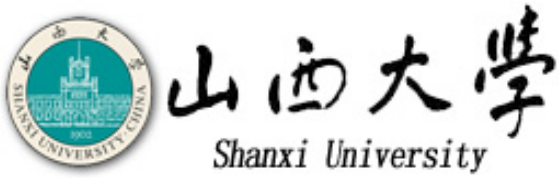


2020 届 硕 士 学 位 论 文

高维数据下的协方差和总体均值检验

作者姓名　　王一静

指导教师　　李顺勇　　教授

学科专业　　统计学

研究方向　　统计机器学习

培养单位　　数学科学学院

学习年限　　2017 年 9 月至 2020 年 6 月

二〇二〇年六月

山西大学

2020 届硕士学位论文

高维数据下的协方差和总体均值检验

作者姓名　　王一静

指导教师　　李顺勇　教授

学科专业　　统计学

研究方向　　统计机器学习

培养单位　　数学科学学院

学习年限　　2017 年 9 月至 2020 年 6 月

二〇二〇年六月

Thesis for Master' s degree, Shanxi University, 2020

Covariance and Population Mean Tests on High-dimensional Data

Student Name	Yi-jing Wang
Supervisor	Prof. Shun-yong Li
Major	Statistics
Specialty	Statistical Machine Learning
Department	School of Mathematical Sciences
Research Duration	2017.09-2020.06

June, 2020

目　录

Contents

中 文 摘 要

随着科技的飞速发展以及大数据时代的到来，统计数据的维数会随着样本量的增加随之增加，有时维数会远远大于样本量，进而造成维数膨胀，使得数学和数据统计分析等学科研究面临重大的挑战压力。

经典的统计理论是在维数固定而样本趋于无穷的假定下推导出来的，而高维数据的稀疏性，这与传统统计理论的假定相违背，因此经典多元统计理论不能直接应用到高维数据上。具体来说，高维数据就是“大 p 小 n”现象。在实际中，仅在维数变高(大 p)这一方面，就需要修正大量传统的多变量分析理论方法和应用程序。因此，在“大 p 小 n”背景下，经典的统计方法与理论很难应用到高维数据问题中，也不再具有强大的效能。所以，对基于高维数据的假设检验问题，寻找优良的检验方法是不可或缺的。

本文在高维数据下，就多元统计分析中的两个基本假设检验问题展开研究：第一个是协方差矩阵相等性检验，第二个是总体均值相等性检验。本篇论文针对这两个高维数据的检验问题，在维数 p 和样本量 n 都趋于无穷的情况下分别提出了新的检验统计方法。

具体来说，在第一个问题中，也就是高维双样本协方差矩阵相等性检验问题，为了使检验方法更有效，本文提出了一个新的检验统计量—T^{New}，同时借助 F-矩阵线性谱统计量的中心极限定理证明了新检验统计量的渐近分布。新的检验方法消除了 Xu 中比例参数的限制，同时在高维数据下正态和非正态的情形都有提高。

对于第二个问题，也就是在高维总体均值假设检验问题中，为了使检验的结果更显著，本文提出了一个新的检验统计量—T_n^*。新的检验统计量不仅摆脱了数据的维度和样本之间大小关系的限制，同时在不同样本所遵循不同分布的 MANOVA 假设检验问题中表现也不错。

数值模拟结果表明，本文提出的新检验统计量T^{New}和T_n^*更具稳健性。

关键词：高维数据；总体均值；协方差矩阵；渐近性；随机矩阵；假设检验；F-矩阵

ABSTRACT

With the rapid development of science and technology and the arrival of the era of big data, the dimension of statistical data will increase with the increase of sample size, sometimes the dimension will be much larger than the sample size, which will cause the dimension expansion, making the research of mathematics, data statistics and analysis facing significant challenges.

The classical statistical theory is derived under the assumption that the dimension is fixed and the sample tends to be infinite. the sparsity of high-dimensional data is contrary to the assumption of traditional statistical theory. Therefore, the classical multivariate statistical theory cannot be directly applied to high dimensional data. Specifically, high dimensional data is the phenomenon of "larger p and small n". In practice, a large number of traditional multivariate analysis theoretical methods and applications need to be modified just to make the dimension higher (larger p). Therefore, under the background of " larger p and small n", it is difficult to apply the classical statistical methods and theories to the problem of high dimensional data, and no longer have strong performance. Therefore, it is necessary to find a good test method for hypothesis testing based on high dimensional data.

In this paper, two basic hypothesis testing problems in multivariate statistical analysis are studied under high dimensional data: the first is the covariance matrix equality test, and the second is the population mean equality test. In this paper, new methods are proposed to test the two high dimensional data under the condition that the dimension and the sample size are both infinite.

Specifically, in the first question, namely high dimensional double sample covariance matrix equality test problems, this paper put forward a new test statistics-T^{New}, and at the same time, based on the F-matrix linear spectrum statistic of the central limit theorem is proved that the new test statistics of progressive distribution. The new test method eliminates the limitation of proportional parameters in Xu, and improves both normal and non-normal conditions in high-dimensional data.

For the second problem, that is the high-dimensional population mean hypothesis test, in order to make the test results more significant, this paper proposes a new test statistic. The new test statistic not only gets rid of the limitation of data dimension and the size relationship between samples, but also performs well in the MANOVA hypothesis test problem with different distribution followed by different samples.

Numerical simulation results show that the two new test statistics T^{New} and T_n^* proposed in this paper are more robust.

Key words： High Dimensional Data; Population Mean; Covariance Matrix; Asymptotic; Random Matrix; Hypothesis Testing; *F*-matrix

第一章 引言

1.1 高维数据分析

随着信息技术的不断发展，不管是工业生产和军事科技，还是日常休闲娱乐，对数据的需求量和数据处理能力的要求越来越高，其中人们在实际操作的过程中，时常碰到的多种多样的海量数据，在统计分析中被称为高维数据。而在高维数据的统计分析过程中遇到最多的问题是样本维数的大量增长，即出现了“维数灾难[1]”的问题。研究表明：通常情况下，分析和处理数据的复杂度和成本会随着维数的升高有着指数增长的趋势。除此之外，对于高维数据，进行统计分析所需的空间样本量也会随维数的增加而急速增长。在处理实际数据时运用传统的多元统计分析方法时，常会碰到非正态分布数据或对数据缺乏先验信息，因此只能采用非参数的方法去处理。而在传统的统计理论中，处理这类问题的非参数方法的前提是大样本数据。另外，通常情形下，高维数据在空间中是非常稀疏的，换句话说，高维数据的样本量与空间的维数相比显得非常少。总的来说，大样本理论在高维数据中并不适用。除此之外，回归分析、主成分分析以及聚类算法中的划分方法和层次方法等经典的数据统计方法，在处理高维数据时也会产生许多的困难，具体来说，维数的增加在很大程度上会加重数据的计算量；高维数据致使数据空间样本量变少，进而使得某些统计量的渐近性质在样本量较小的情况下无法实现；传统的数据统计方法在高维数据下无法保证稳健性等。比如，Hotelling-T^2检验[2]在高维条件下不再具有强大的效能。接下来，通过以下的简单例子来说明高维数据的影响。

例 设 $\mathbf{X}_1,\cdots,\mathbf{X}_n$ 为来自正态分布 $N_p(\boldsymbol{\mu},\mathbf{I})$ 的一个样本。记样本均值为

$$\bar{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i .$$

令 $\bar{x}_i$ 和 μ_i 分别为 $\bar{\mathbf{X}}$ 和 $\boldsymbol{\mu}$ 的第 i 个分量，那么

$$\bar{x}_i-\mu_i \sim N(0,\frac{1}{\sqrt{n}}).$$

所以

$$\bar{x}_i-\mu_i=O_p(\frac{1}{\sqrt{n}}).$$

$$\left\|\bar{\mathbf{X}}-\boldsymbol{\mu}\right\|^2=\sum_{i=1}^{p}(\bar{x}_i-\mu_i)^2=\sum_{i=1}^{p}O_p(\frac{1}{n})=p\cdot O_p(\frac{1}{n}).$$

可以看出，在固定样本量n，维数p逐渐增大的情况下，样本均值$\bar{\mathbf{X}}$作为总体均值$\boldsymbol{\mu}$的估计会变得愈来愈差。所以要使得在p增加的情况下，$\left\|\bar{\mathbf{X}}-\boldsymbol{\mu}\right\|^2\xrightarrow{p}0$依旧成立，样本量$n$就要必须以大于等于$p$的速度增加。这就是常说的维数诅咒，即为了使$p$维的参数有一个更好的估计值，其需要的样本量$n$要随着$p$指数增长。

1.2 随机矩阵理论的研究历程

随机矩阵理论可以追溯到 20 世纪 40 年代和 50 年代初的量子力学中，量子的能级是通过 Hilbert 空间上的厄米特算子(Hermitian)的特征值来表示的，称为哈密顿量。为了避免使用无限维算子，通常通过离散化来逼近量子，相当于截断之后只保留希尔伯特空间中对于所考虑问题重要的部分。因此，大维随机矩阵的极限性质引起了量子力学工作者的特别关注，并在这期间发现了许多规律。自 20 世纪 50 年代末以来，对大维随机矩阵的极限谱分析的研究同样引起了数学家、统计学家以及其他众多专家学者的极大兴趣。其中具有开创性意义的一项工作是 Wigner 提出的高斯矩阵(Wigner 矩阵)的半圆律，证明了大维 Wigner 矩阵的经验谱分布趋向于所谓的半圆律。在这之后，Arnold[3,4]概括了这项具有重要意义的工作，并发现了半圆律以完全确定性的形式可以应用于概率以及削弱了 Wigner 的极强假设的半圆律。Bai 和 Yin[5]证明了当维数相对小于样本量时，进行适当归一化的样本协方差矩阵的谱分布趋于半圆律。继 Marcenko 和 Pastur[6]和 Pastur[7]的工作之后，许多研究人员对大维样本协方差矩阵谱分析的渐近理论进行了发展和完善，其中包括 Bai[8]、Silverstein[9]、Jonsson[10]、Wachter[11,12]以及 Yin[8,13,14]、Krishnaiah[8,14]等。此外，Yin[13,14,15,16]，Bai[15,16]，Krishnaiah[14,15,16]和 Wachter[11]几位学者研究了多变量 *F*-矩阵的极限谱分布，或者更为常见的随机矩阵乘积的极限谱分布。20 世纪 80 年代初，对于某些随机矩阵的极限谱分布(LSD)的存在及其显式形式有了很大进展。近年来，随机矩阵理论的研究方向主要围绕二阶极限定理，其中包括了线性谱统计量的中心极限定理，谱间距的极限分布和极值特征值等研究内容。并且 Bai 等[17]已经提出样本方差矩阵的线性谱统计量的中心极限定理。之后 Zheng[18]基于这个基础研究了随机矩阵中 *F*-矩阵线性谱统计量的中心极限定理。在这些随机矩阵理论的基础上，本文对第一个问题协方差矩阵相

等性检验展开了研究。

1.3 本文的主要工作和结构安排

本文主要针对数据维数远大于样本量的高维总体均值假设检验问题和双样本协方差矩阵相等性两个检验问题，分别提出了新的检验统计量T_n^*和T^{New}，同时分别证明了两个检验统计量的渐近性质。最后都进行了实验模拟，结果表明，本文提出的新检验统计量都有很好的表现。

以下是本文主要内容的结构安排：

第一章绪论主要介绍本文研究背景和进展。首先简要说明了高维数据分析方法在许多科学领域的应用中，所具有的重要意义和重大挑战，进而需要提出新的统计工具和极限理论来解决的总体研究现状。接下来，又对随机矩阵的起源和发展进行了简单的阐述，具体介绍了大维随机矩阵谱分析的主要研究成果，并对大维随机矩阵谱分析的研究现状和发展方向进行了大致说明。

第二章围绕大维随机矩阵理论一些基本概念和主要理论进行具体介绍。其中包含了经验谱分布、线性谱统计量的具体表达形式，还有重要随机矩阵的极限谱分布以及它们各自的线性谱统计量的中心极限定理。

第三章中在数据维数和样本量都趋于无穷的高维双样本协方差矩阵相等性检验问题中，为了使检验方法更有效，本文提出了一个新的检验统计量——T^{New}，同时证明了新检验统计量的渐进分布。新的检验方法消除了 Xu 提出的统计量中比例参数的制约，同时在高维数据下正态和非正态的模型中的效果都有提升。实验模拟的结果表明，本章节提出的新检验统计量更稳健。

第四章中针对不等协方差下的高维多样本均值矩阵的相等性检验问题提出了一个新的检验统计量—T_n^*。新的检验统计量不仅消除了对维数和样本数的限制，也说明了新检验统计量T_n^*分别在原假设和备择假设下的渐近分布，另外在不使用 U 统计量的情况下简化了运算。章节最后进行了数值模拟分析，结果表明新的检验统计量在高维总体均值假设检验方面表现更优。

第五章中对本文的主要工作进行了总结，并对后续研究工作做了一些设想和展望。

第二章 随机矩阵理论

本章主要针对大维随机矩阵理论中的一些基本概念、定理和结果进行了展开介绍。在一些高维数据下的关于协方差矩阵相等性的检验方法中，或多或少都运用了随机矩阵理论，也相应的推动了随机矩阵理论的进一步发展，而随机矩阵理论也扩展到其它领域中，比如电力[21,22]、通信[23,24]以及故障诊断[25]等热门的行业中都运用了随机矩阵理论。以下为经验谱分布与线性谱统计量的定义：

定义 2.1[26]（经验谱分布与线性谱统计量）对任意具有实特征根 λ_i^M，$i=1,2,\cdots,p$ 的 p 阶方阵 M，其矩阵 M 的**经验谱分布**(Empirical spectrum distribution, ESD)可以定义为一个一维的分布函数 F_n^M：

$$F_n^M(x)=\frac{1}{p}\sum_{i=1}^{p}I_{\{\lambda_i^M\le x\}}，\quad x\in\mathbb{R}. \tag{2.1}$$

其中 $I_{\{.\}}$ 代表示性函数，$\mathbb{R}$ 表示实数集合。

随机矩阵 M 的经验谱分布 F_n^M 在一定条件下可以收敛到 F^M，称为极限谱分布(LSD)。为了能更好地对参数 $\theta=\int g(x)dF^M(x)$ 进行统计推断，其中 $g(x)$ 是随机变量函数，一般使用以下统计量：

$$\hat{\theta}=\int g(x)dF_n^M(x)=\frac{1}{p}\sum_{i=1}^{p}f(\lambda_i^M). \tag{2.2}$$

来估计，这里的 $\hat{\theta}$ 称为随机矩阵 M 的**线性谱统计量**(Linear spectral statistics, LSS)。

2.1 F-矩阵极限谱分布

2.1.1 样本协方差矩阵极限谱分布

在多元统计分析中，样本协方差矩阵不仅是假设检验、主成分分析、因子分析和判别分析的基础，另外许多检验统计量是由其特征值定义的。由此可见，样本协方差矩阵是很重要的随机矩阵。其中，样本协方差矩阵的定义如下：

假设存在来自服从 p 维正态分布 $N(\mu_p,\Sigma_p)$ 的一个样本 $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n)$，其中 $\mathbf{x}_i=(x_{1i},x_{2i},\cdots,x_{pi})'$。则该样本的样本协方差矩阵为：

$$\mathbf{S}=\frac{1}{n}\sum_{i=1}^{p}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'.$$

其中：$\bar{\mathbf{x}}=1/n\sum\mathbf{x}_i$。在大维随机矩阵谱分析理论中，样本协方差矩阵经过简化可

写为：

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{p}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\mathbf{X}\mathbf{X}'.$$

这是由于$\bar{\mathbf{x}}\,\bar{\mathbf{x}}'$是秩为 1 的矩阵，协方差矩阵的极限谱分布与$\bar{\mathbf{x}}\,\bar{\mathbf{x}}'$并无联系，故可以省去。

在大维样本协方差矩阵的谱分析中，通常假设维数 p 与样本量 n 成正比，即 $p/n \to y \in (0,\infty)$。在 1967 年，Marcenko 和 Pastur[6]就提出了样本协方差矩阵的极限谱分布(MP 律)。而现在最新的理论结果是 Bai[26]在 1999 年提出的大维样本协方差矩阵的极限谱分布，是在 Yin[13]的基础上考虑了复随机变量，对样本协方差矩阵极限谱分布进行了一个由实向复的推广。具体表述如下：

定理 2.1[26]假设 $\mathbf{X} = \{\xi_{lj}, 1 \le l \le p, 1 \le j \le n\}$ 是由均值为 0，方差为 σ^2 的独立同分布的复随机变量数组，则当 $p/n \to c \in (0,1)$ 时，样本协方差矩阵 $\mathbf{S}_n$ 的经验谱分布以概率 1 收敛到极限谱分布 $F_c(x)$，其中 $F_c(x)$ 的密度函数为：

$$f_c(x) = \begin{cases} \dfrac{1}{2\pi x c\sigma^2}\sqrt{(b(\sigma^2,c)-x)(x-a(\sigma^2,c))}, & a(\sigma^2,c) \le x \le b(\sigma^2,c), \\ 0, & \text{其他}. \end{cases}$$

其中：$a(\sigma^2,c) = \sigma^2(1-\sqrt{c})^2$，$b(\sigma^2,c) = \sigma^2(1+\sqrt{c})^2$；且当 $c>1$ 时，在原点处有点测度 $1-1/y$。

当 σ^2=1时，定理 2.1 就是标准的 MP 律。之后，Bai 和 Silberstein[27]对上述定理进行了扩展，在样本协方差矩阵元素独立但不同分布的情况下，在定理 2.1 中同样的假设基础上，增加了一个假设条件：对任意 $\varepsilon > 0$，

$$\frac{1}{\varepsilon np}\sum_{lj} E\left|\xi_{lj}\right|^2 I\left(\left|\xi_{lj}\right| \ge \varepsilon\sqrt{n}\right) \to 0. \tag{2.3}$$

使定理 2.1 的结论(MP 律)仍然成立。

2.1.2 *F*−矩阵极限谱分布

定义 2.2[27]：（*F*-矩阵）考虑两个 p 维独立总体

$$\mathbf{X} = (\xi_1, \xi_2, \cdots, \xi_p)', \quad \mathbf{Y} = (\eta_1, \eta_2, \cdots, \eta_p)'.$$

其均值向量和协方差矩阵分别为：

$$E(\mathbf{X}) = E(\mathbf{Y}) = \mathbf{0}_p, \quad Cov(\mathbf{X}) = Cov(\mathbf{Y}) = \mathbf{I}_p.$$

设 $\boldsymbol{\xi}_j = (\xi_{1j}, \xi_{2j}, \cdots, \xi_{pj})'$，$j = 1,2,\cdots,n_1$，$\boldsymbol{\eta}_k = (\eta_{1k}, \eta_{2k}, \cdots, \eta_{pk})'$，$k = 1,2,\cdots,n_2$ 分别为

来自总体 $\mathbf{X}$ 和 $\mathbf{Y}$ 的样本，则两者的样本协方差矩阵分别为：

$$\mathbf{S}_{n_1}=\frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{\xi}_j\boldsymbol{\xi}_j',\quad \mathbf{S}_{n_2}=\frac{1}{n_2}\sum_{k=1}^{n_2}\boldsymbol{\eta}_k\boldsymbol{\eta}_k'.$$

定义 ***F*-矩阵**为：

$$\mathbf{F}_{n_1,n_2}=\mathbf{S}_{n_1}\mathbf{S}_{n_2}^{-1}.$$

其中 $n_2>p$，是因为要使 $\mathbf{S}_{n_2}$ 的逆矩阵存在。

关于 *F*-矩阵的极限谱分布的研究，最早是 Wachter[11]于 1980 年在 $\mathbf{S}_{n_1}$ 和 $\mathbf{S}_{n_2}$ 是相互独立的 Wishart 矩阵的条件下提出的，其具体形式在 Bai,Yin 和 Krishnaiah[28] 以及 Silberstein[29]的研究结果中都有描述。另外，去掉 $\mathbf{S}_{n_1}$ 和 $\mathbf{S}_{n_2}$ 是相互独立的 Wishart 矩阵的这个条件，在 Bai 和 Yin[5]对样本协方差最小特征值的强极限和 Yin(1986)[13]的结果基础上，并运用 Bai,Yin 和 Krishnaiah[17]在 1987 年工作中的方法，证明了 $\mathbf{S}_{n_1}$ 和 $\mathbf{S}_{n_2}$ 在适当的矩条件下，*F*-矩阵的极限谱分布与 $\mathbf{S}_{n_1}$ 和 $\mathbf{S}_{n_2}$ 是 Wishart 矩阵的假设条件下有相同的结果。下面给出 *F*-矩阵的极限谱分布的密度函数：

记 $F_{n_1,n_2}^{\mathbf{F}_{n_1,n_2}}$ 是矩阵 $\mathbf{F}_{n_1,n_2}$ 的经验谱分布，F^{c_1,c_2} 是矩阵 $\mathbf{F}_{n_1,n_2}$ 的极限谱分布。Bai[28] 给出了 *F*-矩阵的极限谱分布的具体形式，其密度函数如下：

$$f^{c_1,c_2}(x)=\begin{cases}\dfrac{(1-c_2)\sqrt{(b-x)(x-a)}}{2\pi x(c_1+c_1x)}, & a\le x\le b,\\ 0, & \text{其他.}\end{cases}\tag{2.4}$$

其中：$a=(1-h)^2/(1-c_2)^2$，$b=(1+h)^2/(1-c_2)^2$，$h=\sqrt{c_1+c_2-c_1c_2}$。

2.2*F*-矩阵线性谱统计量的中心极限定理

2.2.1 样本协方差矩阵线性谱统计量的中心极限定理

考虑一个 p 维独立总体

$$\mathbf{X}=(\xi_1,\xi_2,\cdots,\xi_p)'.$$

其均值向量和协方差矩阵分别为：

$$E(\mathbf{X})=\mathbf{0}_p,\quad Cov(\mathbf{X})=\mathbf{I}_p.$$

设 $\boldsymbol{\xi}_j=(\xi_{1j},\xi_{2j},\cdots,\xi_{pj})'$，$j=1,2,\cdots,n$，为来自总体 $\mathbf{X}$ 的样本，则其样本协方差矩阵为：

$$\mathbf{S}_n = \frac{1}{n}\sum_{j=1}^{n_1}\boldsymbol{\xi}_j\boldsymbol{\xi}_j' .$$

令

$$c_n = \frac{p}{n} \to c \in (0,1).$$

F^c 和 F^{c_n} 分别表示参数为 c 和 c_n 的样本协方差矩阵 $\mathbf{S}_n$ 的 MP 律。定义 u 是复平面上的一个开集合，其中包含 $[I_{(0,1)}(y)(1-\sqrt{y})^2,(1+\sqrt{y})^2]$，定义 $\mathcal{A}$ 为一族函数的集合，且 $\mathcal{A}=\{g_1,g_2,\cdots\}$。函数 $g_1,g_2,\cdots$ 均在开集合 u 上解析。在 $\mathcal{A}$ 上定义随机过程 $G_n := \{G_n(g_1),G_n(g_2),\cdots,G_n(g_p)\}$：

$$G_n(g) = p\cdot\int_{-\infty}^{+\infty} g(x)[F_n - F^{c_n}]d(x), \qquad g \in \mathcal{A}.$$

其中 F_n 是样本协方差矩阵 $\mathbf{S}_n$ 的经验谱分布。

之后，对于大维样本协方差矩阵的线性谱统计量的中心极限定理，Bai 和 Silberstein[30]在 2004 年提出了具体的表达形式。

定理 2.3[30]假设：$f_1,f_2,\cdots,f_k \in \mathcal{A}$，$\{\xi_{ij}\}$ 是独立同分布的随机变量，同时 $E\xi_{11}=0$，$E|\xi_{11}|^2=1$，$E|\xi_{11}|^4<\infty$，当 $n\to\infty, p\to\infty$ 时，有 $p/n\to c\in(0,1)$，那么

(i) 设 $\{\xi_{ij}\}$ 是实随机变量并且 $E(\xi_{11}^4)=3$。那么随机变量 $\big(G_n(g_1),G_n(g_2),\cdots,G_n(g_k)\big)$ 弱收敛于 p 维的正态变量，其均值为

$$E[G_n(g_l)] = m(g_l) = \frac{g_l(a(c))+g_l(b(c))}{4} - \frac{1}{2\pi}\int_{a(c)}^{b(c)}\frac{g_l(x)}{\sqrt{4y-(x-1-c)^2}}dx, \qquad l=1,\cdots,p$$

其中 $a(c)=(1-\sqrt{c})^2$，$b(c)=(1+\sqrt{c})^2$。$G_n(g_l)$ 和 $G_n(g_m)$ 的协方差为：

$$Cov(g_l,g_k) = -\frac{1}{2\pi^2}\oint\oint\frac{g_l(z_1)+g_k(z_2)}{(\hat{m}(z_1)-\hat{m}(z_2))^2}d\hat{m}(z_1)d\hat{m}(z_2), \qquad l,k=\{1,\cdots,p\}. \quad (2.5)$$

其中：$\hat{m}(z)\equiv m_{\hat{F}^c}(z)$ 是 $\hat{F}^c \equiv (1-c)I_{[0,\infty)}+cF^c$ 的 Stieltjes 变换；“$\oint$”表示对封闭曲线求积分。(2.5)中的积分曲线互不重叠且都包含 F^c 的支撑集。

(ii) 设 $\{\xi_{ij}\}$ 是实随机变量并且 $E(\xi_{11}^2)=0$，$E(|\xi_{11}|^4)=2$。那么(i)中的结论依然成立，只是均值为 $\mathbf{0}$，协方差矩阵为(2.5)的一半，即：

$$Cov(g_l,g_k) = -\frac{1}{2\pi^2}\oint\oint\frac{g_l(z_1)+g_k(z_2)}{(\hat{m}(z_1)-\hat{m}(z_2))^2}d\hat{m}(z_1)d\hat{m}(z_2), \qquad l,k=\{1,\cdots,p\}. \quad (2.6)$$

2.2.2 *F*-矩阵线性谱统计量的中心极限定理

基于 Bai 和 Silverstein 的定理 2.3 以及 *F*-矩阵的经验谱分布，给出了大维 *F*-

矩阵的线性谱统计量的中心极限定理，具体形式在 Zheng[18]的成果中都有描述。

F^{c_1,c_2} 所对应的支撑集是复平面上一个包含区间

$$[a'=\left((1-\sqrt{c_1})/(1+\sqrt{c_2})\right)^2, b'=\left((1+\sqrt{c_1})/(1-\sqrt{c_2})\right)^2].$$

的开集合，定义 $\tilde{\mathcal{A}}$ 为一族函数的集合，且 $\tilde{\mathcal{A}}=\{g_1,g_2,\cdots\}$。函数 $g_1,g_2,\cdots$ 均在包含区间 $[a',b']$ 的开区域上解析。在 $\tilde{\mathcal{A}}$ 上定义经验过程 $\tilde{G}_n:=\{\tilde{G}_n(g_1),\tilde{G}_n(g_2),\cdots,\tilde{G}_n(g_p)\}$：

$$\tilde{G}_n(g)=p\cdot\int_{-\infty}^{+\infty}g(x)[F_n^{\mathbf{F}_{n_1,n_2}}(x)-F^{c_{n_1},c_{n_2}}(x)]d(x),\qquad g\in\tilde{\mathcal{A}}.$$

这里 $F^{c_{n_1},c_{n_2}}(x)$ 是把(2.4)中参数 c_1,c_2 换成 c_{n_1},c_{n_2} 的极限分布。

为研究 *F*-矩阵的线性谱统计量的中心极限定理，给出如下两个假设[27]：

(I) 对于任意固定的 $\varepsilon_0>0$，有

$$\frac{1}{n_1p}\sum_{l=1}^{p}\sum_{j=1}^{n_1}E\left|\xi_{jl}\right|^4I(\left|\xi_{jl}\right|\geq\varepsilon_0\sqrt{n_1})\to0,$$

$$\frac{1}{n_2p}\sum_{l=1}^{p}\sum_{k=1}^{n_2}E\left|\eta_{jk}\right|^4I(\left|\eta_{jk}\right|\geq\varepsilon_0\sqrt{n_2})\to0,$$

其中 $I(\bullet)$ 为示性函数。

(II) $p\to+\infty,n_1\to+\infty,c_{n_1}=\dfrac{p}{n_1}\to c_1\in(0,+\infty)$；

$$p\to+\infty,n_2\to+\infty,c_{n_2}=\frac{p}{n_2}\to c_2\in(0,1)$$

下面介绍 Zheng[18]提出的 *F*-矩阵的线性谱统计量的中心极限定理(Central limit theorem, CLT)。

定理 2.4[18]：在 *F*-矩阵定义的记号下，如果假设(I)和(II)成立，对于所有的 j,k，有

$$E\xi_{jk}=E\eta_{jk}=0,\qquad E\left|\xi_{jk}\right|^2=E\left|\eta_{jk}\right|^2=\tau-1.$$

$$E\left|\xi_{jk}\right|^4=\beta_x+\tau+1<\infty,\quad E\left|\eta_{jk}\right|^4=\beta_y+\tau+1<\infty.$$

其中 β_x 和 β_y 是与四阶矩相关的常数，τ 满足以下条件：

$$\tau=\begin{cases}2,\ \xi_{jk},\eta_{jk}\text{为实变量},\\ 1,\ \xi_{jk},\eta_{jk}\text{为复变量}.\end{cases}$$

取 $\tau=2$，令 $g_1,g_2,\cdots,g_p\in\mathcal{A}$，则随机向量 $(\tilde{G}_n(g_1),\tilde{G}_n(g_2),\cdots,\tilde{G}_n(g_p))$ 依分布收敛于 p 维正态向量。且有：

$$E[\tilde{G}_n(g_l)]=\mu(g_l)=\lim_{r\to 1_+}\frac{1}{4\pi i}\oint_{|\zeta|=1} g_l\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta)}{(1-c_2)^2}\right)\left(\frac{1}{\zeta-\frac{1}{r}}+\frac{1}{\zeta+\frac{1}{r}}-\frac{2}{\zeta+\frac{c_2}{hr}}\right)d\zeta$$

$$+\frac{\beta_x c_1(1-c_2)^2}{2\pi i\cdot h^2}\oint_{|\zeta|=1} g_l\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta)}{(1-c_2)^2}\right)\frac{1}{(\zeta+\frac{c_2}{h})^3}d\zeta$$

$$+\frac{\beta_y c_1(1-c_2)}{2\pi i\cdot h}\oint_{|\zeta|=1} g_l\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta)}{(1-c_2)^2}\right)\frac{\zeta+1/h}{(\zeta+\frac{c_2}{h})^3}d\zeta\ ,l=1,2,\cdots,p\,.\ (2.7)$$

其中：$\mathrm{Re}(\zeta)=(\zeta+\bar{\zeta})/2$，$\bar{\zeta}=1/\zeta$，对任意 ζ 满足 $|\zeta|=1$；$h=\sqrt{c_1+c_2-c_1c_2}$；“$\oint$”表示对封闭曲线求积分。

$\tilde{G}_n(g_l)$ 和 $\tilde{G}_n(g_m)$ 的协方差为：

$$Cov(g_l,g_m)=-\lim_{r\to 1_+}\frac{1}{2\pi}\oint_{|\zeta_1|=1}\oint_{|\zeta_2|=1}\frac{g_l\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta_1)}{(1-c_2)^2}\right)g_m\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta_2)}{(1-c_2)^2}\right)}{(\zeta_1+r\zeta_2)^2}$$

$$-\frac{(\beta_x c_1+\beta_y c_2)(1-c_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{g_l\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta_1)}{(1-c_2)^2}\right)}{(\zeta_1+\frac{c_2}{h})^2}d\zeta_1\oint_{|\zeta_2|=1}\frac{g_m\left(\frac{1+h^2+2h\,\mathrm{Re}(\zeta_2)}{(1-c_2)^2}\right)}{(\zeta_2+\frac{c_2}{h})^2}d\zeta_2$$

$$l,m=1,2,\cdots,p\,.\qquad(2.8)$$

其中：$r\to 1_+$ 表示 r 从右趋于 1。

第三章　基于 *F*-矩阵的高维协方差矩阵相等性检验

现实生活中会遇到方方面面的问题，譬如生物学、神经科学以及刑事技术和物证检验等问题[31,32]，这些问题经常会涉及到多元回归分析中的多元随机向量及其相关统计分析。总体的协方差矩阵之间的相等性检验是进行多个多元分布的总体统计推断的一个重要过程，因为许多多元统计分析方法都是需要满足多个总体协方差矩阵的齐次性，即要在协方差矩阵相等的前提下展开。综上所述，协方差矩阵的齐次性检验是一个具有重要意义的研究内容。

3.1 协方差矩阵相等性检验问题介绍

考虑两个 p 维独立总体 $\mathbf{X}=(\xi_1,\xi_2,\cdots,\xi_p)'$ 和 $\mathbf{Y}=(\eta_1,\eta_2,\cdots,\eta_p)'$，其均值向量和协方差矩阵满足以下：

$$E(\mathbf{X})=\mu_1,\quad Cov(\mathbf{X})=\Sigma_1.$$

$$E(\mathbf{Y})=\mu_2,\quad Cov(\mathbf{Y})=\Sigma_2.$$

假设 $\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\ldots,\boldsymbol{\xi}_{n_1}$ 是来自总体 $\mathbf{X}$ 的样本，其中记 $\boldsymbol{\xi}_j=(\xi_{1j},\xi_{2j},\cdots,\xi_{pj})',j=1,2,\cdots,n_1$，而 $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_{n_2}$ 是来自总体 $\mathbf{Y}$ 的样本，其中 $\boldsymbol{\eta}_k=(\eta_{1k},\eta_{2k},\cdots,\eta_{pk})'$，$k=1,2,\cdots,n_2$。下面首先提出两个总体协方差矩阵的比例性检验，即如下假设检验：

$$H_0:\Sigma_1=e\Sigma_2 \qquad \text{Vs} \qquad H_1:\Sigma_1\neq e\Sigma_2 \tag{3.1}$$

关于协方差矩阵的比例性检验由来已久，其中大多数研究都是基于维数 p 固定的经典极限定理展开的。Federer(1951)[33]对两个维数 p 小于 3 的独立同分布的样本利用极大似然方法进行了检验。Pillai[34]等(1969)就一个样本协方差矩阵和另一个样本协方差矩阵的逆矩阵乘积，研究了其特征根的比值分布。但是，对于 $k>2$ 的总体没有较为满意的处理方法。之后，Flury(1986)[35]针对 $k\geq 2$ 的总体进行推广并构造了似然比统计量，另外证明了该统计量在维数 p 固定和样本量 n_1,n_2 均为无穷大的情况下的渐近性质。但对于非正态的数据，检验统计量的总体性能并不优。为了考虑非正态情形，Schott(1999)[36]对 k 个具有有限四阶矩的维数固定样本进行了 Wald 检验。

当(3.1)中的参数 $e=1$ 时，就是本章节主要的研究内容，高维数据下两个总体

协方差矩阵的相等性检验问题，即如下的假设检验问题：

$$H_0:\Sigma_1=\Sigma_2 \qquad \text{Vs} \qquad H_1:\Sigma_1\neq\Sigma_2 \tag{3.2}$$

目前关于双样本协方差矩阵的相等性检验的研究已有不少成果。其中，在低维数据下的双样本协方差矩阵相等性检验的研究已经大致成熟，其研究成果在 Gupta D S [37, 38]、Gupta A K 和 Tang J(1984) [39]、O'Brien P C[40]、Perlman 和 Michael D[41]、以及 Sugiura N 和 Nagao H[42, 43]等学者的工作中都有展现。另外，John S[44]和 Hisao N[45]两位学者分别讨论了在固定维数下的经典假设（即 $p<n$ ）和在正态总体下的检验方法，不过这类方法在高维数据的情况下并不适用。

随机矩阵理论的兴起不断发展给大维数据下的统计分析问题提供了新的研究思路。其中被广为认可的有：Bai 等(2009)[27]和 Jiang 等(2012)[46]提出的大维协方差矩阵的相关检验方法；Chen 等(2010)[47]给出的在维数和样本量同时趋于无穷大情况下的检验方法。随后，Xu[19]就高维数据下的两个大维总体协方差矩阵的比例性及相等性的检验问题提出了一种伪似然比检验方法，但美中不足的是检验结果受比例参数的限制。Jiang[20]又提出了适用于高维情形下非正态分布的修正得分检验方法。除此之外，何冰等人[48]根据 Jiang[20]中统计量的形式提出一个检验统计量，并证明了检验统计量的渐近性，但遇到维数相对于样本量较大的情况，何冰等人[48]提出的方法效果并不理想。基于这些研究理论结果，本章节就检验问题(3.2)提出了新的检验统计量。

3.2 新检验统计量及其中心极限定理

3.2.1 T^X 和 T^J

关于检验问题(3.2)，Xu[19]等在 2014 年提出了一种伪似然比检验，其统计量记为 T^X ，如下所示：

$$T^X=p\log\Big(\frac{1}{p}tr\big(\frac{1}{c}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}\big)\Big)-\log\left|\frac{1}{c}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}\right|,\ \forall c>0. \tag{3.3}$$

其中， c 是比例参数； $\hat{\boldsymbol{\Sigma}}_i(i=1,2)$ 分别为各自样本协方差矩阵的极大似然估计，即：

$$\hat{\boldsymbol{\Sigma}}_1=\frac{1}{n_1-1}\sum_{j=1}^{n_1}(\boldsymbol{\xi}_j-\bar{\boldsymbol{\xi}})(\boldsymbol{\xi}_j-\bar{\boldsymbol{\xi}})',\ \ \bar{\boldsymbol{\xi}}=\frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{\xi}_j.$$

$$\hat{\boldsymbol{\Sigma}}_2=\frac{1}{n_2-1}\sum_{k=1}^{n_2}(\boldsymbol{\eta}_k-\bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_k-\bar{\boldsymbol{\eta}})',\ \ \bar{\boldsymbol{\eta}}=\frac{1}{n_2}\sum_{k=1}^{n_2}\boldsymbol{\eta}_k.$$

Xu[19]等提出的检验统计量 T^X 是主要是针对两个总体协方差矩阵的比例性检验

问题(3.1)，但对于协方差矩阵相等性检验来说，比例参数 c 的存在会降低检验性能。也就是说，在检验问题(3.2)中，T^X 需要先用一些样本来消除比例参数 c 的影响，再进行检验，这样就会导致检验效果受到影响。

同样是协方差矩阵相等性检验问题(3.2)，Jiang[20]对协方差矩阵的经典得分检验进行修正得到了修正 Rao 得分检验，记为 T^J，具体表达如下：

$$T^J = tr[(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1} - \mathbf{I}_p)^2]. \tag{3.4}$$

值得我们注意的是，修正的 Rao 得分检验运用的大维样本协方差矩阵的线性谱统计量的中心极限定理是在更为宽泛条件，即当 $p, n_1 \to +\infty$ 时，$c_{n_1} = p/n_1 \to c_1 \in (0, +\infty)$ 下提出的，因此，这样的修正得分检验可以更广泛地应用，并得到很好的检验效果。

3.2.2 新检验统计量 T^{New}

本章节就双样本的协方差相等性检验问题，利用随机矩阵理论，对 Xu[19]等提出的 T^X 中的第一部分借鉴 Jiang[20]提出的修正 Rao 得分检验，同时去除 T^X 中的比例参数 c，以此对 Xu[19]等提出的伪似然比检验 T^X 进行改进。显而易见，$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ 等价于 $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1} = \mathrm{I}_p$，故本章借用一个总体下的协方差矩阵球度检验思想，构造出一个新的检验统计量，具体表达式如下：

$$T^{New} = tr(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1})^2 - \log(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1}). \tag{3.5}$$

为了方便之后叙述，设两个 p 维独立总体 $\mathbf{X}$ 和 $\mathbf{Y}$ 满足：

$$E(\mathbf{X}) = \boldsymbol{\mu}_1, \quad Cov(\mathbf{X}) = \mathbf{\Sigma}_1.$$

$$E(\mathbf{Y}) = \boldsymbol{\mu}_2, \quad Cov(\mathbf{Y}) = \mathbf{\Sigma}_2.$$

记 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{n_1}$ 为来自于 $\mathbf{X}$ 的样本，$\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{n_2}$ 为来自于 $\mathbf{Y}$ 的样本。接下来通过一个定理给出了 T^{New} 的渐近性质。

定理 3.1 假设定理 2.4 中的假设条件均成立，令

$$\boldsymbol{\xi}_j = \mathbf{\Sigma}_1^{-1/2}(\mathbf{x}_j - \boldsymbol{\mu}_1),\ j = 1, 2, \cdots, n_1; \quad \boldsymbol{\eta}_k = \mathbf{\Sigma}_2^{-1/2}(\mathbf{y}_k - \boldsymbol{\mu}_2),\ k = 1, 2, \cdots, n_2.$$

则在原假设 $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ 下，当 $n \to \infty$ 时，检验统计量 T^{New} 有以下结论：

$$T^{New} - p \cdot F^{c_{n_1}, c_{n_2}}(g) \xrightarrow{d} N(\mu(g), Cov(g)). \tag{3.6}$$

即：

$$Cov(g)^{-1/2}\{T^{New}-p\cdot F^{c_{n_1},c_{n_2}}(g)-\mu(g)\}\xrightarrow{d}N(0,1). \tag{3.7}$$

其中： $g(x)=x^2-\log(x)$； $p\cdot F^{c_{n_1},c_{n_2}}(g)=p\cdot\int_a^b g(x)\cdot f^{c_{n_1},c_{n_2}}(x)d(x)$。

证明：根据线性谱统计量的定义对检验统计量T^{New}进行变形得到：

$$tr(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1})^2-\log(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1})=\sum_{i=1}^{p}\left((\lambda_i^{\mathbf{F}_{n_1,n_2}})^2-\log(\lambda_i^{\mathbf{F}_{n_1,n_2}})\right)=p\cdot\int(x^2-\log(x))dF_{n_1,n_2}^{\mathbf{F}_{n_1,n_2}}(x)$$

$$=p\cdot\int g(x)d\{F_{n_1,n_2}^{\mathbf{F}_{n_1,n_2}}(x)-F^{c_{n_1},c_{n_2}}(x)\}+p\cdot F^{c_{n_1},c_{n_2}}(g). \tag{3.8}$$

其中 $\lambda_i^{\mathbf{F}_{n_1,n_2}}$ 和 $F_{n_1,n_2}^{\mathbf{F}_{n_1,n_2}}(x)$ 分别为 F 矩阵的特征值和经验谱分布。

则由定理 2.4 可知：随机向量

$$p\cdot\int g(x)d\{F_{n_1,n_2}^{\mathbf{F}_{n_1,n_2}}(x)-F^{c_{n_1},c_{n_2}}(x)\}=tr(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1})^2-\log(\hat{\mathbf{\Sigma}}_1\hat{\mathbf{\Sigma}}_2^{-1})-p\cdot F^{c_{n_1},c_{n_2}}(g).$$

依分布收敛于一个 p 维的正态向量 $N(\mu(g),Cov(g))$。

令 $g(x)=g_1(x)-g_2(x)$，且 $g_1(x)=x^2$， $g_2(x)=\log(x)$。下面就 $g_1(x)$ 和 $g_2(x)$ 分别展开证明。

(i) 考虑 $g_1(x)=x^2$，令 $x(\zeta)=\dfrac{1+h^2+2h\,\mathrm{Re}(\zeta)}{(1-c_2)^2}$。

根据柯西积分公式，对给出的任意 $r>1$，有

$$\oint_{|\xi|=1}\left(\frac{1}{\xi-\frac{1}{r}}+\frac{1}{\xi+\frac{1}{r}}-\frac{2}{\xi+\frac{c_2}{hr}}\right)d\xi=0,\quad \oint_{|\xi|=1}\frac{1}{(\xi+\frac{c_2}{h})^3}d\xi=0.$$

另外，经简化后， $g_1(x)$ 的均值如下：

$$\mu(g_1)=\lim_{r\to 1_+}\frac{1}{4\pi i}\oint_{|\zeta|=1}\left(x(\zeta)\right)^2\left(\frac{1}{\zeta-\frac{1}{r}}+\frac{1}{\zeta+\frac{1}{r}}-\frac{2}{\zeta+\frac{c_2}{hr}}\right)d\zeta$$

$$+\frac{\beta_x c_1(1-c_2)^2}{2\pi i\cdot h^2}\oint_{|\zeta|=1}\left(x(\zeta)\right)^2\frac{1}{(\zeta+\frac{c_2}{h})^3}d\zeta$$

$$+\frac{\beta_y c_1(1-c_2)}{2\pi i\cdot h}\oint_{|\zeta|=1}\left(x(\zeta)\right)^2\frac{\zeta+1/h}{(\zeta+\frac{c_2}{h})^3}d\zeta.$$

则 $g_1(x)$ 的均值可以分成以下三个部分。

其中第一部分：

$$I_1=\oint_{|\zeta|=1}\big(x(\zeta)\big)^2\Big(\frac{1}{\zeta-\frac{1}{r}}+\frac{1}{\zeta+\frac{1}{r}}-\frac{2}{\zeta+\frac{c_2}{hr}}\Big)d\zeta$$
$$=\frac{(c_2+3)h^2}{(1-c_2)^4}+\frac{c_1+4c_2}{(1-c_2)^2}-\frac{c_1}{(1-c_2)^3}.$$

第二部分：

$$\beta_x I_2=\frac{\beta_x c_1(1-c_2)^2}{2\pi i\cdot h^2}\oint_{|\zeta|=1}\big(x(\zeta)\big)^2\frac{1}{(\zeta+\frac{c_2}{h})^3}d\zeta$$

$$=\beta_x c_1\Big(\frac{2c_1h^2+(c_1+c_2)^2}{c_2^4}+\frac{2h^4+2h^2}{c_2^3(1-c_2)^2}-\frac{3h^4}{c_2^4(1-c_2)^2}\Big).$$

其中 $I_2=c_1\Big(\frac{2c_1h^2+(c_1+c_2)^2}{c_2^4}+\frac{2h^4+2h^2}{c_2^3(1-c_2)^2}-\frac{3h^4}{c_2^4(1-c_2)^2}\Big)$。

第三部分：

$$\beta_y I_3=\frac{\beta_y c_1(1-c_2)}{2\pi i\cdot h}\oint_{|\zeta|=1}\big(x(\zeta)\big)^2\frac{\zeta+1/h}{(\zeta+\frac{c_2}{h})^3}d\zeta$$

$$=\beta_y\Big(\frac{h^4}{c_2(1-c_2)^3}-\frac{2c_1c_2^2+(c_1+c_2)^2}{c_2^2(1-c_2)}+\frac{c_2I_2}{c_2(1-c_2)}\Big).$$

其中 $I_3=\frac{h^4}{c_2(1-c_2)^3}-\frac{2c_1c_2^2+(c_1+c_2)^2}{c_2^2(1-c_2)}+\frac{c_2I_2}{c_2(1-c_2)}$。

综合以上三个部分可得：

$$\mu(g_2)=I_1+\beta_x I_2+\beta_y I_3.$$

(ii)考虑 $g_2(x)=\log x$，可知 $\mu(g_2)=\frac{1}{2}\log\frac{1-h^2}{(1-c_2)^2}-\frac{\beta_x c_1}{2}+\frac{\beta_y c_2}{2}$.

综合(i)(ii)可得：

$$\mu(g)=\mathrm{I}_1-\frac{1}{2}\log\frac{1-h^2}{(1-c_2)^2}+\beta_x(\mathrm{I}_2+\frac{c_1}{2})+\beta_y(\mathrm{I}_3-\frac{c_2}{2}).$$

以上完成了均值的证明，以下进行协方差阵的证明。由定理 2.4，其协方差 $Cov(g)$ 同样也可分成三部分，具体如下。

第一部分：

$$Cov(g_1,g_1)=-\lim_{r\downarrow 1}\frac{1}{2\pi}\oint_{|\zeta_1|=1}\int_{|\zeta_2|=1}\frac{g_1(x(\zeta_1))g_1(x(\zeta_2))}{(\zeta_1-r\zeta_2)^2}d\zeta_1 d\zeta_2$$

$$-\frac{(\beta_x c_1+\beta_y c_2)(1-c_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{g_1(x(\zeta_1))}{(\zeta_1+\frac{c_2}{h})^2}d\zeta_1\times\oint_{|\zeta_2|=1}\frac{g_1(x(\zeta_2))}{(\zeta_2+\frac{c_2}{h})^2}d\zeta_2$$

$$=\mathrm{I}_4+(\beta_x c_1+\beta_y c_2)\mathrm{I}_5$$

其中：

$$\mathrm{I}_4=\frac{8h^2(1+h^2)^2+8h^4}{(1-c_2)^8}-\frac{8h^2(1+h^2)}{(1-c_2)^6}-\frac{4h^2}{(1-c_2)^4}$$

$$\mathrm{I}_5=\frac{4h^4}{c_2^2(1-c_2)^6}-\frac{4(c_1+c_2)^2-6c_2^2}{c_2^2(1-c_2)^2}-\frac{8(c_1+c_2)-8hc_2^2(1+h^2)}{c_2^2(1-c_2)^4}$$

第二部分：

$$Cov(g_1,g_2)=-\lim_{r\to 1_+}\frac{1}{2\pi}\oint_{|\zeta_1|=1}\oint_{|\zeta_2|=1}\frac{g_1(x(\zeta_1))g_2(x(\zeta_2))}{(\zeta_1-r\zeta_2)^2}d\zeta_1 d\zeta_2$$

$$-\frac{(\beta_x c_1+\beta_y c_2)(1-c_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{g_1(x(\zeta_1))}{(\zeta_1+\frac{c_2}{h})^2}d\zeta_1\oint_{|\zeta_2|=1}\frac{g_2(x(\zeta_2))}{(\xi_2+\frac{c_2}{h})^2}d\zeta_2$$

$$=\frac{4h^2(1+h^2)}{(1-c_2)^4}+(\beta_x c_1+\beta_y c_2)\frac{2(1+h^2)}{(1-c_2)^2}=I_6+(\beta_x c_1+\beta_y c_2)I_7$$

其中：$I_6=\frac{4h^2(1+h^2)}{(1-c_2)^4}$，$I_7=\frac{2(1+h^2)}{(1-c_2)^2}$。

第三部分：

$$Cov(g_2,g_2)=-\lim_{r\to 1_+}\frac{1}{2\pi}\oint_{|\zeta_1|=1}\int_{|\zeta_2|=1}\frac{g_2(x(\zeta_1))g_2(x(\zeta_2))}{(\zeta_1-r\zeta_2)^2}d\zeta_1 d\zeta_2$$

$$-\frac{(\beta_x c_1+\beta_y c_2)(1-c_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{g_2(x(\zeta_1))}{(\zeta_1+\frac{c_2}{h})^2}d\zeta_1\oint_{|\zeta_2|=1}\frac{g_2(x(\zeta_2))}{(\zeta_2+\frac{c_2}{h})^2}d\zeta_2$$

$$=-2\log(1-h^2)+(\beta_x c_1+\beta_y c_2)$$

综合三部分可得：

$$\begin{aligned} Cov(g) &= Cov(g_1, g_1) + Cov(g_2, g_2) - 2Cov(g_1, g_2) \\ &= I_4 - I_6 - 2\log(1+h^2) + (\beta_x c_1 + \beta_y c_2)(I_5 + 1 - I_7). \end{aligned}$$

从而由(3.8)可得

$$tr(\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1})^2 - \log(\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}) - p \cdot F^{c_{n_1,n_2}}(g) \xrightarrow{d} N(\mu(g), Cov(g)).$$

即式(3.6)成立，进而有：

$$Cov(g)^{-1/2}\{T^{New} - p \cdot F^{c_{n_1},c_{n_2}}(g) - \mu(g)\} \xrightarrow{d} N(0,1).$$

即式(3.7)成立，证毕。

至此，针对两个总体协方差矩阵的相等性检验问题，提出的新检验统计量的中心极限定理得到了证明，下一小节针对新统计量的性能进行仿真模拟对比。

3.3 模拟实验

本小节通过模拟试验首先对检验统计量T^{New}的表现进行了评估，之后主要在高维多元正态分布和非正态分布下用T^{X}和T^{New}比较。模拟实验主要是在 R 软件下进行。在本小节中采用的评价标准有如下两个：

(i) AVOD：$d = |\hat{\alpha} - \alpha|$

AVOD 是检验水平$\hat{\alpha}$与给定的显著性水平α差的绝对值：$d = |\hat{\alpha} - \alpha|$。具体到实验中，$d$越小，说明模拟实验越好，所对应的估计方法就越好。

在这里，检验水平$\hat{\alpha}$为：

$$\hat{\alpha} = \frac{(\#\{t_{H_0} \geq z_{1-\alpha}\})}{m}.$$

其中t_{H_0}表示统计量在原假设下模拟数据中的值，$z_{1-\alpha}$表示标准正态分布的上侧α分位数，$\#$表示$t_{H_0} \geq z_{1-\alpha}$的总次数。在我们试验中，选择$m = 1000$且选择显著性水平$\alpha = 0.05$。

(ii) 经验功效$\hat{\beta}$

经验功效$\hat{\beta}$是本次实验采用的另一个评价指标。具体是在备择假设下进行模拟，定义经验功效$\hat{\beta}$为：

$$\hat{\beta} = \frac{(\#\{t_{H_1} \geq \hat{z}_{1-\alpha}\})}{m}.$$

其中t_{H_1}表示统计量在备择假设下模拟数据中的值，$\#$表示$t_{H_1} \geq \hat{z}_{1-\alpha}$的总次数。特别地，$\hat{z}_{1-\alpha}$是经验功效的评判点，是对原假设下进行的$m$次试验，其结果中第$m\alpha$

大的点。

此外，在模拟实验中，参数设置如下：

1. 两个总体协方差矩阵分别为 $\mathbf{\Sigma}_1=\sigma_1^2\mathrm{I}_p$ 和 $\mathbf{\Sigma}_1=\sigma_2^2(\rho^{|i-j|})_{p\times p}$，其中参数 $\sigma_1^2,\sigma_2^2>0$，$\rho\in(-1,1)$；

2. 分别设置 ρ=0 或 0.5，$\sigma_1^2=\sigma_2^2=1,2,4$。且当 ρ=0 时，$\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=\mathrm{I}_p$ 用于计算检验水平；当 ρ=0.5 时，$\mathbf{\Sigma}_1=\mathrm{I}_p$，$\mathbf{\Sigma}_1=(0.5^{|i-j|})_{p\times p}$，即 $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}\neq\mathrm{I}_p$，则可以用于检验功效的计算；

对于每一个参数设置 $(\rho,\sigma_1^2,\sigma_2^2,p,n_1,n_2)$，正态分布和非正态分布设置如下：

(i). 对于正态分布，样本 $\xi_1,\xi_2,\ldots,\xi_{n_1}$ 取自正态总体 $N_p(\mathbf{0},\mathbf{\Sigma}_1)$；样本 $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\ldots,\boldsymbol{\eta}_{n_2}$ 取自正态总体 $N_p(\mathbf{0},\mathbf{\Sigma}_2)$。

(ii). 对于 *Gamma* 分布，样本 $\boldsymbol{\xi}_j=\mathbf{\Sigma}_1^{1/2}\mathbf{w}_j$，其中对于每一个 $j=1,2,\cdots,n_1$，$\mathbf{w}_j$ 独立同分布于 *Ga*(2.5,0.5)；样本 $\boldsymbol{\eta}_k=\mathbf{\Sigma}_2^{1/2}\mathbf{z}_k$，其中对于每一个 $k=1,2,\cdots,n_2$，$\mathbf{z}_k$ 独立同分布于 *Ga*(2.5,0.5)。

(iii). 对于 T 分布，样本 $\boldsymbol{\xi}_j=\mathbf{\Sigma}_1^{1/2}\mathbf{w}_j$，其中对于每一个 $j=1,2,\cdots,n_1$，$\mathbf{w}_j$ 独立同分布于 $t(8)$；样本 $\boldsymbol{\eta}_k=\mathbf{\Sigma}_2^{1/2}\mathbf{z}_k$，其中对于每一个 $k=1,2,\cdots,n_2$，$\mathbf{z}_k$ 独立同分布于 $t(8)$。

表 3.1 和表 3.2 列出了正态分布和 Gamma 分布(非正态分布)下的新检验统计量 T^{New} 的检验水平和检验功效模拟结果。表 3.3-3.5 是在 $\sigma_1^2=\sigma_2^2=1$ 时，分别在正态分布、Gamma 分布以及 T 分布三个不同分布下，从检验水平(AVOD)和检验功效两个方面对 T^{New} 和 T^X 进行了比较，其中 T^X 便是 Xu[19]在 2014 年提出的伪似然比检验。

从表 3.1-3.5 中的模拟结果可以看出：

1.从表 3.1 和表 3.2 具体可以看出，随着维数 p 和样本量 (n_1,n_2) 逐渐变大，T^{New} 的检验水平也存在越来越接近于给定的标准水平 5%的趋势。从另外一个评价标准 AVOD 来说，首先在正态分布下的 d，其最大值为 0.0089，最小值仅为 0.0002，其次在 Gamma 分布下，d 的最大值为 0.0101，最小值为 0.0001，说明模拟实验良好，对应的估计方法 T^{New} 效果优。另外，对于经验功效来看，从表 3.1 和表 3.2 可以看出，同样随着维数 p 与样本大小 (n_1,n_2) 成比例地增加，无论在正态分布下，还是 Gamma 分布(非正态分布)下，T^{New} 的检验功效在大体上表现都很稳健。

表 3.1 不同参数设置$(\rho,\sigma_1^2,\sigma_2^2,p,n_1,n_2)$正态分布的检验水平（AVOD）和检验功效(百分比)

(σ_1^2,σ_2^2)	检验水平(AVOD)			检验功效		
	(2,4)	(4,4)	(4,2)	(2,4)	(4,4)	(4,2)
p	$p/n_1, p/n_2 = 0.2$					
40	0.0542(0.0042)	0.0535(0.0035)	0.0541(0.0041)	1	1	1
160	0.0522(0.0022)	0.0530(0.0030)	0.0498(0.0002)	1	1	1
320	0.0589(0.0089)	0.0523(0.0023)	0.0533(0.0033)	1	1	1
640	0.0551(0.0051)	0.0497(0.0003)	0.0521(0.0021)	1	1	1
p	$p/n_1, p/n_2 = 0.5$					
40	0.0514(0.0014)	0.0552(0.0052)	0.0557(0.0057)	0.9979	0.9999	0.9997
160	0.0509(0.0009)	0.0455(0.0045)	0.0568(0.0068)	1	1	1
320	0.0510(0.0010)	0.0534(0.0034)	0.0489(0.0012)	1	1	1
640	0.0506(0.0006)	0.0516(0.0016)	0.0509(0.0009)	1	1	1
p	$p/n_1, p/n_2 = 0.8$					
40	0.0582(0.0082)	0.0572(0.0072)	0.0554(0.0054)	0.9915	0.9604	0.9843
160	0.0449(0.0051)	0.0564(0.0064)	0.0532(0.0032)	1	1	1
320	0.0524(0.0024)	0.0513(0.0013)	0.0512(0.0013)	1	1	1
640	0.0504(0.0004)	0.0504(0.0004)	0.0498(0.0002)	1	1	1

表 3.2 不同参数设置$(\rho,\sigma_1^2,\sigma_2^2,p,n_1,n_2)$Gamma 分布的检验水平（AVOD）和检验功效(百分比)

(σ_1^2,σ_2^2)	检验水平(AVOD)；			检验功效		
	(2,4)	(4,4)	(4,2)	(2,4)	(4,4)	(4,2)
p	$p/n_1, p/n_2 = 0.2$					
40	0.0601(0.0101)	0.0584(0.0084)	0.0586(0.0086)	1	1	1
160	0.0585(0.0085)	0.0553(0.0053)	0.0546(0.0046)	1	1	1
320	0.0551(0.0051)	0.0534(0.0034)	0.0524(0.0024)	1	1	1
640	0.0503(0.0003)	0.0501(0.0001)	0.0466(0.0036)	1	1	1
p	$p/n_1, p/n_2 = 0.5$					
40	0.0593(0.0093)	0.0609(0.0109)	0.0569(0.0069)	1	0.9997	1
160	0.0543(0.0043)	0.0578(0.0078)	0.0532(0.0032)	1	1	1
320	0.0544(0.0044)	0.0568(0.0068)	0.0524(0.0024)	1	1	1
640	0.0506(0.0006)	0.0535(0.0035)	0.0522(0.0022)	1	1	1
p	$p/n_1, p/n_2 = 0.8$					
40	0.0585(0.0085)	0.0570(0.0070)	0.0588(0.0088)	0.9999	1	1
160	0.0531(0.0031)	0.0539(0.0039)	0.0573(0.0073)	1	1	1
320	0.0526(0.0026)	0.0531(0.0031)	0.0528(0.0028)	1	1	1
640	0.0496(0.0096)	0.0459(0.0041)	0.0486(0.0014)	1	1	1

2.当e=1时，比例检验问题等价于两个总体协方差矩阵的等式检验问题。通过表 3.3-3.5 可以看出，在正态分布下T^{New}和T^{X}都有相似的检验水平趋势，从检验水平和标准水平差的绝对值 AVOD 来看，T^{New}的d值更小，更接近于 5%的标

准水平，实验结果更好，检验效果更优。另外在 *Gamma* 分布和 T 分布这种非正态分布下，实验结果与正态分布相似，都表明T^{New}的总体性能比T^{X}要更好。

表 3.3 正态分布下T^{X}和T^{New}的检验水平（AVOD）和检验功效(百分比)

p	检验水平(AVOD)		检验功效	
	T^{X}	T^{New}	T^{X}	T^{New}
		$p/n_1, p/n_2 = 0.2$		
40	0.0549(0.0049)	0.0513(0.0013)	1	1
160	0.0560(0.0060)	0.0487(0.0013)	1	1
320	0.0554(0.0054)	0.0483(0.0017)	1	1
640	0.0532(0.0053)	0.0496(0.0004)	1	1
		$p/n_1, p/n_2 = 0.5$		
40	0.0547(0.0047)	0.0536(0.0036)	0.6274	0.9969
160	0.0538(0.0038)	0.0530(0.0030)	1	1
320	0.0555(0.0055)	0.0526(0.0026)	1	1
640	0.0536(0.0036)	0.0512(0.0012)	1	1
		$p/n_1, p/n_2 = 0.8$		
40	0.0616(0.0116)	0.0580(0.0080)	0.5921	0.9914
160	0.0598(0.0098)	0.0568(0.0068)	0.9978	0.9999
320	0.0576(0.0076)	0.0546(0.0046)	1	1
640	0.0549(0.0049)	0.0492(0.0008)	1	1

表 3.4 T 分布下T^{X}和T^{New}的检验水平（AVOD）和检验功效(百分比)

p	检验水平(AVOD)		检验功效	
	T_X	T_{New}	T_X	T_{New}
		$p/n_1, p/n_2 = 0.2$		
40	0.0558(0.0058)	0.0523(0.0023)	1	1
160	0.0588(0.0088)	0.0483(0.0017)	1	1
320	0.0567(0.0067)	0.0486(0.0014)	1	1
640	0.0522(0.0022)	0.0497(0.0003)	1	1
		$p/n_1, p/n_2 = 0.5$		
40	0.0554(0.0054)	0.0531(0.0031)	0.7257	0.9824
160	0.0551(0.0051)	0.0521(0.0021)	0.9886	1
320	0.0543(0.0043)	0.0547(0.0047)	1	1
640	0.0481(0.0019)	0.0505(0.0005)	1	1
		$p/n_1, p/n_2 = 0.8$		
40	0.0612(0.0112)	0.0596(0.0096)	0.7408	0.8958
160	0.0567(0.0067)	0.0558(0.0058)	0.9999	0.9998
320	0.0558(0.0058)	0.0555(0.0055)	1	1
640	0.0538(0.0038)	0.0524(0.0024)	1	1

3.从检验功效这个方面，可以从表 3.3-3.5 看出，T^{New}检验功效的值绝大部分要比T^{X}检验功效的值要大一些，因为T^{X}是针对任何未知的比例常数c而提出

的，所以很明显，当$c=1$时，T^{New}比T^{X}具有更好的检验效果。这是因为T^{X}首先必须牺牲一些样本来取消未知的c，然后使用剩余的样本来检验。总的来说，T^{New}比T^{X}的表现更佳，更稳健。在 p=40 的情况下，T^{New}的这种优势更为明显。在表 4 可以看出，对于 T 分布，在 p=160 和 p/n_1=$p/n_2=0.8$的情况下，T^{New}检验功效的值要比T^{X}检验功效的值要小，但也仅仅只是 0.001 的差距，并不能否定T^{New}在检验功效方面的优势。

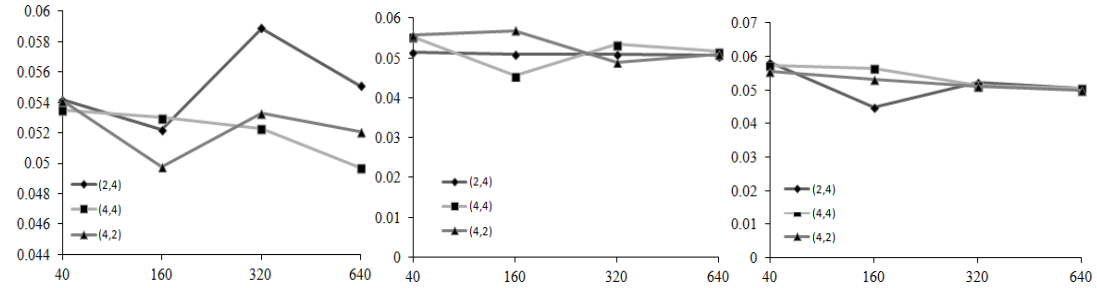
表 3.5 Gamma 分布下T^{X}和T^{New}的检验水平（AVOD）和检验功效(百分比)

p	检验水平(AVOD)		检验功效	
	T_X	T_{New}	T_X	T_{New}
		$p/n_1, p/n_2=0.2$		
40	0.0595(0.0095)	0.0544(0.0044)	1	1
160	0.0588(0.0088)	0.0530(0.0030)	1	1
320	0.0735(0.0235)	0.0527(0.0027)	1	1
640	0.0660(0.0160)	0.0495(0.0005)	1	1
		$p/n_1, p/n_2=0.5$		
40	0.0433(0.0067)	0.0612(0.0112)	0.8501	0.9995
160	0.0732(0.0232)	0.0592(0.0092)	0.9942	1
320	0.0575(0.0075)	0.0529(0.0029)	1	1
640	0.0535(0.0035)	0.0501(0.0001)	1	1
		$p/n_1, p/n_2=0.8$		
40	0.0776(0.0276)	0.0719(0.0219)	0.8992	0.9028
160	0.0668(0.0168)	0.0560(0.0060)	0.9841	0.9998
320	0.0568(0.0068)	0.0537(0.0037)	1	1
640	0.0559(0.0059)	0.0548(0.0048)	1	1

为了有一个更直观的理解，为了使模拟实验结果更直观，首先对部分实验结果(检验水平)用折线图描绘出来。图 3.1 和图 3.2 分别描绘了正态分布和 Gamma 分布下的新检验统计量T^{New}的检验水平模拟结果，图 3.3-3.5 是在$\sigma_1^2=\sigma_2^2=1$时，分别在正态分布、T 分布以及 Gamma 分布下，从检验水平对T^{New}和T^{X}进行了更直观的比较。

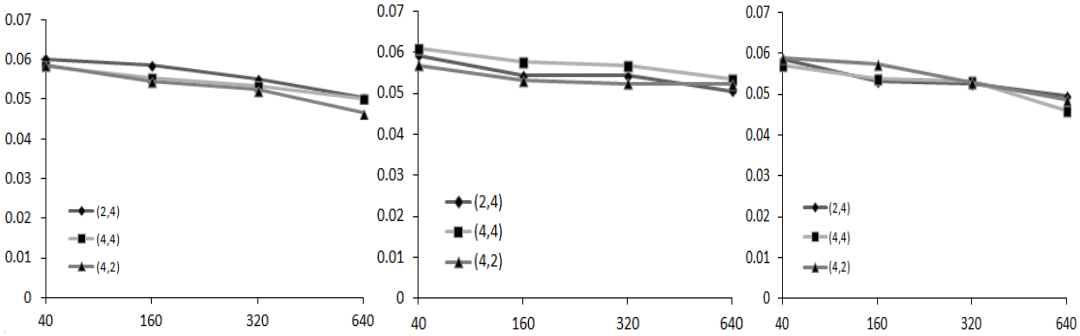
(1) 由图 3.1 和图 3.2 可见，随着维数 p 和样本量(n_1,n_2)逐渐变大，(σ_1^2,σ_2^2)在三种不同取值：(2,4)、(4,4)以及(4,2)下，T^{New}的检验水平基本在标准水平 5%上下波动。例如在正态分布下，当 p/n_1，$p/n_2=0.5$，同时(σ_1^2,σ_2^2)分别取(4,4)和(4,2)时，折线图只有在维数 p 等于 160 的时候波动较大，但整体的结果仍然在5%的周围；其次在(σ_1^2,σ_2^2)取(2,4)时，折线图中检验水平的值都很接近 0.05。另外，在 $p/n_1, p/n_2=0.8$情况下，(σ_1^2,σ_2^2)在三种不同的取值下，折线图中所表示的检验水平都在标准水平 5%周围浮动，并没有太大距离。总之，从图 3.1 和图

3.2 的折线图中可以看出T^{New}的检验水平表现很稳健。



a) $p/n_1, p/n_2 = 0.2$　　b) $p/n_1, p/n_2 = 0.5$　　c) $p/n_1, p/n_2 = 0.8$

图 3.1 不同参数设置$(\rho, \sigma_1^2, \sigma_2^2, p, n_1, n_2)$正态分布下$T^{New}$的检验水平(百分比)



a) $p/n_1, p/n_2 = 0.2$　　b) $p/n_1, p/n_2 = 0.5$　　c) $p/n_1, p/n_2 = 0.8$

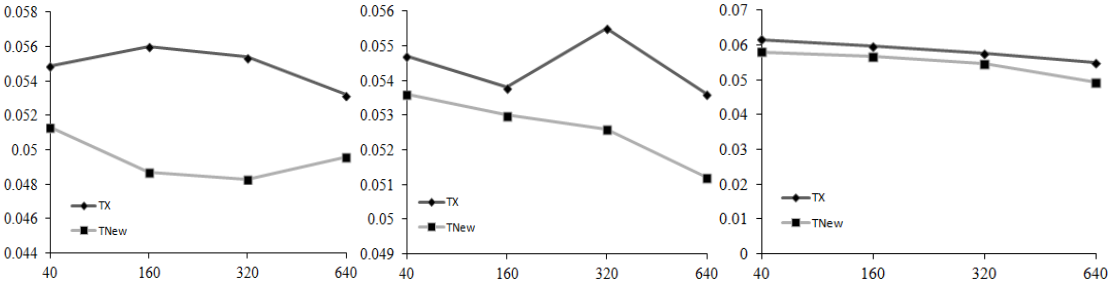
图 3.2 不同参数设置$(\rho, \sigma_1^2, \sigma_2^2, p, n_1, n_2)$Gamma 下$T^{New}$的检验水平(百分比)

(2) 图 3.3-3.5 分别描绘了在σ_1^2=σ_2^2=1时，正态分布、T 分布以及 Gamma 分布下T^{New}和T^X的检验水平趋势。

在正态分布下，可以从图 3.3 中看出，当$p/n_1, p/n_2 = 0.2$时，T^{New}和T^X的检验水平有不同的趋势，但并不影响T^{New}随着样本量(n_1, n_2)逐渐变大，与标准水平 5%越来越接近的趋势。虽然T^X有相似趋势，但与 5%的接近程度比T^{New}要差一些。而当$p/n_1, p/n_2 = 0.5$和$p/n_1, p/n_2 = 0.8$时，T^{New}和T^X的检验水平均呈现下降趋势，在相同的样本量(n_1, n_2)下，T^{New}检验水平的值均小于T^X的检验水平值，且T^{New}的检验水平更接近于标准水平 5%，即T^{New}的检验水平值更接近标准水平 5%。总之，图 3.3 的三个折线图均可表明：在正态分布下，T^{New}的检验水平比T^X表现更好。

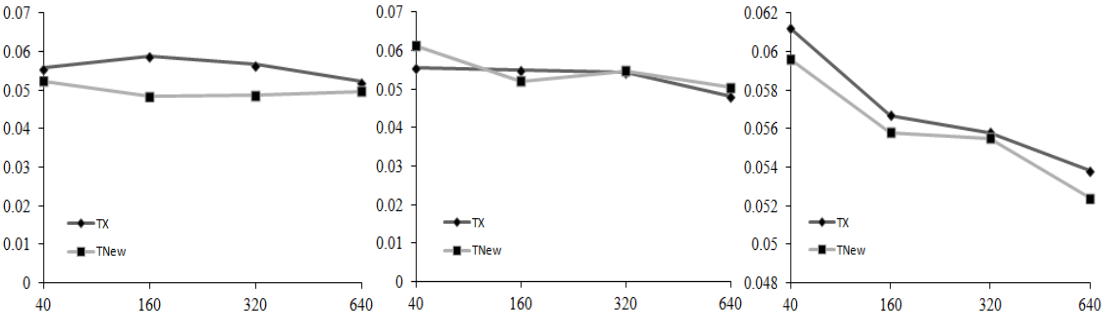
在 T 分布下，在$p/n_1, p/n_2 = 0.2$和$p/n_1, p/n_2 = 0.8$两个折线图中，T^{New}和T^X的检验水平趋势都与正态分布下的结果很相似。但在$p/n_1, p/n_2 = 0.5$，样本量为 40 时，T^{New}与 0.05 的距离要比T^X与 0.05 的距离大些，但并不影响在样本量为 160、320 以及 640 的情况下，T^{New}的整体优势。总之，图 3.4 的三个折线

图均可表明：在 T 分布下，T^{New} 的检验水平比 T^{X} 表现更优。在 Gamma 分布下，当 $p/n_1, p/n_2 = 0.2$ 时，从折线图来看，T^{New} 和 T^{X} 的检验水平有着相似的整体趋势，且相比于 T^{X}，随样本量 (n_1, n_2) 逐渐变大，T^{New} 的检验水平与标准水平 5% 越来越接近的程度更高；当 $p/n_1, p/n_2 = 0.5$ 时，T^{New} 的检验水平呈现出下降的趋势，但 T^{X} 有些波动，在样本量等于 40 时，检验水平的值要小于 T^{New}，但在样本量分别为 160，320，640 时，T^{New} 检验水平的值都小于 T^{X} 检验水平的值，即 T^{New} 的检验水平更接近于标准水平 5%；当 $p/n_1, p/n_2 = 0.8$ 时，T^{New} 和 T^{X} 的检验水平跟相同情况下正态分布情形下的结果类似。但是，对于每一样本量 (n_1, n_2)，T^{New} 的检验水平值都要比 T^{X} 的检验水平值小，且 T^{New} 的检验水平值更接近标准水平 5%。总之，图 3.5 的三个折线图均表明：在 Gamma 分布下，与 T^{X} 相比，T^{New} 的检验水平表现更稳健。



a) $p/n_1, p/n_2 = 0.2$　　b) $p/n_1, p/n_2 = 0.5$　　c) $p/n_1, p/n_2 = 0.8$

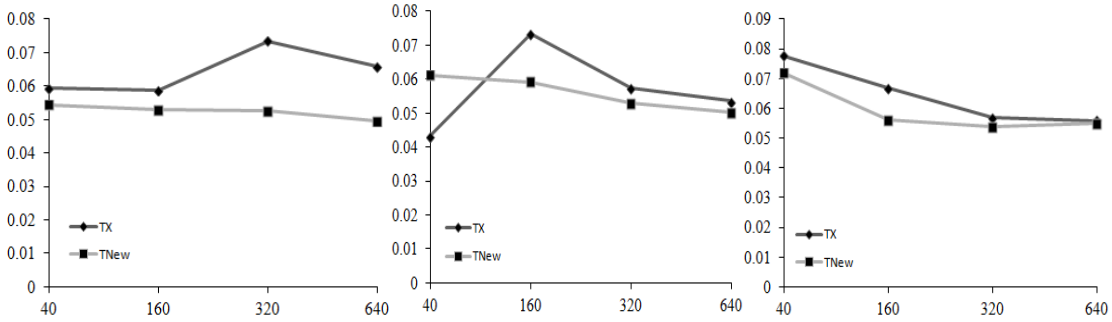
图 3.3 正态分布下 T^{X} 和 T^{New} 的检验水平(百分比)



a) $p/n_1, p/n_2 = 0.2$　　b) $p/n_1, p/n_2 = 0.5$　　c) $p/n_1, p/n_2 = 0.8$

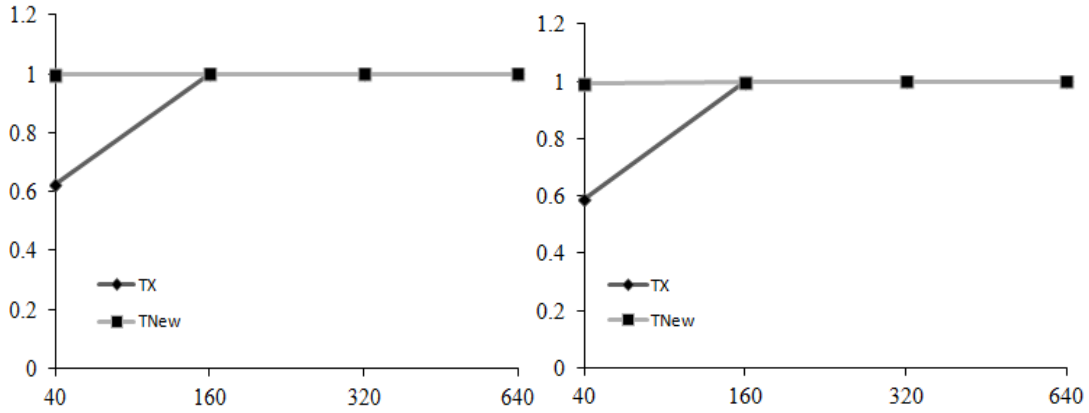
图 3.4 T 分布下 T^{X} 和 T^{New} 的检验水平(百分比)

另外，对部分检验功效的实验结果也用折线图进行了描绘。在 $\sigma_1^2=\sigma_2^2=1$ 时，图 3.6-3.8 分别更加直观地展示了 T^{New} 和 T^{X} 在正态分布、T 分布和 Gamma 分布下的区别。



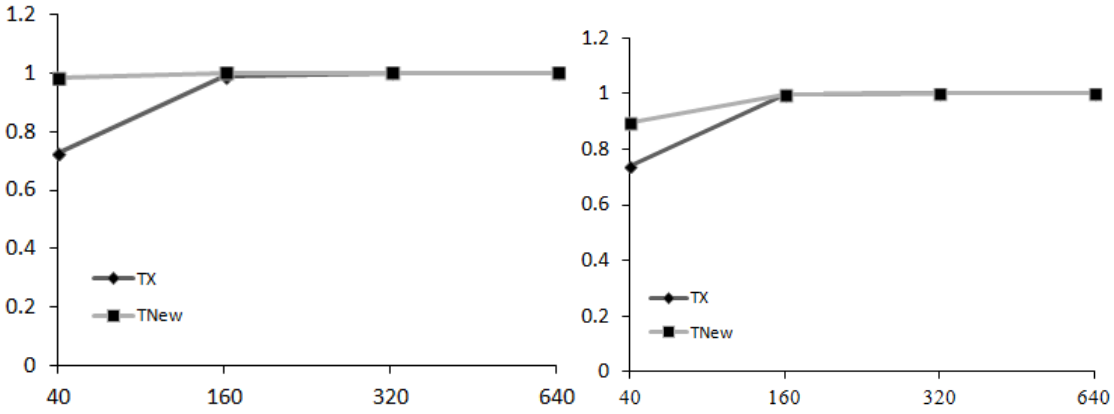
a) $p/n_1, p/n_2 = 0.2$　　b) $p/n_1, p/n_2 = 0.5$　　c) $p/n_1, p/n_2 = 0.8$

图 3.5 Gamma 分布下 T^X 和 T^{New} 的检验水平(百分比)



a) $p/n_1, p/n_2 = 0.5$　　b) $p/n_1, p/n_2 = 0.8$

图 3.6 正态分布下 T^X 和 T^{New} 的检验功效(百分比)



a) $p/n_1, p/n_2 = 0.5$　　b) $p/n_1, p/n_2 = 0.8$

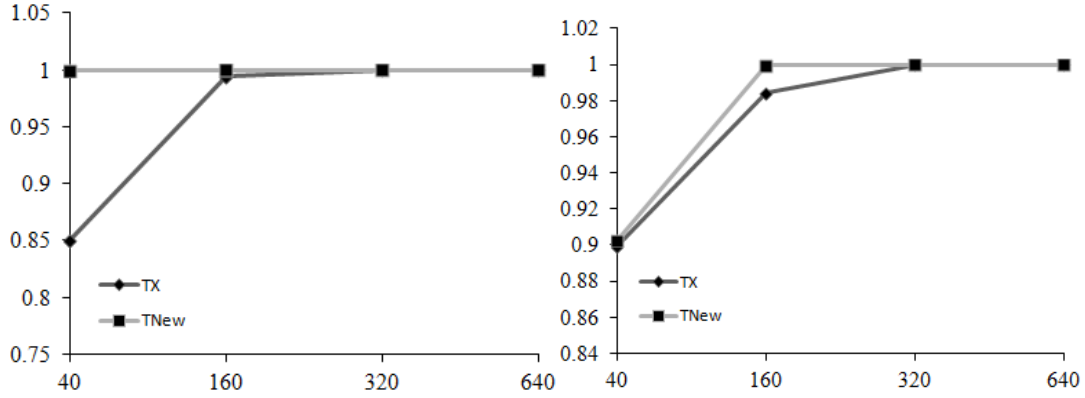
图 3.7 T 分布下 T^X 和 T^{New} 的检验功效(百分比)

(1) 图 3.6 表明，在正态分布下，对于 $p/n_1, p/n_2 = 0.5$ 和 $p/n_1, p/n_2 = 0.8$ 时的 T^{New} 检验功效比 T^X 检验功效更快接近于 1，即：T^{New} 的检验功效表现更稳健。

(2) 图 3.7 表明，在 T 分布下，折线图描绘出的结果同正态分布相似，对于 $p/n_1, p/n_2 = 0.5$ 和 $p/n_1, p/n_2 = 0.8$ 时的 T^{New} 检验功效比 T^X 检验功效也更快地

接近于 1，即：T^{New} 的检验功效表现更稳健。

(3) 图 3.8 显示，在 Gamma 分布下，对于 $p/n_1, p/n_2 = 0.5$ 和 $p/n_1, p/n_2 = 0.8$ 时的 T^{New} 检验功效比 T^{X} 检验功效表现更优。



a) $p/n_1, p/n_2 = 0.5$　　b) $p/n_1, p/n_2 = 0.8$

图 3.8 Gamma 分布下 T^{X} 和 T^{New} 的检验功效(百分比)

3.4 小结

本章针对高维数据情况下的双样本协方差矩阵相等性检验问题，提出了一种新的检验方法，并证明了其渐近正态性。新的检验统计量解决了 Xu[19]检验方法中对于比例参数的限制，并参考 Jiang[46]在更宽泛的条件下的中心极限定理，使得新检验统计量有广泛的应用。另外，通过模拟实验结果表明，新检验统计量 T^{New} 在检验两个高维总体协方差矩阵相等性问题时更具有稳健性，并且在正态分布和非正态分布的情形下都有不错的结果。

第四章 不同协方差下高维数据的 MANOVA 检验问题

近年来，随着数据收集技术的不断发展，高维数据在各个领域的应用异军突起。经典的统计理论是在维数固定而样本趋于无穷的假定下推导出来的，而高维数据的特点之一是有时维数远远大于样本容量，这与传统统计理论的假定相违背，因此经典多元统计理论不能直接应用到高维数据上。具体来说，高维数据就是“大 p 小 n”现象，其中 $p/n\to\infty$。在实际中，仅在维数变高(大 p)这一方面，就需要修正大量传统的多变量分析理论方法和应用程序。因此，在“大 p 小 n”背景下，经典的统计方法与理论很难应用到高维数据问题中。比如，Hotelling-T^2 检验[2]在高维条件下不再具有强大的效能。因此，对基于高维数据的假设检验问题，寻找优良的检验方法是必不可少的。

4.1 总体均值假设检验问题介绍

本章节就高维数据下总体均值假设检验问题寻求更好的检验统计量，做了进一步的工作。下面首先介绍高维数据下多样本的总体均值假设检验问题：

假设 k 个独立同分布的样本 $\mathbf{y}_{\alpha 1},\mathbf{y}_{\alpha 2},\cdots,\mathbf{y}_{\alpha n_\alpha}(\alpha=1,2,\cdots,k)$ 满足：$\mathrm{E}(\mathbf{y}_{\alpha 1})=\boldsymbol{\mu}_\alpha$，$Cov(\mathbf{y}_{\alpha 1})=\boldsymbol{\Sigma}_\alpha$。其中数据维度 p 远大于总体样本数量 $n=\sum_{\alpha=1}^{k}n_\alpha$。在此基础上，考虑以下假设：

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_k \quad VS \quad H_1:H_0\text{不成立} \tag{4.1}$$

该假设问题也称为多变量 k 样本 Behrens-Fisher 问题或 MANOVA 检验问题。近十几年来，基于上述假设检验问题，许多学者进行了大量的研究，其中 Zhang 等[49]及 Xu 等[50]分别在 2009 年和 2015 年都给出了 k-样本 Behrens-Fisher 问题的近似解；Srivastava 等[51]于 2006 年在多元线性回归模型下提出了一个标度不变量的 GLHT 检验方法；2012 年，Yamada 等[52]考虑了在高维线性回归中的 GLHT 假设检验问题；Srivastava 等[53]在 2013 年针对非正态数据推广了 Yamada 的检验方法；2015 年，Yamada 等[54]提出了一种不需要假设共同协方差阵的 MANOVA 检验方法。但当样本协方差不同时，应用上述方法可能会出现错误的结果。为了解决这个问题，曹明响等[55]于 2015 年在 Chen 的思想上进行扩展从而提出了一个新的检验统计量；2017 年，Zhou 等[56]基于 C-Q[57]和 Feng 等[58]的理论并利用 U 统计量[59]也得到一个新检验统计量。

4.2 新检验统计量及其渐近分布

4.2.1 构建检验统计量

本章节构造的新检验统计量是基于 Chen 等[60]的思想提出来的，实现了数据从两样本到多样本的提升。记

$$T_n=\frac{k-1}{nk}\sum_{\alpha=1}^{k}\frac{\sum_{i\neq j}\mathbf{y}_{\alpha i}^T\mathbf{y}_{\alpha j}}{n_\alpha(n_\alpha-1)}-\frac{1}{nk}\sum_{\alpha\neq\beta}\frac{\sum_{i,j}\mathbf{y}_{\alpha i}^T\mathbf{y}_{\beta j}}{n_\alpha n_\beta}.$$

为了简化T_n，令$\mathbf{G}=(g_{\alpha\beta})_{k\times k}$，其中$g_{\alpha\alpha}=\frac{k-1}{nk},g_{\alpha\beta}=\frac{-1}{nk}(\alpha\neq\beta)$，故可以重新写为：

$$T_n=\sum_{\alpha=1}^{k}g_{\alpha\alpha}\frac{\sum_{i\neq j}\mathbf{y}_{\alpha i}^T\mathbf{y}_{\alpha j}}{n_\alpha(n_\alpha-1)}+\sum_{\alpha\neq\beta}g_{\alpha\beta}\frac{\sum_{i,j}\mathbf{y}_{\alpha i}^T\mathbf{y}_{\beta j}}{n_\alpha n_\beta}. \tag{4.2}$$

T_n 的期望为 $E(T_n)=\sum_{\alpha=1}^{k}(\boldsymbol{\mu}_\alpha-\bar{\boldsymbol{\mu}})^T(\boldsymbol{\mu}_\alpha-\bar{\boldsymbol{\mu}})/n=\|\mathbf{G}\boldsymbol{\mu}\|^2$，其中 $\boldsymbol{\mu}=(\boldsymbol{\mu}_1^T,\boldsymbol{\mu}_2^T,\cdots,\boldsymbol{\mu}_k^T)^T$，$\bar{\boldsymbol{\mu}}=\frac{1}{k}\sum_{\alpha=1}^{k}\boldsymbol{\mu}_\alpha$。

本章节基于T_n提出了一个新的检验统计量[61]：

$$T_n^*=\frac{T_n}{\sqrt{\hat{\sigma}^2(T_n)}}. \tag{4.3}$$

其中，

$$\hat{\sigma}^2(T_n)=2\Big(\sum_{\alpha=1}^{k}\frac{(k-1)^2\hat{tr}(\boldsymbol{\Sigma}_\alpha^2)}{n_\alpha(n_\alpha-1)}+\sum_{\alpha\neq\beta}\frac{\hat{tr}(\boldsymbol{\Sigma}_\alpha\boldsymbol{\Sigma}_\beta)}{n_\alpha n_\beta}\Big)/n^2k^2,$$

$$\hat{tr}(\boldsymbol{\Sigma}_\alpha^2)=\frac{(n_\alpha-1)^2}{(n_\alpha+1)(n_\alpha-2)}\Big(tr(\hat{\boldsymbol{\Sigma}}_\alpha^2)+\frac{1}{n_\alpha-1}tr^2(\hat{\boldsymbol{\Sigma}}_\alpha)\Big),$$

$$\hat{tr}(\boldsymbol{\Sigma}_\alpha\boldsymbol{\Sigma}_\beta)=tr\Big(\frac{1}{(n_\alpha-1)(n_\beta-1)}\sum_{i=1}^{n_\alpha}(\mathbf{y}_{\alpha i}-\bar{\mathbf{y}}_\alpha)(\mathbf{y}_{\alpha i}-\bar{\mathbf{y}}_\alpha)^T\sum_{j=1}^{n_\beta}(\mathbf{y}_{\beta i}-\bar{\mathbf{y}}_\beta)(\mathbf{y}_{\beta i}-\bar{\mathbf{y}}_\beta)^T\Big).$$

而估计$\hat{tr}(\boldsymbol{\Sigma}_\alpha^2)$是 Bai 等[62,63,64]提出的$tr(\boldsymbol{\Sigma}_\alpha^2)$的无偏估计，$\hat{tr}(\boldsymbol{\Sigma}_\alpha\boldsymbol{\Sigma}_\beta)$也是$tr(\boldsymbol{\Sigma}_\alpha\boldsymbol{\Sigma}_\beta)$的无偏估计。

具体来说，式(4.2)经过线性运算，可以得到(4.4)和(4.5)两种不同的T_n表达式：

$$\frac{\sum_{i\neq j}\mathbf{y}_{\alpha i}^{T}\mathbf{y}_{\alpha j}}{n_\alpha(n_\alpha-1)}=\overline{\mathbf{y}}_\alpha^{T}\overline{\mathbf{y}}_\alpha-\frac{tr(\hat{\boldsymbol{\Sigma}}_\alpha)}{n_\alpha},\quad \frac{\sum_{i,j}\mathbf{y}_{\alpha i}^{T}\mathbf{y}_{\beta j}}{n_\alpha n_\beta}=\overline{\mathbf{y}}_\alpha^{T}\overline{\mathbf{y}}_\beta.$$

式中，$\overline{\mathbf{y}}_\alpha=n_\alpha^{-1}\sum_{i=1}^{n_\alpha}\mathbf{y}_{\alpha i}$, $\hat{\boldsymbol{\Sigma}}_\alpha=(n_\alpha-1)^{-1}\sum_{i=1}^{n_\alpha}(\mathbf{y}_{\alpha i}-\overline{\mathbf{y}}_\alpha)(\mathbf{y}_{\alpha i}-\overline{\mathbf{y}}_\alpha)^{T}$.

$$T_n=\sum_{\alpha,\beta}g_{\alpha\beta}\overline{\mathbf{y}}_\alpha^{T}\overline{\mathbf{y}}_\beta-\sum_{\alpha=1}^{k}\frac{g_{\alpha\alpha}}{n_\alpha}tr(\hat{\boldsymbol{\Sigma}}_\alpha)=\left\|\mathbf{G}\hat{\boldsymbol{\mu}}\right\|^2-\sum_{\alpha=1}^{k}\frac{g_{\alpha\alpha}}{n_\alpha}tr(\hat{\boldsymbol{\Sigma}}_\alpha).\tag{4.4}$$

其中 $\hat{\boldsymbol{\mu}}=(\overline{\mathbf{y}}_1^{T},\overline{\mathbf{y}}_2^{T},\cdots,\overline{\mathbf{y}}_k^{T})^{T}$ 是 $\boldsymbol{\mu}$ 的无偏估计。

另外，为了更好地解释 T_n 的方差，令 $\mathbf{x}_{\alpha i}=\mathbf{y}_{\alpha i}-\boldsymbol{\mu}_\alpha$, $i=1,2,\cdots,n_\alpha$, $\alpha=1,2,\cdots,k$. 对于每一个 $\alpha=1,2,\cdots,k$， $\mathbf{x}_{\alpha 1},\mathbf{x}_{\alpha 2},\cdots,\mathbf{x}_{\alpha n_\alpha}$ 是独立同分布的，并且满足 $\mathrm{E}(\mathbf{x}_{\alpha 1})=\mathbf{0}$，$Cov(\mathbf{x}_{\alpha 1})=\boldsymbol{\Sigma}_\alpha$。故 T_n 又可写成下式：

$$T_n=\widehat{T}_n+2S_n+\left\|\mathbf{G}\boldsymbol{\mu}\right\|^2.\tag{4.5}$$

式中，

$$S_n=(\mathbf{G}\boldsymbol{\mu})^{T}\mathbf{G}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}).$$

$$\widehat{T}_n=\left\|\mathbf{G}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right\|^2-\sum_{\alpha=1}^{k}\frac{g_{\alpha\alpha}}{n_\alpha}tr(\hat{\boldsymbol{\Sigma}}_\alpha)=\sum_{\alpha,\beta}g_{\alpha\beta}\overline{\mathbf{x}}_\alpha^{T}\overline{\mathbf{x}}_\beta-\sum_{\alpha=1}^{k}\frac{g_{\alpha\alpha}}{n_\alpha}tr(\hat{\boldsymbol{\Sigma}}_\alpha)$$

$$=\sum_{\alpha=1}^{k}g_{\alpha\alpha}\frac{\sum_{i\neq j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\alpha j}}{n_\alpha(n_\alpha-1)}+\sum_{\alpha\neq\beta}g_{\alpha\beta}\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\beta j}}{n_\alpha n_\beta}.$$

从(4.5)中可以看出，T_n 和 $\widehat{T}_n$ 在原假设下具有相同的分布。假设 α，β 和 γ 为不同的整数，则有以下式子成立：

$$\mathrm{E}\Big(\frac{\sum_{i\neq j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\alpha j}}{n_\alpha(n_\alpha-1)}\Big)=\mathrm{E}\Big(\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\beta j}}{n_\alpha n_\beta}\Big)=0,$$

$$Var(\frac{\sum_{i\neq j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\alpha j}}{n_\alpha(n_\alpha-1)})=\frac{2tr(\boldsymbol{\Sigma}_\alpha^2)}{n_\alpha(n_\alpha-1)},Var\Big(\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\beta j}}{n_\alpha n_\beta}\Big)=\frac{tr(\boldsymbol{\Sigma}_\alpha\boldsymbol{\Sigma}_\beta)}{n_\alpha n_\beta},$$

$$Cov(\frac{\sum_{i\neq j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\alpha j}}{n_\alpha(n_\alpha-1)},\frac{\sum_{i\neq j}\mathbf{x}_{\beta i}^{T}\mathbf{x}_{\beta j}}{n_\alpha(n_\alpha-1)})=Cov(\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\beta j}}{n_\alpha n_\beta},\frac{\sum_{i\neq j}\mathbf{x}_{\beta i}^{T}\mathbf{x}_{\beta j}}{n_\alpha(n_\alpha-1)})=Cov(\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\beta j}}{n_\alpha n_\beta},\frac{\sum_{i,j}\mathbf{x}_{\alpha i}^{T}\mathbf{x}_{\gamma j}}{n_\alpha n_\gamma})=0.$$

由此得出了在原假设下 T_n 的方差：

$$Var(T_n)=Var(\widehat{T}_n)=\sigma^2(\widehat{T}_n)=2\Big(\sum_{\alpha=1}^{k}\frac{g_{\alpha\alpha}^2 tr(\mathbf{\Sigma}_\alpha^2)}{n_\alpha(n_\alpha-1)}+\sum_{\alpha\neq\beta}\frac{g_{\alpha\beta}^2 tr(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta)}{n_\alpha n_\beta}\Big)$$

$$=2\Big(\sum_{\alpha=1}^{k}\frac{(k-1)^2 tr(\mathbf{\Sigma}_\alpha^2)}{n_\alpha(n_\alpha-1)}+\sum_{\alpha\neq\beta}\frac{tr(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta)}{n_\alpha n_\beta}\Big)/n^2k^2.$$

4.2.2 新检验统计量的渐近性

在一定条件下，本文提出的T_n^*是渐近分布的。在本小节就新检验统计量T_n^*的渐近正态性进行了具体论证。假设存在 k 个独立的样本满足以下广义多元相依模型[65]：

$$\mathbf{y}_{\alpha i}=\boldsymbol{\mu}_\alpha+\mathbf{\Lambda}_\alpha\mathbf{z}_{\alpha i},\quad i=1,2,\cdots,n_\alpha,\quad \alpha=1,2,\cdots,k. \tag{4.6}$$

其中，$\mathbf{\Lambda}$ 是 $p\times p$ 阶的常数矩阵且满足 $\mathbf{\Lambda}_\alpha\mathbf{\Lambda}_\alpha^T=\mathbf{\Sigma}_\alpha$，$\mathbf{z}_{\alpha i}$ 是独立同分布的 p×1 阶随机向量($i=1,2,\cdots,n_\alpha,\alpha=1,2,\cdots,k$,p 是正整数)，且对 $\alpha=1,2,\cdots,k$ 满足 $\mathrm{E}(\mathbf{z}_{\alpha 1})=\mathbf{0}$，$Cov(\mathbf{z}_{\alpha 1})=\mathbf{I}_p$。经过中心化的样本模型可以写成：

$$\mathbf{x}_{\alpha i}=\mathbf{\Lambda}_\alpha\mathbf{z}_{\alpha i},\quad i=1,2\cdots,n_\alpha,\quad \alpha=1,2,\cdots,k.$$

$$\mathbf{x}_{\alpha i}=\mathbf{\Lambda}_\alpha\mathbf{z}_{\alpha i},\quad i=1,2\cdots,n_\alpha,\quad \alpha=1,2,\cdots,k.$$

为了说明统计量T_n^*的渐近性，本文对总体样本做了下列假设：

I. $\lim\limits_{n\to\infty}\frac{n_\alpha}{n}=\tau_\alpha\in(0,1)$，$\alpha=1,2,\cdots,k$；

II. $\mathrm{E}(z_{\alpha il}^4)=3+\gamma<\infty$，$l=1,2,\cdots,m.$其中 $z_{\alpha il}$ 是 $\mathbf{z}_{\alpha i}$ 的第 l 个元素；

III. $\lim\limits_{n\to\infty}\frac{tr(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\gamma\mathbf{\Sigma}_\beta)}{tr(\mathbf{\Sigma}_\alpha\mathbf{\Sigma}_\beta)tr(\mathbf{\Sigma}_\beta\mathbf{\Sigma}_\gamma)}=0$，$\alpha,\beta,\gamma=1,2,\cdots,k$；

IV. $(\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}_\gamma)^T\mathbf{\Sigma}(\boldsymbol{\mu}_\alpha-\boldsymbol{\mu}_\gamma)=o(\frac{tr(\mathbf{\Sigma}^2)}{n})$，$\alpha,\gamma=1,2,\cdots,k.$

其中假设条件 IV 是高维检验问题中经常会用到的一个表达式，它可以衡量不同总体间均值的差异性。

在这里，主要是通过$\widehat{T}_n$的渐近性来说明T_n^*的渐近性。详细地，用一个鞅差序列把$\widehat{T}_n$表示出来，令$\mathbf{y}_{n'_{\alpha-1}+i}=\mathbf{x}_{\alpha i}$，$n'_{\alpha-1}=n_1+n_2+\cdots+n_\alpha$且$n'_0=0$，$\widehat{T}_n$就可以重新表示为：

$$\widehat{T}_n=\sum_{\alpha=1}^{k}\frac{2g_{\alpha\alpha}}{n_\alpha(n_\alpha-1)}\sum_{i<j}\mathbf{x}_{\alpha i}^T\mathbf{x}_{\alpha j}+\sum_{1\le\alpha<\beta\le k}\frac{g_{\alpha\beta}}{n_\alpha n_\beta}\sum_{i,j}\mathbf{x}_{\alpha i}^T\mathbf{x}_{\beta j}=\sum_{j=2}^{n'_k}\sum_{i=1}^{j-1}\psi_{ij}.$$

其中，

$$\psi_{ij}=\begin{cases}\dfrac{2g_{\alpha\beta}}{n_\alpha n_\beta}\mathbf{y}_i^T\mathbf{y}_j, & i\in\{n'_{\alpha-1}+1,\cdots,n'_\alpha\},\ j\in\{n'_{\beta-1}+1,\cdots,n'_\beta\},\alpha<\beta,\\ \dfrac{2g_{\alpha\alpha}}{n_\alpha(n_\alpha-1)}\mathbf{y}_i^T\mathbf{y}_j, & i,j\in\{n'_{\alpha-1}+1,\cdots,n'_\alpha\},i<j.\end{cases}$$

令 $\mathrm{B}_{nj}=\sum_{i=1}^{j-1}\psi_{ij},\mathrm{A}_{nn'}=\sum_{j=2}^{n'}\sum_{i=1}^{j-1}\psi_{ij}=\sum_{j=2}^{n'}\mathrm{B}_{nj}$，$n'=1,2,\cdots,n'_k$。为了可以更好的解释 T_n^* 的渐近性，Zhang[66]等人提出了以下引理：

引理 4.1[66]：记 $\Pi_{nn'}=\sigma\{\mathbf{y}_1,\cdots,\mathbf{y}_{n'}\}$ 是 $\{\mathbf{y}_1,\cdots,\mathbf{y}_{n'}\}$ 生成的 σ 域，对于每个 n，$\{\mathrm{A}_{nn'}\}_{n'=1}^{n'_k}$ 是关于 σ 域 $\Pi_{nn'}$ 的零均方可积鞅。

引理 4.2[66]：在假设条件 I-III 下，有

$$Var(\sum_{j=n'_{\alpha-1}+1}^{n'_\alpha}\mathrm{E}(\mathrm{B}_{nj}^2\left|\Pi_{n,j-1}\right.))=o(\sigma^4(\widehat{T}_n)).$$

$$\frac{\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^2\left|\Pi_{n,j-1}\right.)}{\sigma^2(\widehat{T}_n)}\xrightarrow{p}1,\quad\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^4)=o(\sigma^4(\widehat{T}_n)).$$

另外，通过式(4.5)的 $T_n=\widehat{T}_n+2S_n+\left\|\mathbf{G}\boldsymbol{\mu}\right\|^2$，可以知道其中的 $\widehat{T}_n$ 和 T_n 在原假设下具有相同的分布。基于 $S_n=(\mathbf{G}\boldsymbol{\mu})^T\mathbf{G}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})$ 可以得出 S_n 的方差，如下所示：

$$Var(S_n)=(\mathbf{G}\boldsymbol{\mu})^T\mathbf{G}diag(\boldsymbol{\Sigma}_1/n_1,\cdots\boldsymbol{\Sigma}_k/n_k)\mathbf{G}^T(\mathbf{G}\boldsymbol{\mu})=\sum_{\alpha=1}^{k}\sum_{\beta=1}^{k}\sum_{\gamma=1}^{k}\frac{g_{\alpha\beta}g_{\beta\gamma}}{n_\beta}\boldsymbol{\mu}_\alpha^T\boldsymbol{\Sigma}_\beta\boldsymbol{\mu}_\gamma.$$

其中：$g_{\alpha\beta}$ 和 $g_{\beta\gamma}$ 是矩阵 $\mathbf{G}$ 中的元素。再根据假设 IV 得到：

$$Var(S_n)=o(\sigma^2(\widehat{T}_n)).\tag{4.7}$$

鉴于上述假设条件以及说明，本文通过以下两个定理证明了新检验统计量 T_n^* 分别在原假设和备择假设下的渐近正态性。

定理 4.1：在上文给定的假设条件 I-IV 下，当 $n,p\to\infty$ 时，$\dfrac{\widehat{T}_n}{\hat{\sigma}(\widehat{T}_n)}\xrightarrow{d}N(0,1).$

证明：通过引理 4.2 中的 $\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^4)=o(\sigma^4(\widehat{T}_n))$ 可得到 $\sigma^{-4}(\widehat{T}_n)\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^4)\to0$，利用 Markov 不等式，对 $\forall\varepsilon>0$

$$\sum_{j=1}^{n}\sigma^{-2}(\widehat{T}_n)\mathrm{E}(\mathrm{B}_{nj}^2I(\left|\mathrm{B}_{nj}\right|>\varepsilon\sigma(\widehat{T}_n))\left|\Pi_{nn'}\right.)$$

$$=\sum_{j=1}^{n}\sigma^{-2}(\widehat{T}_n)\mathrm{E}(\mathrm{B}_{nj}^2\left|\Pi_{nn'}\right.)\mathrm{E}(\left|\mathrm{B}_{nj}\right|>\varepsilon\sigma(\widehat{T}_n))\left|\Pi_{nn'}\right.)$$

$$\le\sum_{j=1}^{n}\sigma^{-4}(\widehat{T}_n)\varepsilon^{-2}\mathrm{E}(\mathrm{B}_{nj}^4\left|\Pi_{nn'}\right.).$$

进而可以得出以下的林德伯格条件：

$$\sigma^{-2}(\widehat{T}_n)\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^2 I(\left|\mathrm{B}_{nj}\right|>\varepsilon\sigma(\widehat{T}_n))\left|\Pi_{nn'}\right.)\xrightarrow{p}0.$$

之后根据引理 4.2 有 $\dfrac{\sum_{j=1}^{n}\mathrm{E}(\mathrm{B}_{nj}^2\left|\Pi_{n,j-1}\right.)}{\sigma^2(\widehat{T}_n)}\xrightarrow{p}1$ 成立。

结合鞅差极限定理[67]便可得出 $\dfrac{\widehat{T}_n}{\sigma(\widehat{T}_n)}\xrightarrow{d}N(0,1)$，

又因为有 $\dfrac{\sigma^2(\widehat{T}_n)}{\hat{\sigma}^2(\widehat{T}_n)}\xrightarrow{p}1$ 明显成立，

所以，综上可得 $\dfrac{\widehat{T}_n}{\hat{\sigma}(\widehat{T}_n)}\xrightarrow{d}N(0,1)$。

定理 4.2：在同定理 4.1 相同的假设条件、广义多元相依模型模型(4.6)以及式(4.7)成立的情况下，当 $n\to\infty, p\to\infty$ 时，可以得到：

$$\mathrm{P}(T_n^*>z_\alpha)=\Phi(-z_\alpha+\frac{\left\|\mathbf{G}\boldsymbol{\mu}\right\|^2}{\sigma(\widehat{T}_n)})(1+o(1)).$$

其中 z_α 和 $\Phi(\bullet)$ 分别表示标准正态分布的上侧 α 分位数和累积分布函数。

证明：由(4.7)可得 $\dfrac{S_n}{\sigma(\widehat{T}_n)}\to o_p(1)$，又因为 $\dfrac{\sigma^2(\widehat{T}_n)}{\hat{\sigma}^2(\widehat{T}_n)}\xrightarrow{p}1$ 成立，再根据(4.5)，

$$\mathrm{P}(T_n^*>z_\alpha)=\mathrm{P}(\frac{T_n}{\hat{\sigma}(\widehat{T}_n)}>\mathrm{z}_\alpha)=\mathrm{P}(\frac{\widehat{T}_n+2S_n+\left\|\mathbf{G}\boldsymbol{\mu}\right\|^2}{\hat{\sigma}(\widehat{T}_n)}>z_\alpha)=\Phi(-\mathrm{z}_\alpha+\frac{\left\|\mathbf{G}\boldsymbol{\mu}\right\|^2}{\sigma(\widehat{T}_n)})(1+o(1)).$$

定理结论得证。

4.3 模拟实验

对于高维总体均值问题(4.1)，Zhou 等[57]提出的 T_Z 可以表示为：

$$T_Z=\sum_{\alpha=1}^{k}\frac{n_\alpha(n-n_\alpha)}{n^2}\overline{\mathbf{y}}_\alpha^T\overline{\mathbf{y}}_\alpha-\sum_{\alpha,\beta}\frac{n_\alpha n_\beta}{n^2}\overline{\mathbf{y}}_\alpha^T\overline{\mathbf{y}}_\beta-\sum_{\alpha=1}^{k}\frac{n_\alpha(n-n_\alpha)}{n^2}\frac{tr(\hat{\boldsymbol{\Sigma}}_\alpha)}{n_\alpha}$$

$$= n^{-1}\Big(\sum_{\alpha=1}^{k} n_\alpha \left\|\bar{\mathbf{y}}_\alpha - \bar{\mathbf{y}}\right\|^2 - \sum_{\alpha=1}^{k}(1-\frac{n_\alpha}{n})tr(\hat{\mathbf{\Sigma}}_\alpha)\Big) = n^{-1}T_{Y\text{-}H}.$$

$T_{Y\text{-}H}$ 是 Yamada 和 Himeno 在 2015 年提出的检验统计量，进行标准化后得到：

$$T_{Y\text{-}H}^{*} = T_{Y\text{-}H} \Big/ \sqrt{Var(\hat{T}_{Y\text{-}H}\,|H_0)}.$$

在这一小节中，新检验统计量的显著性进行了模拟验证。模拟的主要目的是对于问题(4.1)，去比较新检验统计量 T_n^* 和统计量 $T_{Y\text{-}H}^*$ 在不同情形下检验效果的差异。模拟实验中，设置了一些调整参数：

1.选取 4 个样本(k=4)；模拟的样本维度 p 分别取 500,800,1000；样本量分别取三组不同的情形：$\mathbf{n_1} = (n_1, n_2, n_3, n_4)$ =(10,30,40,50)，$\mathbf{n}_2$ =(20,60,80,100)，$\mathbf{n}_3$ =(40,120,160,200)。

2.为了生成不同的样本组均值向量，令 $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\mu}_2 = 1.5\delta\mathbf{h}, \boldsymbol{\mu}_3 = \delta\mathbf{h}, \boldsymbol{\mu}_4 = 2\delta\mathbf{h}$，其中：$\mathbf{h} = \mathbf{u}/\|\mathbf{u}\|$，$\mathbf{u} = (1, 2, \cdots, p)^T$，用区间[1,3]上的值 δ 来控制样本组间均值向量的变化。

3.为了生成样本组协方差矩阵，令 $\mathbf{\Lambda}_\alpha = \lambda_\alpha \mathbf{R}_\alpha^{1/2}\mathbf{M}^{1/2}$ $(\alpha = 1,2,3,4)$，其中：$\lambda_\alpha(\alpha = 1,2,3,4)$ 是正的常数；$\mathbf{M} = diag(1/p, 2/p, \cdots, 1)$；$\mathbf{R}_\alpha$ 是方阵，具体形式为 $\mathbf{R}_\alpha = (r_\alpha^{|i-j|}) : p \times p, \alpha = 1,2,3,4$。

在本文 k 样本在不等协方差下的高维数据假设检验问题中，即 r_α 和 λ_α 在 α 不同时分别取不同的值：$r_1 = 0.8, r_2 = 0.6, r_3 = 0.4, r_4 = 0.1$ 和 $\lambda_1 = \sqrt{2}, \lambda_2 = 1, \lambda_3 = \sqrt{3}, \lambda_4 = 2$。另外模拟实验针对正态模型和非正模型分别展开，具体设置如下：

模型 1：样本 $z_{\alpha il}, l = 1, \ldots, p \overset{i.i.d.}{\sim} N(0,1)$。

模型 2：$z_{\alpha il} = w_{\alpha il}/\sqrt{2}, l = 1, \ldots, p$，其中 $w_{\alpha il}$，$l = 1, \ldots, p \overset{i.i.d.}{\sim} t(4)$。

模型 3：$z_{\alpha il} = (w_{\alpha il} - 1)/\sqrt{2}, l = 1, \ldots, p$，其中 $w_{\alpha il}$，$l = 1, \ldots, p \overset{i.i.d.}{\sim} \chi^2(1)$。

本模拟实验设定模拟实验的评价标准为平均相对误差：

$$ARE = 100M^{-1}\sum_{i=1}^{M}|\alpha_i - \alpha|/\alpha.$$

其中：$\alpha = 5\%$ 是标准水平，α_i 是检验水平，M 是模拟的组数，即样本量和样本维数情形数的乘积(3×3=9)。当检验统计量所对应的平均相对误差 ARE 值越小，则相应测试方法的总体性能越好。模拟重复实验取 1000 次。

表 4.1 模型 1 下的检验水平(百分比)

p	500			800			1000			ARE
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3	
T_n^*	6.06	6.37	5.94	5.62	4.28	4.14	5.08	5.71	4.16	11.28
$T_{Y\text{-}H}^*$	6.81	6.39	7.53	6.89	6.97	5.11	5.89	4.77	6.55	27.48

表 4.2 模型 1 下检验功效(百分比)

p	500			800			1000		
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
δ	2.4	1.6	1.2	2.6	1.8	1.4	2.8	2.0	1.5
T_n^*	96.56	97.51	98.01	96.81	95.71	99.01	96.15	98.74	97.73
$T_{Y\text{-}H}^*$	95.39	89.97	95.04	86.32	93.06	98.01	95.76	96.07	94.33

从表 4.1-4.6 的模拟实验结果可以看出：

1. 表 4.1 和表 4.2 分别列出了在模型 1 下，T_n^*和$T_{Y\text{-}H}^*$检验水平和检验功效的模拟比较结果；表 4.31 和表 4.4 分别列出了在模型 2 下，T_n^*和$T_{Y\text{-}H}^*$检验水平和检验功效的模拟比较结果；表 4.5 和表 4.26 分别列出了在模型 3 下，T_n^*和$T_{Y\text{-}H}^*$检验水平和检验功效的模拟比较结果。

2. 在表 4.1 和表 4.2 中可以看出在模型 1：正态分布下，$T_{Y\text{-}H}^*$的 ARE 值 27.48 明显大于T_n^*的 ARE 值 11.28，说明T_n^*比$T_{Y\text{-}H}^*$的模拟检验水平要更接近于给定的标准水平，统计量T_n^*有更好的总体性能。表 4.2 中可以看出在不同情形下，T_n^*比$T_{Y\text{-}H}^*$的模拟数值要大一点。尤其在维数等于 800，样本量在$\mathbf{n_1}=(10,30,40,50)$的情况下，$T_n^*$的值 96.81 要比$T_{Y\text{-}H}^*$的值 86.32 大。总的来说，$T_n^*$的检验功效更佳。

3. 表 4.3 和表 4.4 记录了在模型 2：T 分布（非正态分布）下T_n^*和$T_{Y\text{-}H}^*$的检验水平和检验功效。在表 4.3 中，在样本量和维数不同的情形下，T_n^*的检验水平比$T_{Y\text{-}H}^*$的检验水平更接近标准水平。进一步从评价标准 ARE 来看，T_n^*的 ARE 值 11.57 比$T_{Y\text{-}H}^*$的 ARE 值 28.00 明显要小，这也说明T_n^*的总体性能比$T_{Y\text{-}H}^*$的总体性能更好。而表 4.4 是T_n^*和$T_{Y\text{-}H}^*$的检验功效比较，在不同的情形下，T_n^*的检验功效比$T_{Y\text{-}H}^*$的检验功效相对较大，这也表明在模型 2 下，T_n^*要比$T_{Y\text{-}H}^*$表现更优。

4. 在表 4.5 和表 4.6 中可以看出在模型 3：Gamma 分布下，$T_{Y\text{-}H}^*$和T_n^*的实验比较结果和模型 2 以及模型 3 相类似。表 4.5 中T_n^*的 ARE 值 7.24 明显小于$T_{Y\text{-}H}^*$的 ARE 值 21.66，说明统计量T_n^*有更好的总体性能。表 4.6 中可以看出在在绝大部分的情形下，T_n^*比$T_{Y\text{-}H}^*$的模拟数值要大一点，表现更稳健。

表 4.3 模型 2 下检验水平(百分比)

p	500			800			1000			ARE
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3	
T_n^*	4.50	5.70	5.43	5.67	5.56	5.54	5.25	6.01	4.45	11.57
$T_{Y\text{-}H}^*$	7.33	6.36	7.03	6.28	6.13	6.32	6.33	6.17	5.65	28.00

表 4.4 模型 2 下检验功效(百分比)

p	500			800			1000		
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
δ	2.4	1.6	1.2	2.6	1.8	1.4	2.8	2.0	1.5
T_n^*	96.56	97.51	98.01	96.87	96.69	98.93	96.73	95.22	97.62
$T_{Y\text{-}H}^*$	95.70	93.62	94.45	94.63	95.98	98.82	95.08	93.99	96.41

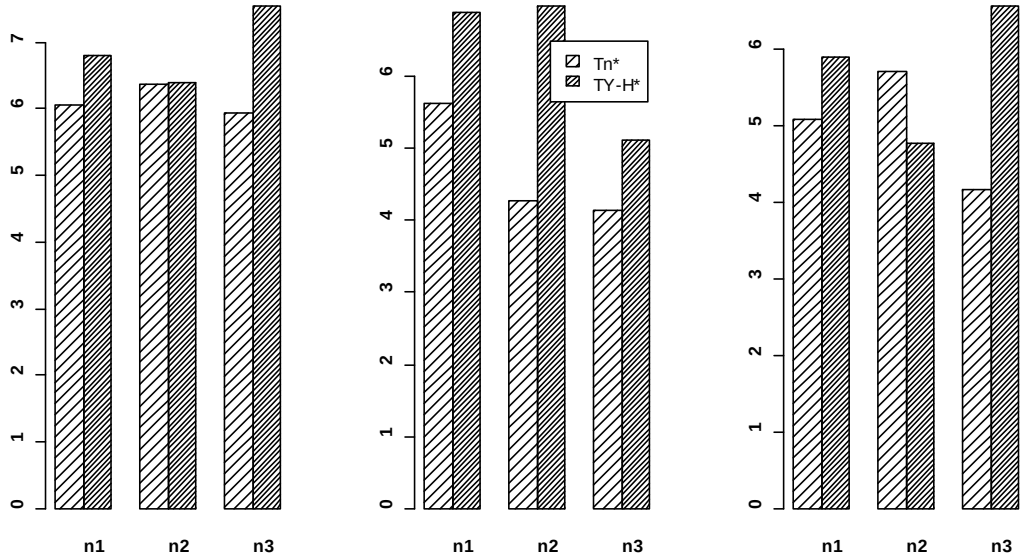
表 4.5 模型 3 下检验水平(百分比)

p	500			800			1000			ARE
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3	
T_n^*	5.41	6.37	4.84	4.63	5.31	5.18	5.03	5.30	4.87	7.24
$T_{Y\text{-}H}^*$	5.86	5.41	5.49	6.64	6.58	5.66	7.05	5.53	6.53	21.66

表 4.6 模型 3 下检验功效(百分比)

p	500			800			1000		
n	n_1	n_2	n_3	n_1	n_2	n_3	n_1	n_2	n_3
δ	2.4	1.6	1.2	2.6	1.8	1.4	2.8	2.0	1.5
T_n^*	93.96	92.16	96.86	95.72	96.79	97.59	97.25	98.38	96.53
$T_{Y\text{-}H}^*$	92.91	91.14	95.72	91.69	93.00	98.19	95.91	94.53	97.49

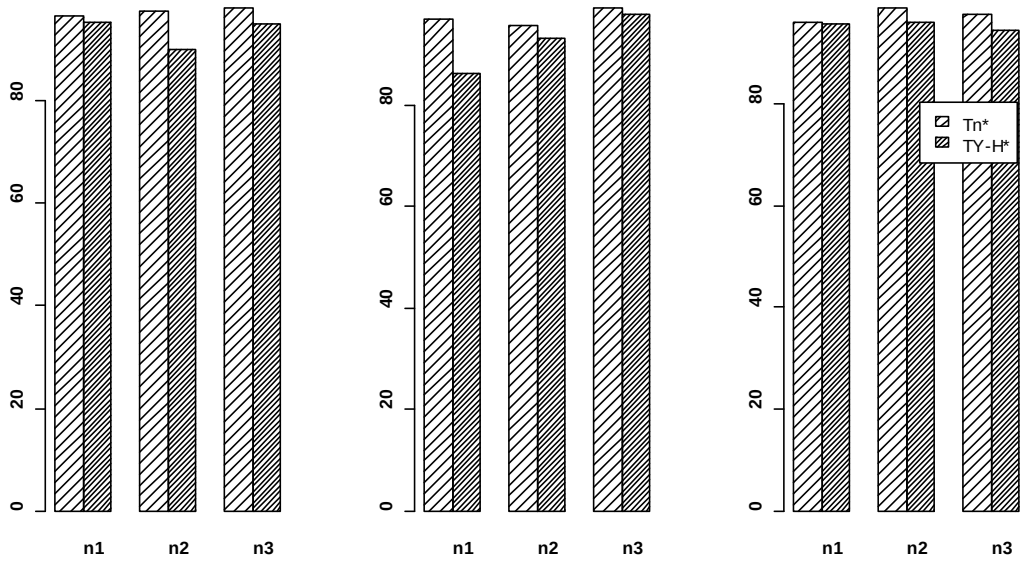
为了更直观的说明实验结果，在这里选取了正态模型，分别绘出了 T_n^* 与 $T_{Y\text{-}H}^*$ 的检验水平对比图和检验功效对比图，见图 4.1，图 4.2。



a)维数 500　　b)维数 800　　c)维数 1000

图 4.1 T_n^* 与 $T_{Y\text{-}H}^*$ 的检验水平对比图

图 4.1 显示在不同维度下，对于不同的样本量，T_n^*要比$T_{Y\text{-}H}^*$更具稳健性，当数据在维度 p 取 1000，样本量为 $\mathbf{n}_1$=(10,30,40,50)时，T_n^*的检验水平明显优于$T_{Y\text{-}H}^*$，而且更接近于给定的标准水平。大部分情形下，T_n^*的表现均比$T_{Y\text{-}H}^*$好。图 4.2 显示，T_n^*的检验功效值均比$T_{Y\text{-}H}^*$大，表明T_n^*要比$T_{Y\text{-}H}^*$优势显著。



a)维数 500　　b)维数 800　　c)维数 1000

图 4.2 T_n^*与$T_{Y\text{-}H}^*$的检验功效对比图

4.4 小结

文章针对高维数据下的 MANOVA 检验问题，提出了一种新的检验统计量，其不仅消除了对数据维数和个数的关系限制，并在一般的条件下，说明了新统计量T_n^*分别在原假设和局部备择假设下的渐近正态性。模拟结果表明：在正态分布、t 分布以及卡方分布三种不同分布的模型下，本文提出的新检验统计量—T_n^*较已有的检验统计量检验水平更高，显著性更强，更具竞争优势，进而为高维 MANOVA 检验问题提供了一种新的有力选择。

第五章　总结与展望

本文主要探讨了两个主要内容，具体如下：

第一个是针对实变量维数和样本量成比例增长的情况，就双样本协方差矩阵的相等性检验，采用大维随机矩阵理论提出一个新的检验统计量，并利用线性谱统计量的中心极限定理证明了新检验统计量的渐近性，并在论文中对新统计量和 Xu 进行了系统的比较研究。通过两种检验统计量的模拟比较，我们可以看到新检验统计量对于正态分布以及非正态分布都有很好的检验效果。

第二个是当样本维数远远大于样本量时，针对多样本总体均值检验问题，借鉴 Chen 等[48]的思想同样提出一个新的检验统计量，做到了检验数据从两样本到多样本的提升和完善。并消除了这类检验问题对样本维数和样本量之间的关系限制，同时对统计量简化了运算。此外，本文证明了所提统计量在原假设和备择假设的渐近性。而且对新统计量和 Yamada 和 Himeno 在 2015 年提出的检验统计量进行了模拟实验比较，通过这种比较，表明了本文所提统计量在不同分布下都展示了更好的稳健性。

通过本文的研究，我们也可以看到通过随机矩阵等理论所提出的一些检验统计量，并不一定是局限在样本维数小于样本量的情况下。这是因为真正决定新检验方法是否适用于“大 p 小 n”的高维数据是被修正的检验统计量。随着随机矩阵相关理论的不断完善，其运用范围越来越广，这为之后运用随机矩阵理论解决高维数据的统计分析问题提供可能，使之拥有更实际的广泛应用。

参 考 文 献

[1] 单洪明, 张军平. 实值多变量维数约简：综述[J]. 自动化学报, 2018.44(02): 193-215.

[2] Hotelling H. The generalization of student's Ratio [J]. The Annals of Mathematical Statistics, 1930, 2(3).

[3] Arnold L. On Wigner's semicircle law for the eigenvalues of random matrices [J].Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete, 1971, 19(3):191-198.

[4] Arnol L. On The asymptotic distirbution of the eigenvalues of random matrices[J]. J.math.anal.appl, 1967, 20(2):262-268.

[5] Bai Z D, Yin Y Q . Convergence to the semicircle law[J]. The Annals of Probability, 1988, 16(2):863-875.

[6] V.A.Marchenko, L.A.Pastur. Distribution of eigenvalues for some sets of random matrices[J]. Mathematics of the USSR-Sbornik, 1967, 1(1):507-536.

[7] Pastur L A. Spectra of random Self-adjoint operators[J]. Journal of Soviet Mathematics, 1989,46(4):1979-2021.

[8] Bai Z D, Yin Y Q , Krishnaiah P R . On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic [J]. Journal of Multivariate Analysis, 1986, 19(1):189-200.

[9] Grenander U, Silverstein J W . Spectral analysis of networks with random topologies [J]. SIAM Journal on Applied Mathematics, 1977, 32(2):499-519.

[10] Jonsson D. Some limit theorems for the eigenvalues of a sample covariance matrix [J]. Journal of Multivariate Analysis, 1982, 12(1):1-38.

[11] Wachter K W. The limiting empirical measure of multiple discriminant ratios[J]. The Annals of Statistics, 1980, 8(5):937-957.

[12] Wachter K W. The strong limits of random matrix spectra for sample matrices of independent elements[J]. The Annals of Probability, 1978, 6(1):1-18.

[13] Yin Y Q. Limiting spectral distribution for a class of random matrices[J]. Journal of Multivariate Analysis, 1986, 20(1):50-68.

[14] Yin Y Q, Krishnaiah P R . A limit theorem for the eigenvalues of product of two random matrices[J]. Journal of Multivariate Analysis, 1983,13(4):489-507.

[15] Yin Y Q, Bai Z D , Krishnaiah P R . On the limit of the largest eigenvalue of the large dimensional sample covariance matrix[J]. Probability Theory and Related Fields, 1988, 78(4):509-521.

[16] Yin Y Q, Bai Z D , Krishnaiah P R . Limiting behavior of the eigenvalues of a multivariate F matrix[J]. Journal of Multivariate Analysis, 1983, 13(4):508-516.

[17] Bai Z, Jiang D, Yao J F , et al. Corrections to LRT on large-dimensional covariance matrix by RMT[J]. The Annals of Statistics, 2009,37(6B):3822-3840.

[18] Zheng, S. Central limit Theorems for linear spectral statistics of large dimensional F-matrices[J]. Ann Inst Henri Poincaré, Probab Statist, 2012, 48(2): 444-476.

[19] Xu L, Liu B, Zheng S, et al. Testing proportionality of two large-dimensional covariance matrices[J]. Computational Statistics & Data Analysis, 2014, 78: 43-55.

[20] Jiang D. Tests for large dimensional covariance structure based on Rao's score Test[J]. Journal of Multivariate Analysis, 2016, 152:28-39.

[21] 王艺霏, 李贤, 来骥等. 基于随机矩阵理论的智能电网大数据体系结构设计[J]. 计算技术与自动化, 2019(3): 17-21.

[22] 贺兴, 艾芊, 邱才明等. 泛在电力物联网数据挖掘体系建设综述及数据驱动认知框架探究[J]. 电器与能效管理技术, 2019(19): 1-14.

[23] 胡旭光, 马大中, 孙秋野等. 基于随机矩阵谱偏离度的微网状态感知方法研究[J]. 中国电机工程学报, 2019, 39(21): 6238-6247.

[24] 林超, 郑霖, 张文辉, 邓小芳. 基于随机矩阵理论的 WSN 异常节点定位算法[J]. 计算机工程, 2020(01): 157-163.

[25] 倪广县, 陈金海, 王恒. 滚动轴承高维随机矩阵状态异常检测算法[J]. 西安交通大学学报, 2019(10): 65-71.

[26] Bai Z D. Methodologies in spectral analysis of large dimensional random matrices, AReview. Statistica Sinica, 1999, 9(3):611-677.

[27] Bai Z D, Silberstein, J. W. Spectral analysis of large-dimensional random matrices[M]. First Edition, Beijing: Science Press, 2006.

[28] Bai Z D , Yin Y Q , Krishnaiah P R . On the limiting empirical distribution function of the eigenvalues of a multivariate *F*-matrix[J]. Theory of Probability and Its Applications, 1987, 32(3):537-548.

[29] Silberstein, J. W. the limiting eigenvalue distribution of a multivariate f matrix[J]. SIAM J.Math Anal. 1985,16(3):641-646.

[30] Silverstein J W , Bai Z D . CLT for linear spectral statistics of large-dimensional sample covariance matrices[J]. The Annals of Probability, 2004,32(1A):553-605.

[31] 丁绍芳, 李从珠, 王建稳. 双样本协方差矩阵和均值向量齐性检验的研究及其应用[J]. 数学的实践与认识, 2004(08): 96-104.

[32] 丁绍芳, 李从珠, 陈明. 双样本的协方差齐性检验及其在刑事技术中的应用[J]. 数理统计与管理, 1998(01): 13-20.

[33] Federer, Walter T . Testing Proportionality of Covariance Matrices[J]. The Annals of Mathematical Statistics, 1951, 22(1):102-106.

[34] Pillai, K.C.S., Al-Ani, S., Jouris, G.M.. On the distribution of the ratios of the roots of a covariance matrix and wilks' crterion for tests of three hypotheses. Ann. Math. Statist. 1969, 40(6), 2033-2040.

[35] Flury, B.K.. Proportionality of k covariance matrices[J].Statistics & Probability Letters. 1986. 4(1): 29-33.

[36] Schott J R . A test for proportional covariance matrices[J]. Computational Statistics & Data Analysis, 1999, 32(2):135-146.

[37] Gupta D S, Giri N. Properties of tests concerning covariance matrices of normal distributions[J]. The Annals of Statistics, 1973(6): 1222-1224.

[38] Gupta, Das S. Properties of power functions of some tests concerning dispersion matrices of multivariate normal distributions[J]. The Annals of Mathematical Statistics, 1969, 40(2):697-701.

[39] Gupta A K, Tang J. Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models[J]. Biometrika, 1984, 71(3): 555-559.

[40] O'Brien P C. Robust procedures for testing equality of covariance matrices[J]. Biometrics, 1992, 48(3):819-827.

[41] Perlman, Michael D. Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations[J]. The Annals of Statistics, 1980, 8(2): 247-263.

[42] Sugiura N, Nagao H. Unbiasedness of some test criteria for the equality of one or two covariance matrices[J]. The Annals of Mathematical Statistics, 1968, 39(5):1686-1692.

[43] Nagao H. Asymptotic non-null distributions of two test criteria for equality of covariance matrices under local alternatives[J]. Annals of the Institute of Statistical Mathematics, 1974, 26(1): 395-402.

[44] John S. Some optimal multivariates tests[J]. Biometrika, 1971, 58(1): 123-127.

[45] Hisao N. On Some test criteria for covariance matrix[J]. The Annals of Statistics, 1973, 1(4): 700-709.

[46] Jiang D, Jiang T, Yang F. Likelihood ratio tests for covariance matrices of high-dimensional normal distributions[J]. Journal of Statistical Planning & Inference, 2012, 142(8): 2241-2256.

[47] Chen S X, Zhang L X, Zhong P S. Tests for high-dimensional covariance matrices[J]. Journal of the American Statistical Association, 2010, 105(490): 810-819.

[48] 何冰, 薄晓玲. 基于随机 *F*-矩阵的高维双样本协方差矩阵相等性检验[J]. 吉林大学学报(理学版), 2019, 57(01): 71-77.

[49] Zhang J, Xu J. On the k-sample Behrens-Fisher Problem for high-dimensional data[J]. Science in China, 2009, 52(6):1285-1304.

[50] Xu F F, Huang J C, Wen Z W. High dimensional covariance matrix estimation using multi-factor models from incomplete information[J]. Science China Mathematics, 2015, 58(4):829-844.

[51] Srivastava M S, Fujikoshi Y. Multivariate analysis of variance with fewer observations than the dimension[J]. Journal of Multivariate Analysis, 2006, 97(9):1927-1940.

[52] Yamada T, Srivastava M S. A test for multivariate analysis of variance in high dimension[J]. Communications in Statistics-Theory and Methods, 2012, 41(13-14):2602-2615.

[53] Srivastava M S, Kubokawa T. Tests for multivariate analysis of variance in high dimension under non-normality[J]. Journal of Multivariate Analysis, 2013, 115:204-216.

[54] Yamada T, Himeno T. Testing homogeneity of mean vectors under heteroscedasticity in high-dimension [J]. Journal of Multivariate Analysis, 2015, 139(02):7-27.

[55] 曹明响, 徐兴忠. 高维数据下 MANOVA 检验[J]. 北京理工大学报, 2015, 35(08): 868-871.

[56] Zhou B, Guo J, Zhang J T. High-dimensional general linear hypothesis testing under heteroscedasticity[J]. Journal of Statistical Planning and Inference, 2017(03):36-54.

[57] Chen S X, Qin Y. A two sample test for high dimensional data with applications to gene-set testing[J]. Mpra Paper, 2012, 38(2):808-835.

[58] Feng, L. Zou, C. Wang, Z. Zhu, L. Two sample Behrens-Fisher problem for high-dimensional data[J]. Statist Sinica, 2015(25): 1297-1312.

[59] 陈敏琼. 单样本 U 统计量及其渐近性质[J]. 数学的实践与认识, 2017, 47(18): 296-303.

[60] Chen S, Qin Y. A two-sample test of high dimensional data with applications to gene-set testing[J]. The Annals of Statistics, 2010(38):808-835.

[61] 李顺勇,王一静.不同协方差下高维数据的 MANOVA 检验问题[J].山西大学学报(自然科学版),2019,42(03): 499-506.

[62] Bai Z D, Saranadasa H . Effect of high dimension: by an example of a two sample problem[J]. Statistica Sinica, 1996, 6(2):311-329.

[63] Himeno T, Yamada T. Estimations for some functions of covariance matrix in high dimension under non-normality and its applications[J]. Journal of Multivariate Analysis, 2014, 130(04): 27-44.

[64] Srivastava M S, Katayama S, Kano Y. A two sample test in high dimensional data[J]. Journal of Multivariate Analysis, 2013, 114(1):349-358.

[65] 梁成扬, 金华, 何美仪, 谭银冰. 基于推断模型的 Behrens-Fisher 问题的精确检验[J]. 统计与决策, 2018, 34(17): 18-20.

[66] Zhang J T, Guo J, Zhou B. Linear hypothesis testing in high-dimensional one-way MANOVA[J]. Journal of Multivariate Analysis,2017,155(01):200-216.

[67] 闫广州, 张丽娜. 任意随机变量序列的一个强极限定理[J]. 山西大学学报(自然科学版), 2008, 31(03): 323-325.

攻读学位期间取得的研究成果

[1] 李顺勇,王一静.不同协方差下高维数据的 MANOVA 检验问题[J].山西大学学报(自然科学版),2019,42(03): 499-506.

[2] 李顺勇,王一静,张晓琴.基于 F-矩阵的高维协方差相等性检验[J].审稿中.

致　谢

时间如白驹过隙，随着论文的收尾，三年的研究生求学生活也接近尾声。回首这三年，自己在学习和生活中都有所进步，并且走向成熟，感谢在我求学路上给我提供帮助的各位老师，同学和亲人。

首先，感谢我的导师李顺勇教授对我学业所倾注的心血。本论文是在李老师的悉心指导下完成的，无论是论文的选题、章节的结构，还是论文的多次修改、最终定稿，李老师都给予了自己悉心的指导和无私的帮助，我的毕业论文才能够得以顺利完成。此外，李老师在生活上也给予了我很多温暖的照顾和帮助，在我们生活中遇到问题时，李老师总是耐心倾听并给出合理性建议，使我的科研生涯能够顺利进行，不再迷茫。李老师始终秉持严谨的科研态度，让我受益匪浅。在今后的学习生活中也会向李老师学习。在此，谨向我的导师李老师表示衷心的感谢。

同时，我要感谢数学科学学院的所有老师们，感谢他们为我们创造良好宁静的学习环境，感谢他们在学习和生活上的关心和帮助；感谢求学期间同门师兄弟、师姐们，感谢他们在生活和学习上的指导与帮助；最后，感谢我的亲人、朋友的关怀包容与支持鼓励，他们是我科研期间最坚实的后盾，陪伴着我顺利完成学业。

短暂的求学道路将画上圆满的句号，但人生新的起点刚刚开始，我会更加严格要求自己，以百倍的热情去工作，去学习，去努力，去奋斗。

个人简况及联系方式

个人简况：

姓　　名　王一静

性　　别　女

籍　　贯　山西省长治市

个人简历：

2017.9-2020.7　　山西大学　　统计学

2012.9-2016.7　　吕梁学院　　数学与应用数学

承　诺　书

本人郑重声明：所呈交的学位论文，是在导师指导下独立完成的，学位论文的知识产权属于山西大学。如果今后以其他单位名义发表与在读期间学位论文相关的内容，将承担法律责任。除文中已经注明引用的文献资料外，本学位论文不包括任何其他个人或集体已经发表或撰写过的成果。

作者签名：王一静

2020 年 5 月 30 日

学位论文使用授权声明

本人完全了解山西大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关机关或机构送交论文的复印件和电子文档，允许论文被查阅和借阅，可以采用影印、缩印或扫描等手段保存、汇编学位论文。同意山西大学可以用不同方式在不同媒体上发表、传播论文的全部或部分内容。

保密的学位论文在解密后遵守此协议。

作者签名：王一静

导师签名：李顺勇

2020年 5 月 31日