right|_{\min} \ge y_p^{1/2}\right) \le \Pr\left(|\mathbf{N}_d|_{min} \ge y_p^{1/2} - \epsilon_n(\log p)^{-1/2}\right)$$
$$+ c_1 d^{5/2}\exp\left(-\frac{n^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\right),$$

here c_1 and c_2 are two positive constant, $\mathbf{N}_d = (N_1, \cdots, N_d)$ is a d-dimensional normal vector with zero mean and $\mathrm{Cov}(\mathbf{N}_d) = \frac{n_1}{n_2}\mathrm{Cov}(\mathbf{M}_1) + \mathrm{Cov}(\mathbf{M}_{n_1+1})$. And

$$\epsilon_n = \begin{cases} (\log p)^{3/2}n^{-\frac{3}{10}}, \text{if } \mathbf{(C2)} \text{ holds}, \\ \dfrac{1}{(\log p)^{1/2}}, \text{ if } (\mathbf{C2}^*) \text{ holds}, \end{cases}$$

Thus, ϵ_n tends to 0 and we can see

$$c_1 d^{5/2}\exp\left(-\frac{n^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\right) = O(p^{-M}),$$

for any $M > 0$ and fixed s. Similarly, we have

$$\Pr\left(\left|n_2^{-1/2}\sum_{k=1}^{n_1+n_2}\mathbf{M}_k\right|_{\min} \ge y_p^{1/2}\right) \ge \Pr\left(|\mathbf{N}_d|_{min} \ge y_p^{1/2} + \epsilon_n(\log p)^{-1/2}\right)$$
$$- c_1 d^{5/2}\exp\left(-\frac{n^{1/2}\epsilon_n}{c_2 d^3\tau_n(\log p)^{1/2}}\right).$$

Thus, it follows that

$$\sum_{1\le m_1<\cdots<m_d\le q} \Pr\left(|\mathbf{N}_d|_{min} \ge y_p^{1/2} + \epsilon_n(\log p)^{-1/2}\right) + O(p^{-M}) \le \sum_{1\le m_1<\cdots<m_d\le q} \Pr\left(\bigcap_{j=1}^{d} E_{m_j}\right)$$

$$\le \sum_{1\le m_1<\cdots<m_d\le q} \Pr\left(|\mathbf{N}_d|_{min} \ge y_p^{1/2} - \epsilon_n(\log p)^{-1/2}\right) + O(p^{-M}). \tag{4-16}$$

For $0 < t, s < d$, let

$$C_0 = \frac{1}{\sqrt{\frac{n_2}{n_1}\eta_{t1} + \eta_{t2}}}\frac{1}{\sqrt{\frac{n_2}{n_1}\eta_{s1} + \eta_{s2}}}.$$

We define $F_{m_j} = \{N_{m_j} \ge y_p^{1/2} - \epsilon_n(\log p)^{-1/2}\}$ By elementary calculation, we can see that when **(C2)** holds

$$\left|\mathrm{Cov}(N_t, N_s) - \left(\frac{n_1}{n_2}C_0\mathrm{Cov}(V_{1t}, V_{1s}) + C_0\mathrm{Cov}(V_{n_1+1,t}, V_{n_1+1,s})\right)\right| \le C\tau_n^{-1}. \tag{4-17}$$

If **(C2*)** holds, we have

$$\left|\mathrm{Cov}(N_t, N_s) - \left(\frac{n_1}{n_2}C_0\mathrm{Cov}(V_{1t}, V_{1s}) + C_0\mathrm{Cov}(V_{n_1+1,t}, V_{n_1+1,s})\right)\right| \le C\tau_n^{-(2\gamma_0+\epsilon/2)}. \tag{4-18}$$

Because of condition **(C1)** (**(C1*)**) and (4-17), (4-18) we can see that for each fixed d, $\mathrm{Cov}(\mathbf{N}_d)$ satisfies **(B1)** either **(B2)**. Thus, by Lemma 4.1 we have

$$\lim_{n\to\infty} \sum_{1\le m_1<\cdots<m_d\le q} \Pr\left(\bigcap_{j=1}^{d} F_{m_j}\right) = \frac{e^{-dx/2+d\log\frac{1}{\sqrt{8\pi}}}}{d!}. \tag{4-19}$$

Take (4-19) into (4-16) we obtain

$$\sum_{1\le m_1<\cdots<m_d\le q} \Pr\left(\bigcap_{j=1}^{d} E_{m_j}\right) = \frac{e^{-dx/2+d\log\frac{1}{\sqrt{8\pi}}}}{d!}. \tag{4-20}$$

Then combine (4-20) with (4-14) we got

$$\Pr\left(\max_{m\in\mathcal{I}_n} \hat{Q}_m^2 - 4\log p + \log\log p \le x\right) \to \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right].$$

4.4 Proof of Lemma 4.2

We only prove (4-1) and (4-2) can be proved similarly. We fisrt prove the lemma under assumption **(C2)**. Without loss of generaliy we assume that $\mathrm{EX} = 0$ and $\mathrm{Var}\,(X_i) = 1$ for $1 \le i \le p$.

Write

$$\hat{\eta}_{ij1} = \frac{\frac{1}{n_1}\sum_{k=1}^{n_1}\left(X_{ki} - \bar{X}_i\right)^2\left(X_{kj} - \bar{X}_j\right)^2}{\hat{\sigma}_{ii1}\hat{\sigma}_{jj1}}(1 + \frac{\hat{r}_{ij1}^2}{2}) + \frac{\left(\hat{r}_{ij1}^2\right)\frac{1}{n_1}\sum_{k=1}^{n_1}\left(X_{ki} - \bar{X}_i\right)^4}{4\hat{\sigma}_{ii1}^2}$$

$$
\begin{aligned}
&+\frac{\left(\hat{r}_{ij1}^{2}\right)\frac{1}{n_1}\sum\limits_{k=1}^{n_1}\left(X_{kj}-\bar{X}_j\right)^4}{4\hat{\sigma}_{jj1}^{2}}\\
&-\left(\frac{\hat{\sigma}_{ij1}}{\hat{\sigma}_{ii1}\hat{\sigma}_{jj1}}\right)\left[\frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1}\left(X_{ki}-\bar{X}_i\right)^3\left(X_{kj}-\bar{X}_j\right)}{\sigma_{ii1}}+\frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1}\left(X_{ki}-\bar{X}_i\right)\left(X_{kj}-\bar{X}_j\right)^3}{\sigma_{jj1}}\right]\\
&\triangleq A\left(1+\frac{B^2}{2}\right)+\frac{B^2C_i}{4}+\frac{B^2C_j}{4}-D\left(E_i+E_j\right). \qquad (4\text{-}21)
\end{aligned}
$$

Based on the first order Taylor expansion of 3-variable and 2-variable function $\frac{x}{yz}$, $\frac{x^2}{(yz)}$ and $\frac{x}{y}$ for $x \in \mathbb{R}$, and $y, z > 0$

$$
\frac{x^2}{(yz)}=\frac{x_0^2}{y_0z_0}+\frac{2x_0(x-x_0)}{y_0z_0}-\frac{x_0^2}{y_0z_0}\left(\frac{y-y_0}{y_0}+\frac{z-z_0}{z_0}\right)+o(x-x_0)+o(y-y_0)+o(z-z_0),
$$

$$
\frac{x}{(yz)}=\frac{x_0}{(y_0z_0)}+\frac{x-x_0}{(y_0z_0)}-\frac{x_0}{(y_0z_0)}\left(\frac{y-y_0}{y_0}+\frac{z-z_0}{z_0}\right)+o(x-x_0)+o(y-y_0)+o(z-z_0),
$$

$$
\frac{x}{y}=\frac{x_0}{y_0}+\frac{x-x_0}{y_0}-\frac{x_0}{y_0^2}\left(y-y_0\right),
$$

A could be approximated by

$$
\begin{aligned}
A=&\frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1}X_{ki}^2X_{kj}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}\\
&+\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{-2X_{ki}^2X_{kj}\bar{X}_j}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}+\frac{-2X_{ki}X_{kj}^2\bar{X}_i}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}+\frac{X_{ki}^2\bar{X}_j^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}\right)\\
&+\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{X_{kj}^2\bar{X}_i^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}+\frac{4X_{ki}X_{kj}\bar{X}_i\bar{X}_j}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}-\frac{3\bar{X}_i^2\bar{X}_j^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}\right)\\
&+\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{\bar{X}_i^2X_{ki}^2X_{kj}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}+\frac{\bar{X}_j^2X_{kj}^2X_{ki}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}\right)+o_p(n^{-1/2})\\
&\triangleq\tilde{A}+A^*,
\end{aligned}
$$

where

$$
\tilde{A}=\frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1}X_{ki}^2X_{kj}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}},
$$

and A^* equals to the rest of A. B could be approximated by

$$\begin{aligned} B =& \frac{\tilde{\sigma}_{ij1}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} + \frac{1}{n_1}\sum_{k=1}^{n_1}\frac{-2X_{ki}X_{kj}\bar{X}_i\bar{X}_j}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} + \frac{1}{n_1^2}\sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_1}\frac{X_{k_1i}X_{k_1j}X_{k_2i}X_{k_2j}\bar{X}_i^2}{\tilde{\sigma}_{ii1}^2\tilde{\sigma}_{jj1}} \\ &+ \frac{1}{n_1^2}\sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_1}\frac{X_{k_1i}X_{k_1j}X_{k_2i}X_{k_2j}\bar{X}_j^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}^2} + o_p(n^{-1}) \\ &\triangleq \tilde{B} + B^*, \end{aligned}$$

where

$$\tilde{B} = \frac{\tilde{\sigma}_{ij1}^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}$$

and B^* equals to the rest of B. C_i and C_j could be approximated by

$$\begin{aligned} C_i =& \frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1} X_{ki}^4}{\tilde{\sigma}_{ii1}^2} - \frac{1}{n_1}\sum_{k=1}^{n_1}\left(-\frac{2}{n_1}\sum_{k_1=1}^{n_1}\frac{X_{k_1i}^2\bar{X}_i{}^2X_{ki}^4}{\tilde{\sigma}_{ii1}^4} + \frac{X_{ki}^4\bar{X}_i^4}{\tilde{\sigma}_{ii1}^4}\right) \\ &+ \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{-4X_{ki}^3\bar{X}_i}{\tilde{\sigma}_{ii1}^2} + \frac{6X_{ki}^2\bar{X}_i^2}{\tilde{\sigma}_{ii1}^2} - \frac{4X_{ki}\bar{X}_i^3}{\tilde{\sigma}_{ii1}^2} + \frac{\bar{X}_i^4}{\tilde{\sigma}_{ii1}^2}\right) + o_p(n^{-1/2}) \\ &\triangleq \tilde{C}_i + C_i^*, \end{aligned} \tag{4-22}$$

and

$$\begin{aligned} C_j =& \frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1} X_{kj}^4}{\tilde{\sigma}_{jj1}^2} - \frac{1}{n_1}\sum_{k=1}^{n_1}\left(-\frac{2}{n_1}\sum_{k_1=1}^{n_1}\frac{X_{k_1j}^2\bar{X}_j{}^2X_{kj}^4}{\tilde{\sigma}_{jj1}^4} + \frac{X_{kj}^4\bar{X}_j^4}{\tilde{\sigma}_{jj1}^4}\right) \\ &+ \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{-4X_{kj}^3\bar{X}_j}{\tilde{\sigma}_{jj1}^2} + \frac{6X_{kj}^2\bar{X}_j^2}{\tilde{\sigma}_{jj1}^2} - \frac{4X_{kj}\bar{X}_j^3}{\tilde{\sigma}_{jj1}^2} + \frac{\bar{X}_j^4}{\tilde{\sigma}_{jj1}^2}\right) + o_p(n^{-1/2}) \\ &\triangleq \tilde{C}_j + C_j^*, \end{aligned}$$

where

$$\tilde{C}_i = \frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1} X_{ki}^4}{\tilde{\sigma}_{ii1}^2},$$

$$\tilde{C}_j = \frac{\frac{1}{n_1}\sum\limits_{k=1}^{n_1} X_{kj}^4}{\tilde{\sigma}_{jj1}^2},$$

and C_i^*, C_j^* are the rest of C_i and C_j respectively. D could be approximated by

$$D = \frac{\tilde{\sigma}_{ij1}}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} - \frac{\bar{X}_i\bar{X}_j}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} + \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{X_{ki}X_{kj}\bar{X}_i^2}{\tilde{\sigma}_{ii1}^2\tilde{\sigma}_{jj1}} + \frac{X_{kj}X_{kj}\bar{X}_j^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}^2}\right) + o_p(n^{-1})$$

$$\triangleq \tilde{D} + D^*$$

where

$$\tilde{D} = \frac{\tilde{\sigma}_{ij1}}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}},$$

and D^* is the rest of D. E_i and E_j could be approximated by

$$\begin{aligned}
E_i = {} & \frac{\frac{1}{n_1}\sum_{k=1}^{n_1} X_{ki}^3 X_{kj}}{\tilde{\sigma}_{ii1}} + \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{X_{ki}^3 X_{kj}\bar{X}_i^2}{\tilde{\sigma}_{ii1}^2}\right) \\
& + \frac{1}{n_1}\sum_{k=1}^{n_1}\left(-\frac{X_{ki}^3\bar{X}_j}{\tilde{\sigma}_{ii1}} - \frac{3X_{ki}^2 X_{kj}\bar{X}_i}{\tilde{\sigma}_{ii1}} + \frac{3X_{ki}^2\bar{X}_i\bar{X}_j}{\tilde{\sigma}_{ii1}}\right) \\
& + \frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{3X_{ki}X_{kj}\bar{X}_i}{\tilde{\sigma}_{ii1}} - \frac{3X_{ki}\bar{X}_i^2\bar{X}_j}{\tilde{\sigma}_{ii1}} + \frac{\bar{X}_i^3 X_{kj}}{\tilde{\sigma}_{ii1}} - \frac{\bar{X}_i^3\bar{X}_j}{\tilde{\sigma}_{ii1}}\right) \\
& + o_p(1) \\
& \triangleq \tilde{E}_i + E_i^*,
\end{aligned}$$

where

$$\tilde{E}_i = \frac{\frac{1}{n_1}\sum_{k=1}^{n_1} X_{ki}^3 X_{kj}}{\tilde{\sigma}_{ii1}},$$

$$\tilde{E}_j = \frac{\frac{1}{n_1}\sum_{k=1}^{n_1} X_{kj}^3 X_{ki}}{\tilde{\sigma}_{jj1}},$$

and E_i^*, E_j^* equals to the rest of E_i and E_j respectively. Thus, we can write

$$\begin{aligned}
\hat{\eta}_{ij1} = {} & \tilde{\eta}_{ij1} + \frac{1}{2}\tilde{A}B^* + A^*\left(1 + \frac{1}{2}\tilde{B}\right) + \frac{1}{2}B^*A^* + \frac{C_iB^* + C_i^*\tilde{B}}{4} + \frac{C_jB^* + C_j^*\tilde{B}}{4} \\
& - \tilde{D}\left(E_i^* + E_j^*\right) - D^*\left(\tilde{E}_i + \tilde{E}_j\right) - D^*\left(E_i^* + E_j^*\right). \qquad (4\text{-}23)
\end{aligned}$$

By taking the taylor expansion

$$\frac{1}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} = 1 - (\tilde{\sigma}_{ii1} - 1) - (\tilde{\sigma}_{jj1} - 1) + o_p(n^{-1/2})$$

and combine with the fact that $\max\limits_{i,j}|\tilde{\sigma}_{ij1} - \sigma_{ij1}| = O_p(\sqrt{\log p/n})$ we can see that in order to prove

$$\Pr\left(\max_{i,j}\frac{1}{n}\left|\sum_{k=1}^{n}\frac{X_{ki}^2 X_{kj}\bar{X}_j}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}\right| \geq C\sqrt{\frac{\log p}{n}}\right) = O\left(p^{-M}\right),$$

we only need to prove the result

$$\Pr\left(\max_{i,j}\frac{1}{n}\left|\sum_{k=1}^{n}X_{ki}^{2}X_{kj}\bar{X}_{j}\right|\geq C_{5}\sqrt{\frac{\log p}{n}}\right)=O\left(p^{-M}\right),$$

which has been proved in the supplementary material of [32]. Other terms in (4-23) can be proved similarily. Here we complete the proof.

4.5 Proof of Lemma 4.3

Write

$$\begin{aligned}\tilde{\eta}_{ij1}-\eta_{ij1}=&\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\frac{\left(X_{ki}X_{kj}\right)^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}-\mathrm{E}\left(X_iX_j\right)^2\right)\\&+\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\sigma_{ij1}\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-\tilde{r}_{ij1}\frac{X_{ki}X_{kj}}{\left(\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}\right)^{1/2}}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)\right]\\&+\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\tilde{r}_{ij1}^2}{4}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right].\end{aligned}$$

We first assume that **(C2)** holds. It suffices to show that

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\left(X_{ki}X_{kj}\right)^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}-\mathrm{E}\left(X_iX_j\right)^2\right]\right|\geq C\sqrt{\log p/n_1}\right)=O(p^{-M}),\tag{4-24}$$

$$\begin{aligned}&\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\sigma_{ij1}\mathrm{E}\left(X_i^3X_j+X_iX_j^3\right)-\tilde{r}_{ij1}\left(\frac{X_{ki}^3X_{kj}}{\tilde{\sigma}_{ii1}^{3/2}\tilde{\sigma}_{jj1}^{1/2}}+\frac{X_{ki}X_{kj}^3}{\tilde{\sigma}_{ii1}^{1/2}\tilde{\sigma}_{jj1}^{3/2}}\right)\right]\right|\geq C\sqrt{\log p/n_1}\right)\\&=O(p^{-M}),\end{aligned}\tag{4-25}$$

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\tilde{r}_{ij1}^2}{4}\left(\frac{X_{ki}^2}{\tilde{\sigma}_{ii1}}+\frac{X_{kj}^2}{\tilde{\sigma}_{jj1}}\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\right|\geq C\sqrt{\log p/n_1}\right)=O(p^{-M}).\tag{4-26}$$

Combine the following taylor expansions

$$\frac{\tilde{\sigma}_{ij1}}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}=\sigma_{ij1}+(\tilde{\sigma}_{ij1}-\sigma_{ij1})-\sigma_{ij1}(\tilde{\sigma}_{ii1}-1)-\sigma_{ij1}(\tilde{\sigma}_{jj1}-1)+o_p(n^{-1/2}),$$

$$\frac{1}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}}=1-(\tilde{\sigma}_{ii1}-1)-(\tilde{\sigma}_{jj1}-1)+o_p(n^{-1/2}),$$

$$\frac{1}{\tilde{\sigma}_{ii1}}=1-(\tilde{\sigma}_{ii1}-1)+o_p(n^{-1/2}),$$

with the fact that $\max_{i,j}|\tilde{\sigma}_{ij1}-\sigma_{ij1}| = O_p(\sqrt{\frac{\log p}{n}})$, we only need to prove the following results

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left(\left(X_{ki}X_{kj}\right)^2-\mathrm{E}\left(X_iX_j\right)^2\right)\right|\geq C\sqrt{\log p/n_1}\right)=O(p^{-M}), \qquad (4\text{-}27)$$

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\sigma_{ij1}\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-\sigma_{ij1}X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\geq C\sqrt{\log p/n_1}\right)$$
$$=O(p^{-M}), \qquad (4\text{-}28)$$

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\right|\geq C\sqrt{\log p/n_1}\right)=O(p^{-M}).$$
$$(4\text{-}29)$$

Define

$$\hat{X}_{kj}=X_{kj}I\left\{\left|X_{kj}\right|\leq\tau\sqrt{\log(p+n_1)}\right\},\ \check{X}_{kj}=X_{kj}-\hat{X}_{kj}$$

where τ is sufficient large. Firstly, we consider (4-24). We have

$$\begin{aligned}\left|\mathrm{E}\left(X_{ki}X_{kj}\right)^2-\mathrm{E}\left(X_{ki}\hat{X}_{kj}\right)^2\right| &\leq C\left(\mathrm{E}X_{kj}^4I\left\{\left|X_{kj}\right|\geq\tau\sqrt{\log(p+n)}\right\}\right)^{1/2}\\ &\leq C(n+p)^{-\tau^2\eta/2}\left(\mathrm{E}X_{kj}^4\exp\left(2^{-1}\eta X_{kj}^2\right)\right)^{1/2}\\ &\leq C(n+p)^{-\tau^2\eta/2}.\end{aligned}$$

Thus it follows that

$$\begin{aligned}&\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\left(X_{ki}X_{kj}\right)^2-\mathrm{E}\left(X_iX_j\right)^2\right]\right|\geq C\sqrt{\log p/n_1}\right)\\ &\leq\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\left(X_{ki}\hat{X}_{kj}\right)^2-\mathrm{E}\left(X_{ki}\hat{X}_{kj}\right)^2\right]\right|\geq\frac{1}{2}C\sqrt{\log p/n}\right)\\ &+n_1p\max_i\Pr\left(|X_i|\geq\tau\sqrt{\log(p+n_1)}\right). \qquad (4\text{-}30)\end{aligned}$$

Note that

$$n_1p\Pr\left(|X_{11}|\geq\tau\sqrt{\log(p+n)}\right)\leq n_1p(n_1+p)^{-\tau^2\eta}\mathrm{E}\exp\left(\eta X_{11}^2\right)=O\left(p^{-M}\right). \qquad (4\text{-}31)$$

Let $t=\eta\left(8\tau^4\right)^{-1}\sqrt{\log p/n_1}$ and $\hat{Z}_{kij}=\left(X_{ki}\hat{X}_{kj}\right)^2-\mathrm{E}\left(X_i\hat{X}_j\right)^2$. Then we have the following estimation

$$
\begin{aligned}
&\sum_{k=1}^{n_1} \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t\left|\hat{Z}_{kij}\right|\right) \\
&\leq \log p \left\{\mathrm{E}\left[\eta^2(8\tau^4)^{-2} 2\left(\left|X_{ki}\hat{X}_{kj}\right|^4 + \mathrm{E}\left|X_{ki}\hat{X}_{kj}\right|^4\right)\right]^2\right\}^{1/2} \times \\
&\left\{\mathrm{E}\exp\left[2\eta(8\tau^4)^{-1}\sqrt{\frac{\log p}{n_1}}\left(\left(X_{ki}\hat{X}_{kj}\right)^2 + \mathrm{E}\left(\hat{X}_{ki}\hat{X}_{kj}\right)^2\right)\right]\right\}^{1/2} \\
&\leq \log p \left\{\mathrm{E}\left[\eta^4(8\tau^4)^{-4} 8\left(\left|X_{ki}X_{kj}\right|^8 + \mathrm{E}\left|X_{ki}X_{kj}\right|^8\right)\right]\right\}^{1/2} \times \\
&\left\{\mathrm{E}\exp\left[2\eta(8\tau^4)^{-1}\sqrt{\frac{\log p}{n_1}}\left(\left(X_{ki}\hat{X}_{kj}\right)^2 + \mathrm{E}\left(X_{ki}\hat{X}_{kj}\right)^2\right)\right]\right\}^{1/2} \\
&\leq \log p \left\{\mathrm{E}\left[\eta^4(8\tau^4)^{-4} 8\left(\left|X_{ki}X_{kj}\right|^8 + \mathrm{E}\left|X_{ki}X_{kj}\right|^8\right)\right]\right\}^{1/2} \times \\
&\left\{\mathrm{E}\exp\left[2\eta(8\tau^4)^{-1}\sqrt{\frac{\log p}{n_1}}\tau^2 \log(p+n_1)\left((X_{ki})^2 + \mathrm{E}(X_{ki})^2\right)\right]\right\}^{1/2} \qquad \text{(4-32)}
\end{aligned}
$$

Thus, by (4-32) and ($C2$) we can see that

$$
\sum_{k=1}^{n_1} \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t\left|\hat{Z}_{kij}\right|\right) \leq C_{\tau,\eta} \log p, \qquad \text{(4-33)}
$$

where $C_{\tau,\eta}$ is positive a constant only relate to τ and η. Then we have

$$
\begin{aligned}
&\Pr\left(\max_{i,j} \frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\left(X_{ki}\hat{X}_{kj}\right)^2}{\tilde{\sigma}_{ii1}\tilde{\sigma}_{jj1}} - \mathrm{E}\left(X_{ki}\hat{X}_{kj}\right)^2\right] \geq \frac{1}{2}C\sqrt{\log p/n}\right) \\
&\leq \exp(-Ct\sqrt{n_1 \log p}) \prod_{k=1}^{n_1} \mathrm{E}\exp\left(t\hat{Z}_{kij}\right) \\
&\leq \exp(-Ct\sqrt{n_1 \log p}) \prod_{k=1}^{n_1} \left(1 + \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t\left|\hat{Z}_{kij}\right|\right)\right) \\
&\leq \exp\left(-Ct\sqrt{n_1 \log p} + \sum_{k=1}^{n_1} \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t\left|\hat{Z}_{kij}\right|\right)\right) \\
&\leq \exp\left(-C\eta\left(8\tau^4\right)^{-1} \log p + c_{\tau,\eta} \log p\right) \\
&\leq Cp^{-M}.
\end{aligned}
$$

Similarily, we can show that

$$
\Pr\left(\max_{i,j} \frac{1}{n_1}\sum_{k=1}^{n_1}\left[\left(X_{ki}\hat{X}_{kj}\right) - \mathrm{E}\left(X_{ki}\hat{X}_{kj}\right)^2\right] \leq -\frac{1}{2}C\sqrt{\log p/n}\right) \leq Cp^{-M}.
$$

Thus, (4-24) is proved. As for (4-25), we define the following notation

$$\begin{aligned} I\{|X_i| \le \tau\sqrt{\log(p+n_1)}\} &\triangleq \hat{I}_{X_i}, \\ I\{|X_i| > \tau\sqrt{\log(p+n_1)}\} &\triangleq \check{I}_{X_i}, \end{aligned}$$

and we use the same technique as above. We consider

$$\begin{aligned} &\max_{i,j}\left|\sigma_{ij1}\mathrm{E}X_iX_j\left(X_i^2+X_j^2\right)-\sigma_{ij1}\mathrm{E}X_iX_j\left(X_i^2+X_j^2\right)\hat{I}_{X_i}\hat{I}_{X_j}\right| \\ &= \max_{i,j}\left|\sigma_{ij1}\mathrm{E}X_iX_j\left(X_i^2+X_j^2\right)(1-\hat{I}_{X_i}\hat{I}_{X_j})\right| \\ &\le \max_{i,j}\left|\sigma_{ij1}\mathrm{E}X_iX_j\left(X_i^2+X_j^2\right)\left(\check{I}_{X_i}+\check{I}_{X_j}\right)\right| \\ &\le 2\max_{i,j}\left|\sigma_{ij1}\left(\mathrm{E}X_i^2X_j^2(X_i^2+X_j^2)^2\right)^{1/2}\left(\mathrm{E}\left(\check{I}_{X_i}+\check{I}_{X_j}\right)\right)^{1/2}\right| \qquad (4\text{-}34) \\ &\le C(p+n_1)^{-\eta\tau^2/2}, \end{aligned}$$

Thus, it follows that

$$\begin{aligned} &\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\sigma_{ij1}\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-\sigma_{ij1}X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\ge C\sqrt{\log p/n_1}\right) \\ &\le \Pr\left(\max_{i,j}\left|\sum_{k=1}^{n_1}\left[\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\sigma_{ij1}\hat{I}_{X_i}\hat{I}_{X_j}\ge\frac{\sqrt{n_1\log p}}{2}\right) \\ &+ n_1\Pr\left(\max_{i,j}\left|\sigma_{ij1}X_iX_j\left(X_i^2+X_j^2\right)(1-\hat{I}_{X_i}\hat{I}_{X_j})\right|\ge\frac{1}{2}C\sqrt{\log p/n_1}-C(n_1+p)^{-\eta\tau^2/2}\right) \\ &\le \Pr\left(\max_{i,j}\left|\sum_{k=1}^{n_1}\left[\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\sigma_{ij1}\hat{I}_{X_i}\hat{I}_{X_j}\ge\frac{\sqrt{n_1\log p}}{2}\right) \\ &+ n_1\Pr\left(\max_{j}|X_{1j}|\ge\tau\sqrt{\log(p+n_1)}\right) \\ &\le \Pr\left(\max_{i,j}\left|\sum_{k=1}^{n_1}\left[\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\sigma_{ij1}\hat{I}_{X_i}\hat{I}_{X_j}\ge\frac{\sqrt{n_1\log p}}{2}\right) \end{aligned}$$

Let

$$t=\eta\left(8\tau^4\right)^{-1}\sqrt{\log p/n_1}$$

and

$$\hat{Z}_{kij}=\sigma_{ij1}\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)\hat{I}_{X_i}\hat{I}_{X_j}-\sigma_{ij1}X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\hat{I}_{X_i}\hat{I}_{X_j}.$$

Then we consider

$$\mathrm{E}t^2\hat{Z}_{kij}^2\exp\left(t\left|\hat{Z}_{kij}\right|\right)$$

$$
\begin{aligned}
&\leq \log p \left\{ \mathrm{E}\left[\left(\eta^2 \left(8\tau^4\right)^{-2}\right) \hat{Z}_{kij}^2 \right]^2 \right\}^{1/2} \\
&\times \left\{ \mathrm{E} \exp \left[2\eta(8\tau^4)^{-1} \sqrt{\frac{\log p}{n_1}} \left(2\tau^4 \log^2(p+n_1)\right) \right] \right\}^{1/2} \\
&\leq \log p \left\{ 2\mathrm{E}\left[\left(\eta^2 \left(8\tau^4\right)^{-2}\right) \left(\sigma_{ij1}^2 \mathrm{E} X_i^2 X_j^2 \left(X_i^2 + X_j^2\right)^2 + \sigma_{ij1}^2 X_{ki}^2 X_{kj}^2 \left(X_{ki}^2 + X_{kj}^2\right)^2 \right) \right]^2 \right\}^{1/2} \\
&\times \left\{ \mathrm{E} \exp \left[2\eta(8\tau^4)^{-1} \sqrt{\frac{\log p}{n_1}} \left(2\tau^4 \log^2(p+n_1)\right) \right] \right\}^{1/2} \qquad (4\text{-}35)
\end{aligned}
$$

Thus, by (4-35) and condition $(C2)$ we can deduced that

$$
\sum_{k=1}^{n_1} \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t \left|\hat{Z}_{kij}\right|\right) \leq C_{\eta,\tau} \log p. \qquad (4\text{-}36)
$$

Then we have

$$
\begin{aligned}
&\Pr\left(\max_{i,j} \frac{1}{n_1} \sum_{k=1}^{n_1} \hat{Z}_{kij} \geq \frac{1}{2} C \sqrt{\log p / n_1} \right) \\
&\leq \exp(-Ct\sqrt{n_1 \log p}) \prod_{k=1}^{n_1} \mathrm{E} \exp\left(t \hat{Z}_{kij}\right) \\
&\leq \exp(-Ct\sqrt{n_1 \log p}) \prod_{k=1}^{n_1} \left[1 + \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t \left|\hat{Z}_{kij}\right|\right) \right] \\
&\leq \exp\left[-Ct\sqrt{n_1 \log p} + \sum_{k=1}^{n_1} \mathrm{E} t^2 \hat{Z}_{kij}^2 \exp\left(t \left|\hat{Z}_{kij}\right|\right) \right] \\
&\leq \exp\left(-C\eta \left(8\tau^4\right)^{-1} \log p + c_{\tau,\eta} \log p \right) \\
&\leq C p^{-M}.
\end{aligned}
$$

Similarily, we can prove that

$$
\begin{aligned}
&\Pr\left(\max_{i,j} \frac{1}{n_1} \sum_{k=1}^{n_1} \hat{Z}_{kij} \leq -\frac{1}{2} C \sqrt{\log p / n_1} \right) \\
&\leq C p^{-M}.
\end{aligned}
$$

As for (4-26) We consider

$$
\begin{aligned}
&\max_{i,j} \left| \frac{\sigma_{ij1}^2}{4} \mathrm{E} \left(X_i^2 + X_j^2\right)^2 \hat{I}_{X_i} \hat{I}_{X_j} - \frac{\sigma_{ij1}^2}{4} \mathrm{E} \left(X_i^2 + X_j^2\right)^2 \right| \\
&\leq \max_{i,j} \frac{\sigma_{ij1}^2}{2} \mathrm{E} \left(X_i^2 + X_j^2\right)^2 \check{I}_{X_i}
\end{aligned}
$$

$$\begin{aligned}
&\le \max_{i,j} \frac{\sigma_{ij1}^2}{2} \left[\mathrm{E}\left(X_i^2 + X_j^2\right)^4\right]^{1/2} \left(\mathrm{E}\check{I}_{X_i}\right)^{1/2} \\
&\le C\left(p+n_1\right)^{-\eta\tau^2/2}.
\end{aligned}$$

Then, we can see

$$\begin{aligned}
&\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\right|\ge C\sqrt{\log p/n_1}\right) \\
&\le \Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\hat{I}_{X_i}\hat{I}_{X_j}\right|\ge \frac{1}{2}C\sqrt{\log p/n_1}\right) \\
&+n_1\Pr\left(\max_{i,j}\left(X_{ki}^2+X_{kj}^2\right)^2\frac{\sigma_{ij1}^2}{4}\left(1-\hat{I}_{X_i}\hat{I}_{X_j}\right)\ge\frac{1}{2}C\sqrt{\log p/n_1}-C\left(p+n_1\right)^{-\eta\tau^2/2}\right) \\
&+n_1\Pr\left(\max_j|X_{1j}|\ge\tau\sqrt{\log(p+n_1)}\right) \\
&=\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\hat{I}_{X_i}\hat{I}_{X_j}\right|\ge \frac{1}{2}C\sqrt{\log p/n_1}\right) \\
&+O(p^{-M}).
\end{aligned}$$

Let

$$t=\eta\left(8\tau^4\right)^{-1}\sqrt{\log p/n_1}$$

and

$$\hat{Z}_{kij}=\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\hat{I}_{X_i}\hat{I}_{X_j}.$$

Then we consider

$$\begin{aligned}
&\sum_{k=1}^{n_1}\mathrm{E}t^2\hat{Z}_{kij}^2\exp\left(t\left|\hat{Z}_{kij}\right|\right) \\
&\le \log p\,\mathrm{E}\frac{2\eta^2}{(8\tau^4)^2}\left[\frac{\sigma_{ij1}^4}{16}\left(X_{ki}^2+X_{kj}^2\right)^4+\frac{\sigma_{ij1}^4}{16}\mathrm{E}\left(X_i^2+X_j^2\right)^4\right] \\
&\times\exp\left[\frac{2\eta}{8\tau^4}\sqrt{\frac{\log p}{n_1}}\sigma_{ij1}^2\tau^4\log^2(p+n_1)\right] \\
&\le \log p\left\{\mathrm{E}\frac{8\eta^4}{(8\tau^4)^4}\frac{\sigma_{ij1}^8}{16^2}\left[\left(X_{ki}^2+X_{kj}^2\right)^8+\mathrm{E}\left(X_i^2+X_j^2\right)^8\right]\right\}^{1/2} \\
&\times\left[\mathrm{E}\frac{\eta}{4}\sqrt{\frac{\log p}{n_1}}\sigma_{ij1}^2\log^2(p+n_1)\right]^{1/2}
\end{aligned}$$

$$\begin{aligned}&\le \log p\left\{\mathrm{E}\frac{8\eta^4}{(8\tau^4)^4}\frac{\sigma_{ij1}^8}{16^2}\left[\left(X_{ki}^2+X_{kj}^2\right)^8+\mathrm{E}\left(X_i^2+X_j^2\right)^8\right]\right\}^{1/2}\\&\times\left[\mathrm{E}\frac{\eta}{4}\sqrt{\frac{\log p}{n_1}}\sigma_{ij1}^2\log^2(p+n_1)\right]^{1/2}\end{aligned}\tag{4-37}$$

Therefore, based (4-37) and **(C2)**, there exists a constant $C_{\tau,\eta}$ that only depend on τ and η such that

$$\sum_{k=1}^{n_1}\mathrm{E}t^2\hat{Z}_{kij}^2\exp\left(t\left|\hat{Z}_{kij}\right|\right)\le C_{\eta,\tau}\log p.$$

Thus, we have

$$\begin{aligned}&\Pr\left(\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\hat{I}_{X_i}\hat{I}_{X_j}\ge\frac{1}{2}C\sqrt{\log p/n_1}\right)\\&\le\exp(-Ct\sqrt{n_1\log p})\prod_{k=1}^{n_1}\mathrm{E}\exp\left(t\hat{Z}_{kij}\right)\\&\le\exp(-Ct\sqrt{n_1\log p})\prod_{k=1}^{n_1}\left[1+\mathrm{E}t^2\hat{Z}_{kij}^2\exp\left(t\left|\hat{Z}_{kij}\right|\right)\right]\\&\le\exp(-Ct\sqrt{n_1\log p})\prod_{k=1}^{n_1}\left[1+\mathrm{E}t^2\hat{Z}_{kij}^2\exp\left(t\left|\hat{Z}_{kij}\right|\right)\right]\\&\le\exp\left[-C\eta\left(8\tau^4\right)^{-1}\log p+c_{\tau,\eta}\log p\right]\\&\le Cp^{-M}.\end{aligned}$$

Similarly, we can use the same method to prove

$$\Pr\left(\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\hat{I}_{X_i}\hat{I}_{X_j}\le-\frac{1}{2}C\sqrt{\log p/n_1}\right)\le Cp^{-M}.$$

Here we complete the proor of (4-26).

It remains to prove (4-27) - (4-29) by replacing $O(p^{-M})$ with $O(p^{-M}+n^{-\varepsilon/8})$ under **(C2*****)**. Define

$$Y_{ij,k}=\left(X_{ki}X_{kj}\right)^2,\quad \hat{Y}_{ij,k}=Y_{ij,k}I\left\{\left|Y_{ij,k}\right|\le n/(\log p)^8\right\}.$$

Then we have

$$\begin{aligned}&\left|\mathrm{E}Y_{ij,k}-\mathrm{E}\hat{Y}_{ij,k}\right|=\mathrm{E}X_{ki}^2X_{kj}^2I\{X_{ki}^2X_{kj}^2>n/(\log p)^8\}\\&\le n^{-\gamma_0-\varepsilon/4}(\log p)^{8\gamma_0+2\varepsilon}\mathrm{E}\left(X_{ki}^2X_{kj}^2\right)^{\gamma_0+1+\varepsilon/4}\\&\le Cn^{-\gamma_0}.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}
&\Pr\left(\max_{i,j}\left|\sum_{k=1}^{n_1}\left(Y_{ij,k}-\mathrm{E}Y_{ij,k}\right)\right|\geq\frac{n\varepsilon_n}{\log p}\right)\\
&\leq\Pr\left(\max_{i,j}\left|\sum_{k=1}^{n_1}\left(\hat{Y}_{ij,k}-\mathrm{E}\hat{Y}_{ij,k}\right)\right|\geq 2^{-1}\frac{n\varepsilon_n}{\log p}\right)+\Pr\left(\max_{i,j,k}\left|Y_{ij,k}\right|\geq\frac{n}{(\log p)^8}\right)\\
&\leq Cp^2\exp\left[-C(\log p)^4\right]+Cn^{-\varepsilon/8}
\end{aligned}$$

where the last inequality follows from Bernstein's inequality and **(C2*)**. Here we proved that

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\left(X_{ki}X_{kj}\right)^2-\mathrm{E}\left(X_iX_j\right)^2\right]\right|\geq\frac{\varepsilon_n}{\log p}\right)=O(p^{-M}+n^{\varepsilon/8}). \tag{4-38}$$

We can similarily prove

$$\begin{aligned}
&\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\sigma_{ij1}\mathrm{E}\left(X_iX_j\right)\left(X_i^2+X_j^2\right)-\sigma_{ij1}X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)\right]\right|\geq\frac{\varepsilon_n}{\log p}\right)\\
&=O(p^{-M}+n^{-\varepsilon/8}),
\end{aligned} \tag{4-39}$$

and

$$\Pr\left(\max_{i,j}\left|\frac{1}{n_1}\sum_{k=1}^{n_1}\left[\frac{\sigma_{ij1}^2}{4}\left(X_{ki}^2+X_{kj}^2\right)^2-\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2\right]\right|\geq\frac{\varepsilon_n}{\log p}\right)=O(p^{-M}+n^{-\varepsilon/8}), \tag{4-40}$$

by letting

$$Y_{ij,k}=\sigma_{ij1}X_{ki}X_{kj}\left(X_{ki}^2+X_{kj}^2\right)$$

and

$$Y_{ij,k}=\frac{\sigma_{ij1}^2}{4}\mathrm{E}\left(X_i^2+X_j^2\right)^2.$$

Here we complete the proof of this Lemma.

结 论

1. 主要结果

本文主要讨论了两样本的高维相关性矩阵的检验问题。我们介绍了在低维情况下，传统的似然比检验统计量的不足及其失效的原因，之后我们列举了一些现有的处理高维协方差矩阵的方法，主要分为从随机矩阵的角度处理和统计渐进理论的角度处理，而本文主要针对的是第二种角度。Cai, Liu, Xia 对两样本的高维协方差矩阵提出了对应的极值统计量，并且证明了它的极限分布是第一型极值分布，而 Cai, Zhang 也类似地提出了针对两样本高维相关性矩阵地极值统计量, 但没有给出其渐近行为的理论证明。在本文中我们严格证明了相关性矩阵的极值统计量的极限分布同样是第一型的极值分布，同时我们的证明并不需要任何的分布假定，因此，我们将此结论推广到了无分布假定的情形。

2. 创新点

由于协方差矩阵和相关性矩阵存在着内在性质的不同，所以一些基于协方差矩阵的统计量的极限分布可能和类似的基于相关性矩阵的统计量的极限分布是不同的，正如 Kullback 提出的两种分别针对协方差矩阵检验和相关性矩阵的似然比统计量的极限分布是不同的。因此验证相关性矩阵的极值统计量的极限分布是否和协方差矩阵的极值统计量一致是必要的，而我们的工作就严格的验证了这一点，证实了两种分别针对相关性矩阵和协方差矩阵的极值统计量的极限分布确实是一致的。进一步，Cai, Zhang 的断言是基于一种类椭圆分布的假设之下的，而我们的证明再一定程度上去掉了这一假设，故将结论推广到了无分布假设的情形下。

3. 展望

本文严格地验证了了相关性矩阵的极值统计量的极限分布确实是第一型的极值分布，接下来自然的问题就是关于收敛速度的问题。事实上，利用 Stein's Method，我们可以考虑关于相关性矩阵的极值统计量的收敛速度。通过考虑收敛速度，我们可以修正极限分布，从而得到收敛速度更快的极限过程，使得统计检验更有效。

Conclusions

1. Main Results

In this article, we discuss the test of the two-sample high-dimensional correlation matrices. We talk about the drawbacks of the classic methods and explain why the traditional likelihood ratio test statistic will fail in the case of high-dimesional data. Then, we list some alternative methods to handle on high-dimensional covariance matrices, which are mainly divided into the perspective of random matrix theory and the perspective of establishing some new statistics. Cai, Liu, Xia (2013) proposed an extreme value statistic M_n for the two-sample high-dimensional covariance matrices test, and proved that the limit distribution of M_n is the Type-I extreme value distribution, while Cai, Zhang (2016) similarly proposed an extreme value statistic T_n for two-sample high-dimensional correlation matrices test. However, they did not give a theoretical proof for the asymptotic behavior of T_n. In this paper, we strictly prove that the limit distribution of T_n is also a Type-I extreme value distribution, meanwhile, our proof does not require any distribution assumption.

2. Innovations

Due to the intrinsic differences that exist between the covariance matrix and correlation matrix, the asymptotic behavior of the statistic that based on covariance matrix may differ from the asymptotic behavior of statistic that based on correlation matrix. Therefore, it is necessary to verify whether the limit distribution of the extreme value distribution of the correlation matrix case is consistent with the extreme value distribution of the covariance matrix case. Our work has strictly verified that the limit distributions for the two cases are indeed consistent. Further, Cai and Zhang's assertion is based on the population with a kind of ellptical distribution property, but our proof does not require this assumption, so this conclusion is extended to the case of no distribution assumption.

3. Prospect

This paper proves that the limit distribution of the extreme value distribution in correlation matrix case is the Type-I extreme value distribution, and the following natural problem is about the convergence rate. In fact, by Stein's Method, we can consider the Berry-Esseen bound of the asymptotic behavior, and through the bound, we are able to adjust our limiting distribution which could help us to obtain a better approximation.

References

[1] Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis : Vol 5[M]. [S.l.] : Prentice hall Upper Saddle River, NJ, 2002.

[2] Bai Z, Jiang D, Yao J, et al. Corrections to LRT on Large-dimensional Covariance Matrix by RMT[J]. The Annals of Statistics, 2009, 37(6B) : 3822-3840.

[3] Bai Z, Silverstein J W. CLT for Linear Spectral Statistics of Large-dimensional Sample Covariance Matrices[J]. The Annals of Probability, 2004, 32(1A) : 553-605.

[4] Zheng S, Bai Z, Yao J. Substitution Principle for CLT of Linear Spectral Statistics of High-dimensional Sample Covariance Matrices with Applications to Hypothesis Testing[J]. The Annals of Statistics, 2015, 43(2) : 546-591.

[5] Zheng S. Central Limit Theorems for Linear Spectral Statistics of Large Dimensional F-matrices[J]. Annales de l'IHP Probabilités et statistiques, 2012, 48(2) : 444-476.

[6] Zheng S, Bai Z, Yao J. CLT for Eigenvalue Statistics of Large-dimensional General Fisher Matrices with Applications[J]. Bernoulli, 2017, 23(2) : 1130-1178.

[7] Cai T T, Liu W, Xia Y. Two-sample Test of High Dimensional Means under Dependence[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(2) : 349-372.

[8] Kullback S. On Testing Correlation Matrices[J]. Applied Statistics, 1967 : 80-85.

[9] Jennrich R I. An Asymptotic χ2 Test for the Equality of Two Correlation Matrices[J]. Journal of the American Statistical Association, 1970, 65(330) : 904-912.

[10] Aitkin M A. Some Tests for Correlation Matrices[J]. Biometrika, 1969 : 443-446.

[11] Cai T T, Zhang A. Inference for High-dimensional Differential Correlation Matrices[J]. Journal of multivariate analysis, 2016, 143 : 107-126.

[12] Srivastava M S, Yanagihara H. Testing the Equality of Several Covariance Matrices with Fewer Observations Than the Dimension[J]. Journal of Multivariate Analysis, 2010, 101(6) : 1319-1329.

[13] Schott J R. A Test for the Equality of Covariance Matrices When the Dimension is Large Relative to the Sample Sizes[J]. Computational Statistics & Data Analysis, 2007, 51(12) : 6535-6542.

[14] Cai T, Liu W, Xia Y. Two-sample Covariance Matrix Testing and Support Recovery in High-dimensional and Sparse Settings[J]. Journal of the American Statistical Association, 2013, 108(501) : 265-277.

[15] Aitkin M, Nelson W, Reinfurt K H. Tests for Correlation Matrices[J]. Biometrika, 1968, 55(2) : 327-334.

[16] Browne M. The Likelihood Ratio Test for the Equality of Correlation Matrices[J]. British Journal of Mathematical and Statistical Psychology, 1978, 31(2) : 209-217.

[17] Modarres R, Jernigan R W. Testing the Equality of Correlation Matrices[J]. Communications in Statistics-Theory and Methods, 1992, 21(8) : 2107-2125.

[18] Bai Z, Silverstein J W. Spectral Analysis of Large Dimensional Random Matrices : Vol 20[M]. [S.l.] : Springer, 2010.

[19] Yao J, Zheng S, Bai Z. Sample Covariance Matrices and High-dimensional Data Analysis[M]. [S.l.] : Cambridge University Press Cambridge, 2015.

[20] Jiang T. The Asymptotic Distributions of the Largest Entries of Sample Correlation Matrices[J]. The Annals of Applied Probability, 2004, 14(2) : 865-880.

[21] Liu W-D, Lin Z, Shao Q-M. The Asymptotic Distribution and Berry–Esseen Bound of a New Test for Independence in High Dimension with An Application to Stochastic Optimization[J]. The Annals of Applied Probability, 2008, 18(6) : 2337-2366.

[22] Shao Q-M, Zhou W-X. Necessary and Sufficient Conditions for the Asymptotic Distributions of Coherence of Ultra-high Dimensional Random Matrices[J]. The Annals of Probability, 2014, 42(2) : 623-648.

[23] Cai T T, Jiang T. Limiting Laws of Coherence of Random Matrices with Applications to Testing Covariance Structure and Construction of Compressed Sensing Matrices[J]. The Annals of Statistics, 2011, 39(3) : 1496-1525.

[24] Xiao H, Wu W B. Asymptotic Theory for Maximum Deviations of Sample Covariance Matrix Estimates[J]. Stochastic Processes and their Applications, 2013, 123(7) : 2899-2920.

[25] Li J, Chen S X. Two Sample Tests for High-dimensional Covariance Matrices[J]. The Annals of Statistics, 2012, 40(2) : 908-940.

[26] Zhong P-S, Li R, Santo S. Homogeneity Tests of Covariance Matrices with High-dimensional Longitudinal Data[J]. Biometrika, 2019.

[27] Zheng S, Cheng G, Guo J, et al. Test for High-dimensional Correlation Matrices[J]. The Annals of Statistics, 2019, 47(5) : 2887-2921.

[28] Gao J, Han X, Pan G, et al. High Dimensional Correlation Matrices: the Central Limit Theorem and Its Applications[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(3) : 677-693.

[29] Schott J R. Testing for Complete Independence in High Dimensions[J]. Biometrika, 2005, 92(4) : 951-956.

[30] Morrison D F, Marshall L C, Sahlin H L. Multivariate Statistical Methods[J], 1976.

[31] Zaitsev A Y. On the Gaussian Approximation of Convolutions under Multidimensional Analogues of SN Bernstein's Inequality Conditions[J]. Probability theory and related fields, 1987, 74(4) : 535-566.

[32] Cai T, Liu W. Adaptive Thresholding for Sparse Covariance Matrix Estimation[J]. Journal of the American Statistical Association, 2011, 106(494) : 672-684.

Appendix A Simulation for Gamma Distribution Case

Here is the simulation results for the case of Gamma distribution. We consider the two-sample correlation matrices test problem, and set the dimension p to be $p = 50, 100, 200, 500, 1000$ and the sample size $n_1 = n_2 = 100, 150, 200$. The data were generated according to $\mathbf{x_i} = \mathbf{R}_1^{1/2}\mathbf{w}_{1i}$ and $\mathbf{y_i} = \mathbf{R}_2^{1/2}\mathbf{w}_{2i}$, where each component of $\mathbf{w}_{\ell i}$ independently follows Gamma$-(4, 2)$, for $\ell = 1, 2$. Again we summerize the different models below.

- Model 1: Let $\mathbf{R}_1 = \mathbf{R}_2 = (r^{|i-j|})_{i,j=1}^{p}$, where r = 0.25, 0.5, 0.75, 1.0.
- Model 2: Let $\mathbf{R}_1 = (0.5^{|i-j|})_{i,j=1}^{p}$ and $\mathbf{R}_2 = \mathbf{R}_1 + \epsilon(\mathbf{1}_p\mathbf{1}_p^T)$, where ϵ = 0.25, 0.3, 0.35, 0.4.
- Model 3: Let $\mathbf{R}_1 = \mathbf{I}_p$ and $\mathbf{R}_2 = \mathbf{R}_1 + \mathbf{D}$, where $\mathbf{D} = (d_{ij})_{i,j=1}^{p}$ and $d_{ij} = \epsilon$, if $|i-j| = 1$, for r = 0.05, 0.08, 0.10, 0.12.
- Model 4: Let $\mathbf{R}_1 = \mathbf{I}_p$ and $\mathbf{R}_2 = (r^{|i-j|})_{i,j=1}^{p}$, for r = 0.5, 0.525, 0.55, 0.575.

Table A-1 shows the statistical size under Model 1 and Table A-2 ~ A-4 shows the sizes and powers under Model 2 ~ Model 4, respectively.

Table A-1 Empirical sizes for Model 1 under *Gamma* $-(4, 2)$ population

Empirical size of Model 1							
r	n	p	50	100	200	500	1000
	100		0.041	0.051	0.077	0.087	0.09
0.25	150		0.047	0.062	0.062	0.076	0.083
	200		0.047	0.050	0.051	0.064	0.080
	100		0.059	0.067	0.071	0.079	0.096
0.5	150		0.05	0.046	0.053	0.071	0.072
	200		0.05	0.052	0.06	0.054	0.063
	100		0.067	0.067	0.077	0.08	0.108
0.75	150		0.05	0.051	0.053	0.065	0.081
	200		0.031	0.04	0.042	0.043	0.069
	100		0.0	0.0	0.0	0.0	0.0
1.0	150		0.0	0.001	0.0	0.0	0.0
	200		0.0	0.0	0.0	0.0	0.0

Table A-2 Empirical size and empirical power for Model 2 under $Gamma-(4,2)$ population

			Empirical size of Model 2					Empirical power of Model 2				
ϵ	n	p	50	100	200	500	1000	50	100	200	500	1000
	100		0.043	0.055	0.059	0.063	0.078	0.934	0.995	1.0	1.0	1.0
0.25	150		0.054	0.068	0.049	0.076	0.084	0.992	0.998	1.0	1.0	1.0
	200		0.042	0.043	0.059	0.078	0.075	1.0	1.0	1.0	1.0	1.0
	100		0.052	0.058	0.065	0.077	0.078	0.984	0.993	0.998	1.0	1.0
0.3	150		0.042	0.068	0.073	0.076	0.085	1.0	1.0	1.0	1.0	1.0
	200		0.041	0.047	0.061	0.063	0.071	1.0	1.0	1.0	1.0	1.0
	100		0.054	0.052	0.084	0.093	0.113	0.998	0.999	1.0	1.0	1.0
0.35	150		0.047	0.05	0.061	0.072	0.09	1.0	1.0	1.0	1.0	1.0
	200		0.042	0.047	0.054	0.059	0.071	1.0	1.0	1.0	1.0	1.0
	100		0.049	0.058	0.078	0.083	0.107	0.997	0.998	0.999	1.0	1.0
0.4	150		0.041	0.055	0.063	0.063	0.088	0.998	1.0	1.0	1.0	1.0
	200		0.037	0.055	0.051	0.063	0.074	1.0	1.0	1.0	1.0	1.0

Table A-3 Empirical size and empirical power for Model 3 under $Gamma-(4,2)$ population

			Empirical size of Model 3					Empirical power of Model 3				
ϵ	n	p	50	100	200	500	1000	50	100	200	500	1000
	100		0.056	0.061	0.59	0.073	0.09	0.723	0.719	0.701	0.681	0.631
0.05	150		0.044	0.052	0.059	0.066	0.074	0.977	0.978	0.996	1.0	1.0
	200		0.031	0.065	0.059	0.088	0.075	1.0	1.0	1.0	1.0	1.0
	100		0.065	0.074	0.075	0.097	0.113	0.718	0.724	0.708	0.698	0.673
0.08	150		0.058	0.042	0.051	0.075	0.079	0.981	0.975	0.983	0.969	0.963
	200		0.04	0.048	0.041	0.063	0.072	1.0	1.0	1.0	0.996	0.998
	100		0.054	0.052	0.074	0.073	0.083	0.998	0.999	1.0	1.0	1.0
0.10	150		0.057	0.05	0.061	0.072	0.09	1.0	1.0	1.0	1.0	1.0
	200		0.052	0.057	0.054	0.069	0.071	1.0	1.0	1.0	1.0	1.0
	100		0.062	0.06	0.077	0.093	0.0107	0.699	0.721	0.742	0.772	0.781
0.12	150		0.048	0.056	0.073	0.073	0.098	0.978	0.975	0.968	0.961	0.958
	200		0.05	0.058	0.057	0.063	0.074	1.0	1.0	1.0	1.0	1.0

Table A-4 Empirical size and empirical power for Model 4 under $Gamma-(4,2)$ population

			Empirical size of Model 4					Empirical power of Model 4				
ϵ	n	p	50	100	200	500	1000	50	100	200	500	1000
0.5	100		0.052	0.065	0.065	0.073	0.081	0.949	0.972	0.983	0.989	0.995
	150		0.046	0.061	0.072	0.066	0.081	1.0	1.0	1.0	1.0	1.0
	200		0.05	0.05	0.059	0.068	0.085	1.0	1.0	1.0	1.0	1.0
0.525	100		0.052	0.068	0.079	0.077	0.088	0.969	0.9713	0.981	0.981	0.995
	150		0.038	0.045	0.056	0.07	0.085	1.0	1.0	1.0	1.0	1.0
	200		0.046	0.052	0.068	0.063	0.071	1.0	1.0	1.0	1.0	1.0
0.55	100		0.066	0.071	0.066	0.083	0.093	0.983	0.983	0.986	0.993	0.995
	150		0.047	0.059	0.063	0.082	0.105	0.993	0.998	1.0	1.0	1.0
	200		0.043	0.073	0.054	0.059	0.071	1.0	1.0	1.0	1.0	1.0
0.575	100		0.068	0.061	0.08	0.093	0.107	0.994	1.0	1.0	1.0	1.0
	150		0.046	0.087	0.063	0.053	0.088	0.998	1.0	1.0	1.0	1.0
	200		0.057	0.063	0.069	0.063	0.074	1.0	1.0	1.0	1.0	1.0

哈尔滨工业大学与南方科技大学联合培养研究生

学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《两样本高维稀疏相关性矩阵的检验》，是本人在导师指导下，在学校攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：　　　　　　　　　　日期：　　　年　　月　　日

学位论文使用权限

学位论文是研究生在学校攻读学位期间完成的成果，知识产权归属南方科技大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为南方科技大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：　　　　　　　　　　日期：　　　年　　月　　日

导师签名：　　　　　　　　　　日期：　　　年　　月　　日

Acknowledgments

Time goes by, it is time to say goodbye to my two-year graduate life. My deep gratitude goes first to Professor Guoliang Tian, who is my graduate advior offered me expertly guidiance on research and writing. His great support on my research encourages me a lot to keep moving on this way, and his personal generosity helped me make my time at SUSTech enjoyable.

My appreciation also extends to my cosupervisor, Professor Shurong Zheng, who gave me specific and careful advise on my research topic. Professor Zheng's mentoring and encouragement have been especially valuable, and her insights on extreme value statistic launched the greater part of this dissertation. What's more, she also gave me a lot of advice on how to develop this topic and how to make it into a research paper. This dissertation could never be completed without Professor Zheng's help.

I would also like to acknowledge to the deparment of mathematics and department of statistics and data science of SUSTech. Thanks for the broad platform the departments supplies to us so that we have the oppotunities to contact with great mathematicians and statisticians, and these experience really broaden my horizon and encourage me to pursue some bigger objective. Thanks also goes to my colleagues, who always feel happy to participate my seminar and always give me precious advice on not only my study life but also on my real life.

My deepest appreciation belongs to my family for their patience and understanding. When regards to many questions about my future academic endeavours fomr family and friends I shall answer in the words of Winston Churchill "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning".