分类号_____ 密级_____

UDC_____ 编号_____

華中師範大學

硕士学位论文

高维两样本比对问题的一种新统计检验方法

学位申请人姓名：胡玉婷

申请学位学生类别：全日制硕士

申请学位学科专业：数理统计学

指导教师姓名：李正帮 副教授

硕士学位论文

高维两样本比对问题的一种新统计检验方法

论文作者：胡玉婷

指导教师：李正帮 副教授

学科专业：数理统计学

研究方向：生物医学统计

华中师范大学数学与统计学学院

2022 年 4 月

硕士学位论文
MASTER'S THESIS

A New Statistical Test Method for High-Dimensional Two-Sample Comparison

A thesis

submitted in Partial Fulfillment of the Requirements

for the M.S.Degree in Statistics

By

Yuting Hu

Postgradute Program

School of Mathematics and Statistics

Central China Normal University

Supervisor Zhengbang Li

Academic Title Associate Professor **Signature**

Approved

Apr 2022

华中师范大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果. 除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果. 对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明. 本声明的法律结果由本人承担.

作者签名: 胡玉婷　　日期: 2022年 5 月 30 日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学. 学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文. (保密的学位论文在解密后遵守此规定) 保密论文注释：本学位论文属于保密，在____年解密后适用本授权书. 非保密论文注释：本学位论文不属于保密范围，适用本授权书.

作者签名: 胡玉婷　　导师签名:
日期: 2022 年 5 月 30 日　　日期: 2022 年 5 月 30 日

本人已经认真阅读“CALIS高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS 高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益. 同意论文提交后滞后：□半年；□一年；□二年发布.

作者签名: 胡玉婷　　导师签名:
日期: 2022年 5 月 30 日　　日期: 2022 年 5 月 30 日

摘要

当今时代，高维数据在各个领域出现得越来越频繁，如历史研究、医学、金融、社会发展等方面，给传统的多元分析理论及相关的方法带来了巨大的挑战。在高维数据假设检验的问题中，现有的方法有好的一面，同时也有一定的局限性：一方面，经典的检验统计量多数用正态近似计算 p 值，有很强的局限性；另一方面，经典的检验统计量在样本数据不同分量上的方差不同的情况下检验效果较差，经典的极限理论也就不再实用。故此基于 $Welch-Satterthwaite$ 卡方近似方法提出当多个分量方差不等时的表现较好的一种新检验方法。与此同时，通过 R 语言进行统计模拟，以及实际数据分析，将新的检验方法和以往的检验方法进行比较，结果发现我们提出的统计量在一定的条件下在检验功效和经验水平上都有着比较好的表现，能够较好的处理多种类型的高维数据，并对于不同维度上的方差不同的样本数据的两样本均值检验问题也有着较好的检验效果。

关键词：高维数据；两样本均值检验；二次型；$Welch-Satterthwaite$ 卡方近似

Abstract

In today's era, high-dimensional data appears more and more frequently in various fields, such as historical research, medicine, finance, social development, and so on, they have brought great challenges to the traditional multivariate analysis theory and related methods.

In the problem of hypothesis testing of high-dimensional data, the existing methods have good sides, but also have certain limitations: on the one hand, most of the classical test statistics use the normal approximation to calculate the p value, which has strong limitations. On the other hand, when the variances of different components of the sample data are different, the classical test methods have poor test effect, and thay are not applicable. Therefore, Based on the $Welch-Satterthwaite$ chi-square approximation method,a new test method that performs better when the variances of multiple components are not equal is proposed. a new test statistic is proposed by improving the original statistics. At the same time, through the R language for statistical simulation and actual data analysis, the new test method is compared with the previous test method, and it is found that the statistic we proposed has a good test effect and empirical size under certain conditions. It can handle various types of high-dimensional data, and has a good test effect on the two-sample mean test of sample data with different variances in different dimensions.

Keywords: high-dimensional data ; two sample mean test; the quadratic form ; the Welch-Satterthwaite chi-square approximation method.

目录

第一章 绪论

1.1 研究背景

随着科学技术的发展，人们现在所解决的问题在数据上相比之前更加庞大，在实际工作中收集的数据的类型也更加多样，这种情况给统计学的工作带来了挑战。虽然目前已经有了一些方法，但针对复杂系统的统计建模、高维和不彻底数据的统计分析等问题都有待探索更完美的解决途径。比如，在生物统计问题中，实际所能获得的样本量往往很少，但是每个样本量的生物统计信息却很多，这种数据就是我们常说的高维数据，同时在许多时候这些数据彼此之间还可能有较强的相关性。比如对于基因数据来说，就常常会出现样本量相对于其维数而言反而相当小的情况。在实际生活中统计学工作者和其他从事数据分析的工作者都针对这些问题进行了相应的研究，也已经取得了一定的成绩，但是在其他某些方面还处在探索阶段，还没有形成系统的解决方式。

在研究中，我们所说的高维数据往往是数据的维度高于样本的大小，即大家所熟知的“大 p，小 n”的情况。这种情况下，协方差矩阵不可逆，原来所适用的一些经典的极限理论便不能处理这种情况了。在高维数据假设检验的问题中，一方面，经典的检验统计量多数用正态近似检验 p 值，有很强的局限性；另一方面，经典的检验统计量在样本数据内部多个分量方差不等的情况下检验效果较差，经典的极限理论也就不再实用。

在许多应用领域中，我们需要检验来自两个总体的高维随机向量的两个均值向量是否相等，在此我们规定一下符号：假设随机向量 $\boldsymbol{X}=(X_1,X_2...,X_p)$，其均值向量为 $\boldsymbol{\mu}_1=(\mu_{11},\mu_{12},...,\mu_{1p})$，随机向量 $\boldsymbol{Y}=(Y_1,Y_2...,Y_p)$，其均值向量为 $\boldsymbol{\mu}_2=(\mu_{21},\mu_{22},...,\mu_{2p})$，$X_i=(x_{1i},x_{2i}...,x_{n_1i})^T$，$Y_i=(y_{1i},y_{2i}...,y_{n_2i})^T$，$i=(1,2,...,p)$。共同的协方差矩阵为 $\boldsymbol{\Sigma}=(\sigma_{q,s})_{q,s=1,...,p}$，$n=n_1+n_2$。

我们考虑如下的假设检验问题:

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \qquad vs \qquad H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2.$$

当 p 固定且 n 远大于 p 时, 传统的 $Hotelling\ T^2$ 检验可以很好地解决问题,其检验统计量为

$$T_H=\frac{n_1n_2}{n}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}).$$

而当 p 远大于 n 时，这时的协方差矩阵是奇异的，则有 $\boldsymbol{\Sigma}$ 不可逆，也就无法求出 $\hat{\boldsymbol{\Sigma}}^{-1}$，检验统计量 T_H 就无法计算，传统的 $Hotelling\ T^2$ 检验方法也就不再适

用。针对这个问题，Srivastava, Du (2008) [4] 基于 Σ 矩阵正定的假设下提出了检验方法，一定程度上解决了“大 p，小 n”情况下的假设检验问题，之后 Li, Liu (2021) [3]基于 *bootstrap* 算法提出了一种新的检验方法，也有着较好的检验效果。Chen, Qin (2010) [2] 提出了 L^2-norm 检验方法，能较好的处理高维数据并有理想的检验效果。但是我们认为该检验方法的功效仍有提升的空间，并且已有方法不能较好地处理多个分量上的方差不同的数据，于是我们将在已有的方法上加以改进，提出了检验方法 L^2new 来解决这一问题。

本文从两样本均值检验问题出发，基于 $Welch-Satterthwaite(W-S)$ 卡方近似方法(Satterthwaite 1946; Welch 1947 [5])，在改进原有统计量的基础上，提出了新的检验统计量。与此同时，通过 R 语言进行统计模拟，并对模拟结果进行分析，以此来对两种检验方法进行比较，看我们提出的检验方法 L^2new 与 L^2-norm 方法相比，其在犯第一类错误的概率和检验功效上是否有着比较好的表现。

1.2 预备知识

这一节将对下文中的证明和理论叙述部分将会应用到的相关符号、定义、定理、引理等进行介绍。

定义1. 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_l)^T$ 为 l 维随机变量，称矩阵

$$\mathbf{C}=(c_{ij})_{l\times l}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1l} \\ c_{21} & c_{22} & \cdots & c_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ c_{l1} & c_{l2} & \cdots & c_{ll} \end{pmatrix}.$$

为 $\boldsymbol{X}$ 的协方差矩阵，也写作 $D(\boldsymbol{X})$，内部元素具体为

$$c_{ij}=Cov(X_i,X_j),i,j=1,2,\cdots,l.$$

其含义是分量 X_i 和 X_j 的协方差。

定义2. 原假设指的是要检验的假设，记作 H_0；备择假设指的是在拒绝原假设时所接受的假设，记作 H_1。

定义3. 第一类错误指的是当原假设 H_0 为真时，检验结果却是拒绝 H_0，其概率记作 α，即$P($ 拒绝 $H_0|H_0$ 为真$)=\alpha$，也称作弃真错误。

第二类错误指的是当原假设 H_0 为假时，检验结果却是接受 H_0，其概率记作 β，即$P($接受 $H_0|H_0$ 为假$)=\beta$，也称作受伪错误。

定义4. 如果 $\mathbf{A}$ 为 $a \times b$ 的矩阵，而 $\mathbf{B}$ 为 $c \times d$ 的矩阵，则克罗内克积 $\mathbf{A} \otimes \mathbf{B}$ 为 $ac \times bd$ 的分块矩阵，具体如下

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1b}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{a1}\mathbf{B} & \cdots & a_{ab}\mathbf{B} \end{pmatrix},$$

$$= \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1d} & \cdots & \cdots & a_{1b}b_{11} & a_{1b}b_{12} & \cdots & a_{1b}b_{1d} \\ a_{11}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2d} & \cdots & \cdots & a_{1b}b_{21} & a_{1b}b_{22} & \cdots & a_{1b}b_{2d} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{11}b_{c1} & a_{11}b_{c2} & \cdots & a_{11}b_{cd} & \cdots & \cdots & a_{1b}b_{c1} & a_{1b}b_{c2} & \cdots & a_{1b}b_{cd} \\ \vdots & \vdots & & \vdots & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \ddots & \vdots & \vdots & & \vdots \\ a_{a1}b_{11} & a_{a1}b_{12} & \cdots & a_{a1}b_{1d} & \cdots & \cdots & a_{ab}b_{11} & a_{ab}b_{12} & \cdots & a_{ab}b_{1d} \\ a_{a1}b_{21} & a_{a1}b_{22} & \cdots & a_{a1}b_{2d} & \cdots & \cdots & a_{ab}b_{21} & a_{ab}b_{22} & \cdots & a_{ab}b_{2d} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ a_{a1}b_{c1} & a_{a1}b_{c2} & \cdots & a_{a1}b_{cd} & \cdots & \cdots & a_{ab}b_{c1} & a_{ab}b_{c2} & \cdots & a_{ab}b_{cd} \end{pmatrix}.$$

定义5. 若 $\mathbf{A}$ 是 $a \times b$ 的矩阵，将其元素按列连接，即每一列的元素都按次序接到前一列的元素后面，则最后组成的长向量称为 $\mathbf{A}$ 的按列拉直算子，记作 $vec(\mathbf{A})$。

$$vec(\mathbf{A}) = (a_{11}, a_{21}, a_{31}, ..., a_{12}, a_{22}, a_{32}, ..., a_{1b}, a_{2b}, a_{3b}, ...a_{ab})^T.$$

设文中的随机变量 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$，$\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$，令 $\boldsymbol{\Sigma} = (\sigma_{ij})$，令 $D_\sigma = diag(\sigma_{11}, ..., \sigma_{pp})$，有 $\mathcal{R}_p = D_\sigma^{-1/2}\boldsymbol{\Sigma}D_\sigma^{-1/2} = (\rho_{ij})$，其中 $\mathcal{R}_p$ 为随机变量的相关系数矩阵，其对角线元素为1。

假设1. 随机变量的相关系数矩阵满足如下条件：

$$1 \leq \lim_{p\to\infty} \frac{tr\mathcal{R}_p^i}{p} < \infty \quad (i = 1, 2, 3, 4).$$

为方便以下论文中证明部分的书写，在不改变样本数据的基础上将随机变量 X 和 Y 分别按行整合得到相互独立的高维样本 $m_{11}, ..., m_{1n_1}$ 和 $m_{21}, ...m_{2n_2}$，满足 $E(m_{11}, ..., m_{1n_1}) = \bar{\boldsymbol{m}}_1$，$E(m_{21}, ...m_{2n_2}) = \bar{\boldsymbol{m}}_2$。下面的定理 1 为概率论教材的基础知识，其正确性已得到证明。

定理1. $\boldsymbol{\Sigma}$ 的一个无偏估计量为

$$\hat{\boldsymbol{\Sigma}} = (n-2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (m_{ij} - \bar{\boldsymbol{m}}_i)(m_{ij} - \bar{\boldsymbol{m}}_i)^T.$$

记 $\hat{\mathbf{\Sigma}}_i = (n_i-1)^{-1}\sum_{j=1}^{n_i}(m_{ij}-\bar{\boldsymbol{m}}_i)(m_{ij}-\bar{\boldsymbol{m}}_i)^T$，$i=1,2$。

推论1. *$tr(\mathbf{\Sigma})$ 的一个无偏估计量为*

$$\widehat{tr(\mathbf{\Sigma})} = \frac{(n_1-1)tr(\hat{\mathbf{\Sigma}}_1)+(n_2-1)tr(\hat{\mathbf{\Sigma}}_2)}{n-2}.$$

证明： 结合定理 1 可知$E(\hat{\mathbf{\Sigma}}) = \mathbf{\Sigma}$，则

$$\begin{aligned}
E[\widehat{tr(\mathbf{\Sigma})}] =& E[\frac{(n_1-1)tr(\hat{\mathbf{\Sigma}}_1)+(n_2-1)tr(\hat{\mathbf{\Sigma}}_2)}{n-2}],\\
=& E\{\frac{(n_1-1)tr[(n_1-1)^{-1}\sum_{j=1}^{n_1}(m_{1j}-\bar{\boldsymbol{m}}_1)(m_{1j}-\bar{\boldsymbol{m}}_1)^T]}{n-2}\\
&+\frac{(n_2-1)tr[(n_2-1)^{-1}\sum_{j=1}^{n_2}(m_{2j}-\bar{\boldsymbol{m}}_2)(m_{2j}-\bar{\boldsymbol{m}}_2)^T]}{n-2}\},\\
=& tr[E(\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(m_{ij}-\bar{\boldsymbol{m}}_i)(m_{ij}-\bar{\boldsymbol{m}}_i)^T}{n-2})],\\
=& tr(\mathbf{\Sigma}).
\end{aligned}$$

所以 $\widehat{tr(\mathbf{\Sigma})}$ 是 $tr(\mathbf{\Sigma})$ 的无偏估计量。 □

引理1. *[马尔可夫 $(Markov)$ 不等式] 假设对非负随机变量 $\mathbf{X}$ 都有 $E(\mathbf{X})$ 存在，则对任意常数 $\epsilon>0$ 有*

$$P(\mathbf{X}\geq\epsilon)\leq\frac{E(\mathbf{X})}{\epsilon}.$$

证明： 根据 X 是非负随机变量，得出

$$\begin{aligned}
E(\mathbf{X}) &= \int_0^{\infty} xP(\mathbf{X}=x)dx,\\
&= \int_0^{\epsilon} xP(\mathbf{X}=x)dx + \int_{\epsilon}^{\infty} xP(\mathbf{X}=x)dx,\\
&\geq \int_{\epsilon}^{\infty} xP(\mathbf{X}=x)dx,\\
&\geq \int_{\epsilon}^{\infty} \epsilon P(\mathbf{X}=x)dx,\\
&= \epsilon\int_{\epsilon}^{\infty} P(\mathbf{X}=x)dx,
\end{aligned}$$

$$=\epsilon P(\mathbf{X} \geq \epsilon).$$

于是得到

$$E(\mathbf{X}) \geq \epsilon P(\mathbf{X} \geq \epsilon) \Longrightarrow P(\mathbf{X} \geq \epsilon) \leq \frac{E(\mathbf{X})}{\epsilon}.$$

所以马尔可夫不等式得证。 □

引理2. [切比雪夫 (*Chebyshev*) 不等式] 假设对随机变量 $\mathbf{X}$ 和常数 $\epsilon > 0$，都有 $E(\mathbf{X})$ 和 $D(\mathbf{X})$ 存在，则有

$$P(|\mathbf{X} - E(\mathbf{X})| \geq \epsilon) \leq \frac{1}{\epsilon^2} D(\mathbf{X}).$$

证明： 如果事件A的示性函数用 $I[A]$ 来表示，则 $|\mathbf{X} - E(\mathbf{X})|$ 取任何值，都有

$$\begin{aligned} I[|\mathbf{X} - E(\mathbf{X})| \geq \epsilon] =& I[(\mathbf{X} - E(\mathbf{X}))^T(\mathbf{X} - E(\mathbf{X})) \geq \epsilon^2], \\ \leq& \frac{(\mathbf{X} - E(\mathbf{X}))^T(\mathbf{X} - E(\mathbf{X}))}{\epsilon^2}. \end{aligned}$$

于是对两边取期望，得到

$$\begin{aligned} E\{I[|\mathbf{X} - E(\mathbf{X})| \geq \epsilon]\} &\leq E[\frac{(\mathbf{X} - E(\mathbf{X}))^T(\mathbf{X} - E(\mathbf{X}))}{\epsilon^2}]. \\ \Longrightarrow P(|\mathbf{X} - E(\mathbf{X})| \geq \epsilon) &= \frac{1}{\epsilon^2} D(\mathbf{X}). \end{aligned}$$

所以切比雪夫不等式得证。 □

引理3. 设 $\mathbf{W} \sim W_p(a, \Sigma/a)$，则 $tr(\Sigma)$、$tr^2(\Sigma)$、$tr(\Sigma^2)$ 的一个无偏估计量分别可如下表示：

$$\begin{aligned} \widehat{tr(\Sigma)} &= tr(\mathbf{W}), \\ \widehat{tr^2(\Sigma)} &= \frac{a(a+1)}{(a-1)(a+2)}[tr^2(\mathbf{W}) - \frac{2}{a+1} tr(\mathbf{W}^2)], \\ \widehat{tr(\Sigma^2)} &= \frac{a^2}{(a-1)(a+2)}[tr(\mathbf{W}^2) - \frac{1}{a} tr^2(\mathbf{W})]. \end{aligned}$$

且对所有的 p，当 $a \to \infty$ 时，上述三个估计量对 $tr(\Sigma)$、$tr^2(\Sigma)$、$tr(\Sigma^2)$ 是比率一致的。

证明： 因为 $\mathbf{W} \sim W_p(a, \Sigma/a)$，故由 *Wishart* 分布的性质可以得出

$$E[tr(\mathbf{W})] = tr(\Sigma), \qquad Var[tr(\mathbf{W})] = \frac{2}{a} tr(\Sigma).$$

进一步根据 Schott(2007,p.1829) [18]，得出

$$E[tr^2(\mathbf{W})] = tr^2(\mathbf{\Sigma}) + \frac{2}{a}tr(\mathbf{\Sigma}^2).$$
$$E[tr(\mathbf{W}^2)] = \frac{1}{a}tr^2(\mathbf{\Sigma}) + \frac{a+1}{a}tr(\mathbf{\Sigma}^2).$$

进而可求出无偏估计的均值分别为

$$\begin{aligned}E[\widehat{tr^2(\mathbf{\Sigma})}] &= \frac{a(a+1)}{(a-1)(a+2)}\{tr^2(\mathbf{\Sigma}) + \frac{2}{a}tr(\mathbf{\Sigma}^2) - \frac{2}{a+1}[\frac{1}{a}tr^2(\mathbf{\Sigma}) + \frac{a+1}{a}tr(\mathbf{\Sigma}^2)]\},\\ &= \frac{a(a+1)}{(a-1)(a+2)}[\frac{(a-1)(a+2)}{a(a+1)}tr^2(\mathbf{\Sigma})] = tr^2(\mathbf{\Sigma}).\\ E[\widehat{tr(\mathbf{\Sigma}^2)}] &= \frac{a^2}{(a-1)(a+2)}\{\frac{1}{a}tr^2(\mathbf{\Sigma}) + \frac{a+1}{a}tr(\mathbf{\Sigma}^2) - \frac{1}{a}[tr^2(\mathbf{\Sigma}) + \frac{2}{a}tr(\mathbf{\Sigma}^2)]\},\\ &= \frac{a^2}{(a-1)(a+2)}[\frac{(a-1)(a+2)}{a^2}tr(\mathbf{\Sigma}^2)] = tr(\mathbf{\Sigma}^2).\end{aligned}$$

所以可以得出结论 $\widehat{tr(\mathbf{\Sigma})}$、$\widehat{tr^2(\mathbf{\Sigma})}$、$\widehat{tr(\mathbf{\Sigma}^2)}$分别为 $tr(\mathbf{\Sigma})$、$tr^2(\mathbf{\Sigma})$、$tr(\mathbf{\Sigma}^2)$ 的无偏估计量。

进一步根据 Schott(2007,p.1829-1830) [18]，得出如下结果

$$\begin{aligned}Var[\widehat{tr^2(\mathbf{\Sigma})}] =& \frac{1}{(a-1)^2(a+2)^2}\{[8a+\mathcal{O}(a)]tr(\mathbf{\Sigma}^4) + [16a^2+\mathcal{O}(a^2)]tr(\mathbf{\Sigma}^3)tr(\mathbf{\Sigma})\\ &+ [8a^2+\mathcal{O}(a^2)]tr^2(\mathbf{\Sigma}^2) + [8a^2+\mathcal{O}(a^2)]tr(\mathbf{\Sigma}^2)tr(\mathbf{\Sigma})\}.\\ Var[\widehat{tr(\mathbf{\Sigma}^2)}] =& \frac{1}{(a-1)^2(a+2)^2}\{[8a^3+\mathcal{O}(a^3)]tr(\mathbf{\Sigma}^4) + [4a^2+\mathcal{O}(a^2)]tr(\mathbf{\Sigma}^2)\}.\end{aligned}$$

因为对于非负矩阵 $\boldsymbol{A}$ 有 $tr(\boldsymbol{A}^r)/tr^r(\boldsymbol{A}) \leq 1, r = 1, 2, \ldots$，所以当 $a \to \infty$ 时，可以得出如下结果

$$Var[\frac{tr(\mathbf{W})}{tr(\mathbf{\Sigma})}] = \frac{2tr(\mathbf{\Sigma}^2)}{atr^2(\mathbf{\Sigma})} \leq \frac{2}{a} \to 0.$$

$$\begin{aligned}Var[\frac{\widehat{tr^2(\mathbf{\Sigma})}}{tr^2(\mathbf{\Sigma})}] =& \frac{1}{(a-1)^2(a+2)^2}\{[8a+\mathcal{O}(a)]\frac{tr(\mathbf{\Sigma}^4)}{tr^4(\mathbf{\Sigma})} + [16a^2+\mathcal{O}(a^2)]\frac{tr(\mathbf{\Sigma}^3)}{tr^3(\mathbf{\Sigma})}\\ &+ [8a^2+\mathcal{O}(a^2)][\frac{tr(\mathbf{\Sigma}^2)}{tr^2(\mathbf{\Sigma})}]^2tr^2(\mathbf{\Sigma}^2) + [8a^2+\mathcal{O}(a^2)]\frac{tr(\mathbf{\Sigma}^2)}{tr^2(\mathbf{\Sigma})}\},\\ \leq& \frac{32a^2+\mathcal{O}(a^2)}{(a-1)^2(a+2)^2} \to 0.\end{aligned}$$

$$Var[\frac{\widehat{tr(\mathbf{\Sigma}^2)}}{tr(\mathbf{\Sigma}^2)}] = \frac{1}{(a-1)^2(a+2)^2}\{[8a^3+\mathcal{O}(a^3)]\frac{tr(\mathbf{\Sigma}^4)}{tr^2(\mathbf{\Sigma}^2)} + [4a^2+\mathcal{O}(a^2)]\},$$

$$\leq \frac{8a^3+\mathcal{O}(a^3)}{(a-1)^2(a+2)^2} \to 0.$$

因此可以知道估计量$\widehat{tr(\boldsymbol{\Sigma})}$、$\widehat{tr^2(\boldsymbol{\Sigma})}$、$\widehat{tr(\boldsymbol{\Sigma}^2)}$分别对$tr(\boldsymbol{\Sigma})$、$tr^2(\boldsymbol{\Sigma})$、$tr(\boldsymbol{\Sigma}^2)$是比率一致的。

综上所述，引理 3 证毕。 □

推论2. $tr^2(\boldsymbol{\Sigma})$、$tr(\boldsymbol{\Sigma}^2)$的一个无偏估计量为

$$\widehat{tr^2(\boldsymbol{\Sigma})} = \frac{(n-2)(n-1)}{(n-3)n}[tr^2(\hat{\boldsymbol{\Sigma}}) - \frac{2}{n-1}tr(\hat{\boldsymbol{\Sigma}}^2)].$$
$$\widehat{tr(\boldsymbol{\Sigma}^2)} = \frac{(n-2)^2}{(n-3)n}[tr(\hat{\boldsymbol{\Sigma}}^2) - \frac{tr^2(\hat{\boldsymbol{\Sigma}})}{n-2}].$$

证明： 因为$\hat{\boldsymbol{\Sigma}} = (n-2)^{-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(m_{ij}-\bar{\boldsymbol{m}}_i)(m_{ij}-\bar{\boldsymbol{m}}_i)^T$，所以由$Wishart$分布的定义可知$\hat{\boldsymbol{\Sigma}} \sim W_p[n-2, \boldsymbol{\Sigma}/(n-2)]$，结合引理 1 可知$tr^2(\boldsymbol{\Sigma})$、$tr(\boldsymbol{\Sigma}^2)$的无偏估计量为

$$\begin{aligned}
\widehat{tr^2(\boldsymbol{\Sigma})} =& \frac{(n-2)(n-2+1)}{(n-2-1)(n-2+2)}[tr^2(\hat{\boldsymbol{\Sigma}}) - \frac{2}{n-2+1}tr(\hat{\boldsymbol{\Sigma}}^2)],\\
=& \frac{(n-2)(n-1)}{(n-3)n}[tr^2(\hat{\boldsymbol{\Sigma}}) - \frac{2}{n-1}tr(\hat{\boldsymbol{\Sigma}}^2)].\\
\widehat{tr(\boldsymbol{\Sigma}^2)} =& \frac{(n-2)^2}{(n-2-1)(n-2+2)}[tr(\hat{\boldsymbol{\Sigma}}^2) - \frac{1}{n-2}tr^2(\hat{\boldsymbol{\Sigma}})],\\
=& \frac{(n-2)^2}{(n-3)n}[tr(\hat{\boldsymbol{\Sigma}}^2) - \frac{tr^2(\hat{\boldsymbol{\Sigma}})}{n-2}].
\end{aligned}$$

所以推论 2 得证。 □

引理4. $\mathbf{A}$ 和 $\mathbf{B}$ 均为 $c*c$ 对称矩阵，$\boldsymbol{X} \sim N_c(\mathbf{0}, \boldsymbol{\Omega})$ 且 $\boldsymbol{\Omega}$ 正定，则有

$$E(\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}\boldsymbol{X}) = tr(\mathbf{A}\boldsymbol{\Omega})tr(\mathbf{B}\boldsymbol{\Omega}) + 2tr(\mathbf{A}\boldsymbol{\Omega}\mathbf{B}\boldsymbol{\Omega}).$$

证明： 对 $E(\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}\boldsymbol{X})$ 做出如下变形，将期望转化为迹的形式

$$\begin{aligned}
E(\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}\boldsymbol{X}) =& E\{tr[(\boldsymbol{X}^T\otimes\boldsymbol{X}^T)(\mathbf{A}\otimes\mathbf{B})(\boldsymbol{X}\otimes\boldsymbol{X})]\},\\
=& E\{tr[(\mathbf{A}\otimes\mathbf{B})(\boldsymbol{X}^T\otimes\boldsymbol{X}^T)(\boldsymbol{X}\otimes\boldsymbol{X})]\},\\
=& E\{tr[(\mathbf{A}\otimes\mathbf{B})(\boldsymbol{X}\boldsymbol{X}^T\otimes\boldsymbol{X}\boldsymbol{X}^T)]\},\\
=& tr[(\mathbf{A}\otimes\mathbf{B})E(\boldsymbol{X}\boldsymbol{X}^T\otimes\boldsymbol{X}\boldsymbol{X}^T)],
\end{aligned}$$

$$
\begin{aligned}
&=tr\{(\mathbf{A}\otimes\mathbf{B})[2N_c(\mathbf{\Omega}\otimes\mathbf{\Omega})+vec(\mathbf{\Omega})(vec(\mathbf{\Omega}))^T]\},\\
&=tr\{(\mathbf{A}\otimes\mathbf{B})[(I_{c^2}+K_{cc})(\mathbf{\Omega}\otimes\mathbf{\Omega})+vec(\mathbf{\Omega})(vec(\mathbf{\Omega}))^T]\},\\
&=tr[(\mathbf{A}\otimes\mathbf{B})(\mathbf{\Omega}\otimes\mathbf{\Omega})]+tr[(\mathbf{A}\otimes\mathbf{B})K_{cc}(\mathbf{\Omega}\otimes\mathbf{\Omega})]\\
&\quad+tr[(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{\Omega})(vec(\mathbf{\Omega}))^T].
\end{aligned}
$$

将 $E(\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}\boldsymbol{X})$ 变形后的三个部分分别进行化简计算，得出

$$
\begin{aligned}
&tr[(\mathbf{A}\otimes\mathbf{B})(\mathbf{\Omega}\otimes\mathbf{\Omega})]=tr(\mathbf{A\Omega}\otimes\mathbf{B\Omega})=tr(\mathbf{A\Omega})tr(\mathbf{B\Omega}).\\
&tr[(\mathbf{A}\otimes\mathbf{B})K_{cc}(\mathbf{\Omega}\otimes\mathbf{\Omega})]=tr[\mathbf{A\Omega}\otimes\mathbf{B\Omega}K_{cc}]=tr(\mathbf{A\Omega B\Omega}).\\
&tr[(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{\Omega})(vec(\mathbf{\Omega}))^T]=(vec(\mathbf{\Omega}))^T(\mathbf{A}\otimes\mathbf{B})vec(\mathbf{\Omega}),\\
&\qquad=(vec(\mathbf{\Omega}))^Tvec(\mathbf{B\Omega A})=tr(\mathbf{A\Omega B\Omega}).
\end{aligned}
$$

将上述三个部分相加，得出结论

$$
\begin{aligned}
E(\boldsymbol{X}^T\mathbf{A}\boldsymbol{X}\boldsymbol{X}^T\mathbf{B}\boldsymbol{X})&=tr(\mathbf{A\Omega})tr(\mathbf{B\Omega})+tr(\mathbf{A\Omega B\Omega})+tr(\mathbf{A\Omega B\Omega}),\\
&=tr(\mathbf{A\Omega})tr(\mathbf{B\Omega})+2tr(\mathbf{A\Omega B\Omega}).
\end{aligned}
$$

即引理 4 得证。 □

第二章　两样本均值向量比较问题的一种新假设检验统计方法

2.1　现有的其他检验方法及研究动机

传统的 *Hotelling* T^2 检验能够很好地处理样本量 n 大于维数 p 的问题，其检验统计量为

$$T_H = \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}). \tag{2.1}$$

而 n 大于 p 的高维数据的协方差矩阵是奇异的，此时 *Hotelling* T^2 检验不再适用。针对这个问题，Srivastava, Du (2008) [4] 基于 $\boldsymbol{\Sigma}$ 正定的假设下提出了新的检验方法，该方法有着较好的检验效果但限制性较强。而后 Chen, Qin (2010) [2]提出了 *L2N* 检验方法，该方法能较好地处理高维数据并有理想的检验效果，其检验统计量为

$$T_{n,p} = \frac{n_1 n_2}{n}\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\|^2. \tag{2.2}$$

L2N 检验方法能很好地解决维数 p 大于样本量 n 的问题，但是其主要解决不同维度上的方差大致相同的情况，在不同维度上的方差差异较大时，其检验效果不太理想，所以我们认为该检验方法仍有提升的空间，于是我们在已有的方法上加以改进，在此基础上提出新的检验统计量

$$\begin{aligned} T_{new} &= \frac{n_1 n_2}{n}[\frac{\|\bar{X}_1 - \bar{Y}_1\|^2}{\hat{\sigma_{11}}} + \frac{\|\bar{X}_2 - \bar{Y}_2\|^2}{\hat{\sigma_{22}}} + \ldots + \frac{\|\bar{X}_p - \bar{Y}_p\|^2}{\hat{\sigma_{pp}}}]. \\ &= \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T D_s^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}). \end{aligned} \tag{2.3}$$

其中 $D_s = \frac{\widehat{D(\boldsymbol{X})}}{n_1} + \frac{\widehat{D(\boldsymbol{Y})}}{n_2}$。

因为 T_{new} 中涉及到估计值 D_s，不便于推导其理论分布，所以下文将假设知道其真实值，通过推导得出新统计量近似版本 $\widetilde{T}_{new}$ 的理论分布，进一步可知检验统计量 T_{new} 的理论分布。

$$\widetilde{T}_{new} = \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T D_\sigma^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}).$$

本文首先通过引理了解某一二次型统计量的渐近分布，而后证明 $\widetilde{T}_{new}$ 在 H_0 假设成立的条件下依概率收敛到该二次型，从而得出 $\widetilde{T}_{new}$ 的渐近分布，据此可知检验统计量 T_{new} 的理论分布，进一步得出检验 p 值和功效函数的计算方法，并进行统计模拟，通过对比模拟结果来验证该检验方法的适用性。

2.2 新统计量的渐近分布

以下首先通过引理 5 了解某一二次型统计量的渐近分布，而后合理猜想新统计量的渐近分布，而在后面的证明中可知，在 H_0 假设成立的条件下，T_{new} 依概率收敛到 $\widetilde{T}_{new}$，而套用该引理可知 $\widetilde{T}_{new}$ 的渐近分布，从而猜想的正确性可得到证明。

2.2.1 二次型统计量的 $W-S$ 卡方渐近分布

以下我们引用引理 5 的目的是用著名的 $W-S$ 卡方分布来近似 $\widetilde{T}_{new}$ 的分布，即用 $R \overset{d}{=} \beta\chi_d^2$ 来近似 $\widetilde{T}_{new}$ 的分布，其中 β, d 是未知参数，可通过计算 $\widetilde{T}_{new}$ 与 R 的均值与方差来确定参数，而在证明T_{new} 依概率收敛到 $\widetilde{T}_{new}$ 之后，推论的正确性也就得到证明了。以下引理来自于 Katayama, Kano, Srivastava(2013) [10]，其正确性已经得到证明。

引理5. 当 $(n,p) \to \infty$ 时，在 H_0 假设成立的条件下有

$$T_{n,p,0} = \frac{n_1 n_2}{n}||\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}||^2 \xrightarrow{L} T_{p,0} \overset{d}{=} \sum_{r=1}^{p} \lambda_{p,r} A_r.$$

且有

$$E(T_{p,0}) = tr(\boldsymbol{\Sigma}), Var(T_{p,0}) = 2tr(\boldsymbol{\Sigma}^2).$$
$$E[T_{p,0} - E(T_{p,0})]^3 = 8tr(\boldsymbol{\Sigma}^3).$$

其中 $\boldsymbol{\Sigma}$ 为 $\boldsymbol{X}, \boldsymbol{Y}$ 的加权协方差矩阵 $\boldsymbol{\Sigma} = \frac{D(\boldsymbol{X})}{n_1} + \frac{D(\boldsymbol{Y})}{n_2}$，$\lambda_{p,1}, ..., \lambda_{p,p}$ 为 $\boldsymbol{\Sigma}$ 的按降序排列的特征值，$A_1, ..., A_p$ 为相互独立的服从 χ_1^2 分布的随机变量。

推论3. 当 $n \to \infty$ 时，对任意大的常数 p，在 H_0 假设成立的条件下有

$$T_{new} = \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T D_s^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) \xrightarrow{L} T_{p,0} \overset{d}{=} \sum_{r=1}^{p} \lambda_{p,r} A_r.$$

且有

$$E(T_{p,0}) = tr(D_{1/\sigma}\boldsymbol{\Sigma}), Var(T_{p,0}) = 2tr(D_{1/\sigma}\boldsymbol{\Sigma})^2.$$

其中 $\boldsymbol{\Sigma}$ 为 $\boldsymbol{X}, \boldsymbol{Y}$ 的加权协方差矩阵 $\boldsymbol{\Sigma} = \frac{D(\boldsymbol{X})}{n_1} + \frac{D(\boldsymbol{Y})}{n_2}$，$\lambda_{p,1}, ..., \lambda_{p,p}$ 为 $D_{1/\sigma}\boldsymbol{\Sigma}$ 的按降序排列的特征值，$A_1, ..., A_p$ 为相互独立的服从 χ_1^2 分布的随机变量。

2.2.2 新统计量近似版本的近似分布

以下我们将证明在 H_0 假设成立的条件下，T_{new} 依概率收敛到 $\widetilde{T}_{new}$。

$$\begin{aligned} T_{new} &= \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^T D_s^{-1}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}), \\ &= \sqrt{\frac{n_1 n_2}{n}}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^T D_\sigma^{-1/2} D_{s/\sigma}^{-1} D_\sigma^{-1/2}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\sqrt{\frac{n_1 n_2}{n}}. \end{aligned} \tag{2.4}$$

令 $Z_p = D_\sigma^{-1/2}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\sqrt{\frac{n_1 n_2}{n}}$，则有

$$T_{new} = Z_p^T D_{s/\sigma}^{-1} Z_p.$$

在 H_0 假设成立的条件下，$\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}} \xrightarrow{L} N_p(\mathbf{0}, \frac{n}{n_1 n_2}\boldsymbol{\Sigma})$，故

$$Z_p = D_\sigma^{-1/2}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})\sqrt{\frac{n_1 n_2}{n}}. \tag{2.5}$$

$$Z_p \overset{H_0}{\sim} N_p(\mathbf{0}, \mathcal{R}_p). \tag{2.6}$$

在此，取

$$D_1 = \begin{pmatrix} \frac{s_{11}}{\sigma_{11}}-1 & & & \\ & \frac{s_{22}}{\sigma_{22}}-1 & & \\ & & \vdots & \\ & & & \frac{s_{pp}}{\sigma_{pp}}-1 \end{pmatrix}. \tag{2.7}$$

$$D_2 = \begin{pmatrix} \frac{\sigma_{11}}{s_{11}}+\frac{s_{11}}{\sigma_{11}}-2 & & & \\ & \frac{\sigma_{22}}{s_{22}}+\frac{s_{22}}{\sigma_{22}}-2 & & \\ & & \vdots & \\ & & & \frac{\sigma_{pp}}{s_{pp}}+\frac{s_{pp}}{\sigma_{pp}}-2 \end{pmatrix}. \tag{2.8}$$

则有

$$D_{s/\sigma}^{-1} = \begin{pmatrix} 1-(\frac{s_{11}}{\sigma_{11}}-1)+(\frac{\sigma_{11}}{s_{11}}+\frac{s_{11}}{\sigma_{11}}-2) & & & \\ & 1-(\frac{s_{22}}{\sigma_{22}}-1)+(\frac{\sigma_{22}}{s_{22}}+\frac{s_{22}}{\sigma_{22}}-2) & & \\ & & \vdots & \\ & & & 1-(\frac{s_{pp}}{\sigma_{pp}}-1)+(\frac{\sigma_{pp}}{s_{pp}}+\frac{s_{pp}}{\sigma_{pp}}-2) \end{pmatrix},$$

$$= \mathbf{I}_p - D_1 + D_2. \tag{2.9}$$

所以得出

$$\begin{aligned} T_{new} &= Z_p^T D_{s/\sigma}^{-1} Z_p, \\ &= Z_p^T (\mathbf{I}_p - D_1 + D_2) Z_p, \\ &= Z_p^T Z_p - Z_p^T D_1 Z_p + Z_p^T D_2 Z_p. \end{aligned} \tag{2.10}$$

记 $A = Z_p^T Z_p$, $B = Z_p^T D_1 Z_p$, $C = Z_p^T D_2 Z_p$, 因为 $\boldsymbol{X}, \boldsymbol{Y}$ 与 $s_{ii}(i = 1, 2, ..., p)$ 独立，所以 Z_p 与 $s_{ii}(i = 1, 2, ..., p)$ 独立，推出 Z_p 与 D_1 独立，则对 B 有

$$\because \sqrt{\frac{n}{2}}(\frac{s_{ii}}{\sigma_{ii}} - 1) \xrightarrow{n\to\infty} N(0,1) \qquad (i = 1, 2, ..., p), \tag{2.11}$$

$$\begin{aligned} \therefore E(B) =& E(Z_p^T D_1 Z_p), \\ =& E[E(Z_p^T D_1 Z_p | Z_p)] = 0. \end{aligned} \tag{2.12}$$

$$\begin{aligned} Var(B) &= Var(Z_p^T D_1 Z_p), \\ &= E(Z_p^T D_1 Z_p)^2 - [E(Z_p^T D_1 Z_p)]^2, \\ &= E(Z_p^T D_1 Z_p)^2. \end{aligned} \tag{2.13}$$

应用引理 2 (Schott[11,p.394 Theorem 9.21] [16])和引理 4 的结论，可以进一步将 $Var(B)$ 转化为如下形式

$$\begin{aligned} Var(B) &= E(Z_p^T D_1 Z_p Z_p^T D_1 Z_p), \\ &= tr(D_1 \mathcal{R}_p) tr(D_2 \mathcal{R}_p) + 2tr(D_1 \mathcal{R}_p D_2 \mathcal{R}_p), \\ &= 2 \sum_{i,j=1}^{p} \rho_{i,j}^2 (\frac{s_{ii}}{\sigma_{ii}} - 1)(\frac{s_{jj}}{\sigma_{jj}} - 1) + [\sum_{i=1}^{p} (\frac{s_{ii}}{\sigma_{ii}} - 1)]^2, \\ &= 3 \sum_{i=1}^{p} (\frac{s_{ii}}{\sigma_{ii}} - 1)^2 + \sum_{i \neq j} (1 + 2\rho_{i,j}^2)(\frac{s_{ii}}{\sigma_{ii}} - 1)(\frac{s_{jj}}{\sigma_{jj}} - 1), \\ &= 3Var(\frac{s_{ii}}{\sigma_{ii}} - 1) + \sum_{i \neq j} (1 + 2\rho_{i,j}^2)(\frac{s_{ii}}{\sigma_{ii}} - 1)(\frac{s_{jj}}{\sigma_{jj}} - 1), \\ &= \frac{6p}{n} + \sum_{i \neq j} (1 + 2\rho_{i,j}^2)(\frac{s_{ii}}{\sigma_{ii}} - 1)(\frac{s_{jj}}{\sigma_{jj}} - 1). \end{aligned} \tag{2.14}$$

而根据 Anderson[2,p.39] [7]，可得出如下结论

$$cov(\frac{s_{ii}}{\sigma_{ii}} - 1, \frac{s_{jj}}{\sigma_{jj}} - 1) = cov(\frac{s_{ii}}{\sigma_{ii}}, \frac{s_{jj}}{\sigma_{jj}}),$$

$$= \frac{1}{\sigma_{ii}\sigma_{jj}} cov(\sigma_{ii}, \sigma_{jj}) = \frac{2}{n}\rho_{i,j}^2. \tag{2.15}$$

所以 B 的方差可转化为如下形式，进而可计算出 $\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}$ 的方差为：

$$Var(B) = \frac{6p}{n} + \frac{2}{n}\sum_{i\neq j}(1+2\rho_{i,j}^2)\rho_{i,j}^2. \tag{2.16}$$

$$\begin{aligned} Var(\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}) &= Var(\frac{\sqrt{p}}{\sqrt{2tr\mathcal{R}_p^2}}\frac{B}{\sqrt{p}}), \\ &= \frac{p}{2tr\mathcal{R}_p^2} + \frac{Var(B)}{p}, \\ &= \frac{p}{2tr\mathcal{R}_p^2}[\frac{1}{p}(\frac{6p}{n} + \frac{2}{n}\sum_{i\neq j}(1+2\rho_{i,j}^2)\rho_{i,j}^2)], \\ &= \frac{p}{2tr\mathcal{R}_p^2}(\frac{6}{n} + \frac{2}{np}(tr\mathcal{R}_p^2 - p) + \frac{4}{np}\sum_{i\neq j}\rho_{i,j}^4), \\ &= \frac{p}{2tr\mathcal{R}_p^2}(\frac{4}{n} + \frac{2}{np}tr\mathcal{R}_p^2 + \frac{4}{np}\sum_{i\neq j}\rho_{i,j}^4). \end{aligned} \tag{2.17}$$

又因为

$$\begin{aligned} \sum_{i\neq j}\rho_{i,j}^4 &\leq \sum_{i=1}^{p}\sum_{j=1}^{p}\rho_{i,j}^4 \leq \sum_{i=1}^{p}(\sum_{j=1}^{p}\rho_{i,j}^2)^2, \\ &\leq \sum_{k=1}^{p}\sum_{i=1}^{p}(\sum_{j=1}^{p}\rho_{i,j}\rho_{k,j})^2 = tr\mathcal{R}_p^4. \end{aligned} \tag{2.18}$$

所以得到 $\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}$ 的方差的上限为

$$Var(\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}) \leq \frac{2}{n}\frac{1}{\frac{tr\mathcal{R}_p^2}{p}} + \frac{1}{n} + \frac{2}{n}\frac{1}{\frac{tr\mathcal{R}_p^2}{p}}\frac{1}{\frac{tr\mathcal{R}_p^4}{p}}. \tag{2.19}$$

根据预备知识中所写的假设 1，可知 $1 \leq \lim_{p\to\infty}\frac{tr\mathcal{R}_p^i}{p} < \infty \quad (i = 1,2,3,4)$，代入上式得出如下结论

$$Var(\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}) \leq \frac{2}{n} + \frac{1}{n} + \frac{2}{n} = \frac{5}{n}. \tag{2.20}$$

由 *Chebyshev* 不等式可知

$$P(|\frac{B}{\sqrt{2tr\mathcal{R}_p^2}} - E(\frac{B}{\sqrt{2tr\mathcal{R}_p^2}})| \geq \epsilon) \leq Var(\frac{B}{\sqrt{2tr\mathcal{R}_p^2}})/\epsilon^2. \tag{2.21}$$

$$\Longrightarrow P(|\frac{B}{\sqrt{2tr\mathcal{R}_p^2}}| \geq \epsilon) \leq \frac{5}{n\epsilon^2}. \tag{2.22}$$

所以得出 $\frac{B}{\sqrt{2tr\mathcal{R}_p^2}} = \mathcal{O}(\frac{1}{\sqrt{n}})$。

而对 C 可作如下变换

$$\begin{aligned} C &= Z_p^T D_2 Z_p, \\ &= \frac{n_1 n_2}{n}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T D_\sigma^{-1/2} D_2 D_\sigma^{-1/2}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}), \\ &= \frac{n_1 n_2}{n}[\sum_{i=1}^{p} \frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}}(\frac{\sigma_{ii}}{s_{ii}} + \frac{s_{ii}}{\sigma_{ii}} - 2)], \\ &= [\sum_{i=1}^{p} \frac{(\bar{X}_i - \bar{Y}_i)^2}{s_{ii}\frac{n}{n_1 n_2}} + \sum_{i=1}^{p} \frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}}\frac{s_{ii}}{\sigma_{ii}} - 2\sum_{i=1}^{p} \frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}}]. \end{aligned} \tag{2.23}$$

因为 $\sqrt{n-1}\frac{\bar{X}_i - \bar{Y}_i}{s_{ii}\frac{n}{n_1 n_2}} \overset{H_0}{\sim} t(n-1)$，所以有

$$E(\sqrt{n-1}\frac{\bar{X}_i - \bar{Y}_i}{s_{ii}\frac{n}{n_1 n_2}}) = 0. \tag{2.24}$$

$$Var(\sqrt{n-1}\frac{\bar{X}_i - \bar{Y}_i}{s_{ii}\frac{n}{n_1 n_2}}) = \frac{n-1}{n-3}. \tag{2.25}$$

结合上述结论，可以求出 C 的均值为

$$\begin{aligned} E(C) &= [\sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{s_{ii}\frac{n}{n_1 n_2}}) + \sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}})E(\frac{s_{ii}}{\sigma_{ii}}) - 2\sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}})], \\ &= [\sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{s_{ii}\frac{n}{n_1 n_2}}) + \sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}}) - 2\sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}})], \\ &= [\sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{s_{ii}\frac{n}{n_1 n_2}}) - \sum_{i=1}^{p} E(\frac{(\bar{X}_i - \bar{Y}_i)^2}{\sigma_{ii}\frac{n}{n_1 n_2}})], \end{aligned}$$

$$= \frac{n-1}{n-3}p - p = \frac{2p}{n-3}. \tag{2.26}$$

又因为 $\frac{\sigma_{ii}}{s_{ii}} + \frac{s_{ii}}{\sigma_{ii}} - 2 \geq 0 \quad (i = 1, 2, ..., p)$，且易知 D_2, C 均为非负定矩阵，所以有 $E(|C|) = E(C) = \frac{2p}{n-3}$，故可得如下结果

$$\begin{aligned} E(|\frac{C}{\sqrt{2tr\mathcal{R}_p^2}}|) &= \frac{1}{\sqrt{2tr\mathcal{R}_p^2}} E(|C|), \\ &= \frac{2p}{n-3} \frac{1}{\sqrt{2tr\mathcal{R}_p^2}}, \\ &= \frac{\sqrt{2}p^{1/2}}{n-3} (\frac{tr\mathcal{R}_p^2}{p})^{-1/2}. \end{aligned} \tag{2.27}$$

因为 $|\frac{C}{\sqrt{2tr\mathcal{R}_p^2}}|$ 是非负随机变量，满足 $Markov$ 不等式的应用条件，则可以得出对任意 $\epsilon > 0$ 有

$$\begin{aligned} P(|\frac{C}{\sqrt{2tr\mathcal{R}_p^2}}| \geq \epsilon) &\leq \frac{1}{\epsilon} E(|\frac{C}{\sqrt{2tr\mathcal{R}_p^2}}|), \\ &= \frac{1}{\epsilon} \frac{\sqrt{2}p^{1/2}}{n-3} (\frac{tr\mathcal{R}_p^2}{p})^{-1/2}, \\ &\leq \frac{1}{\epsilon} \frac{\sqrt{2p}}{n-3}, \\ &= \frac{1}{\epsilon} \mathcal{O}_p(\frac{p^{1/2}}{n-3}). \end{aligned} \tag{2.28}$$

可以得出 $\frac{C}{\sqrt{2tr\mathcal{R}_p^2}} = \mathcal{O}_p(\frac{p^{1/2}}{n-3})$，故

$$\frac{T_{new}}{\sqrt{2tr\mathcal{R}_p^2}} = \frac{A}{\sqrt{2tr\mathcal{R}_p^2}} + \mathcal{O}_p\{max(\frac{1}{\sqrt{n}}, \frac{p^{1/2}}{n-3})\}. \tag{2.29}$$

而

$$A = Z_p^T Z_p = \frac{n_1 n_2}{n} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^T D_\sigma^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) = \widetilde{T}_{new} \tag{2.30}$$

所以在 H_0 假设成立的条件下，$T_{new} \xrightarrow{P} \widetilde{T}_{new}$，即得出 H_0 假设成立的条件下检验统计量 T_{new} 的理论分布。

应用引理 5 的结论结合 A 的定义和 $E(R) = \beta d, Var(R) = 2\beta^2 d$ 可得出检验统计量 $\widetilde{T}_{new}$ 的渐进分布的分布参数为

$$A = \frac{n_1 n_2}{n} || \frac{\bar{\boldsymbol{X}}}{D_\sigma^{1/2}} - \frac{\bar{\boldsymbol{Y}}}{D_\sigma^{1/2}} ||^2 \xrightarrow{L} R \overset{d}{=} \beta \chi_d^2. \Longrightarrow \begin{cases} tr(D_\sigma^{-1}\boldsymbol{\Sigma}) = \beta d. \\ 2tr(D_\sigma^{-1}\boldsymbol{\Sigma})^2 = 2\beta^2 d. \end{cases} \tag{2.31}$$

$$\Longrightarrow \begin{cases} \beta = \dfrac{tr(D_\sigma^{-1}\boldsymbol{\Sigma})^2}{tr(D_\sigma^{-1}\boldsymbol{\Sigma})}, \\ d = \dfrac{tr^2(D_\sigma^{-1}\boldsymbol{\Sigma})}{tr(D_\sigma^{-1}\boldsymbol{\Sigma})^2}. \end{cases} \tag{2.32}$$

而已知 $T_{new} \xrightarrow{P} \widetilde{T}_{new}$，故 T_{new} 的渐进分布的分布参数计算方法同上。而以上为理论上的结果，实际代码编写中将 $D_\sigma^{-1}\boldsymbol{\Sigma}$ 视为一个整体，这里用 M 表示，再应用推论 2 的结论来计算 $\hat{\beta}$ 和 $\hat{d}$ 的值，即

$$\begin{cases} \hat{\beta} = \dfrac{\widehat{tr(M^2)}}{tr(\hat{M})} = \dfrac{\dfrac{(n-2)^2}{(n-3)n}[tr(\hat{M}^2) - \dfrac{tr^2(\hat{M})}{n-2}]}{tr(\hat{M})}, \\ \hat{d} = \dfrac{\widehat{tr^2(M)}}{\widehat{tr(M^2)}} = \dfrac{\dfrac{(n-2)(n-1)}{(n-3)n}[tr^2(\hat{M}) - \dfrac{2}{n-1}tr(\hat{M}^2)]}{\dfrac{(n-2)^2}{(n-3)n}[tr(\hat{M}^2) - \dfrac{tr^2(\hat{M})}{n-2}]}. \end{cases} \tag{2.33}$$

2.2.3 新统计方法的假设检验和功效函数的计算

假设检验是强有力的统计检验工具，用样本信息去判断命题正确与否的思想能有效避免主观印象对结果的影响，尽可能地保证判断结果的客观性和公正性，为正确决策提供科学依据。基于我们的统计量，提出的假设检验如下：

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \qquad vs \qquad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

由上述证明可知检验统计量 T_{new} 的理论渐近分布为卡方分布，所以可得出拒绝域为：$T_{new} \leq \hat{\beta}\chi^2_{\hat{d}}$，由此可以计算出来功效函数为：

$$Power = P(\frac{T_{new}}{\dfrac{(n-2)^2}{(n-3)n}[tr(\hat{M}^2) - \dfrac{tr^2(\hat{M})}{n-2}] \Big/ tr(\hat{M})} \geq \chi^2_{\hat{d}}(\alpha)). \tag{2.34}$$

也就是说在原假设 H_0 为假，显著性水平 $\alpha = 0.05$ 时，一个服从自由度为 $\hat{d}$ 的卡方分布的随机变量，其小于检验统计量 T_{new} 除以 $\hat{\beta}$ 的值的概率即为拒绝原假设的概率，也就是检验功效。而在原假设 H_0 为真时，自由度为 $\hat{d}$ 的卡方分布随机变量小于检验统计量 T_{new} 除以 $\hat{\beta}$ 的值的概率，即为落入拒绝域的概率，也就是犯第一类错误的概率。所以在进行统计模拟的程序编写时，犯第一类错误的概率和检验功效的代码编写大致相同，只用在运行时修改下控制原假设 H_0 真假的参数即可。

第三章 统计模拟

3.1 统计模拟

这一节将对理论部分进行模拟，观察模拟的结果，并通过分析运算结果来对我们提出的检验方法 L^2new 与前文提到的 L^2-norm 方法进行比较。为了更详细地比较两种检验方法，以下将分成协方差矩阵不同维度上的方差相同和不同两种情况来进行数据模拟。

我们将采用 Zhang, Guo, Zhou, Cheng(2020) [18] 的论文中所提到的模型，分别产生 p 维多元正态分布数据、t 分布数据和卡方分布数据三种不同类型的模拟数据。基于以上数据，首先在协方差矩阵不同维度上的方差相同的情况下产生样本数据，再进行两样本均值假设检验，观察运算结果并进行分析比较，然后再在协方差矩阵不同维度上的方差不同的情况下产生样本数据，再进行两样本均值假设检验，观察运算结果并进行分析比较。

3.1.1 协方差矩阵不同维度上的方差相同情况下的统计模拟

这里我们采用 Zhang, Guo, Zhou, Cheng(2020) [18] 的论文中所提到的模型来产生模拟数据，即使用 $\mathbf{y}_{ij}=\boldsymbol{\mu}_i+\boldsymbol{\Sigma}^{1/2}\mathbf{z}_{ij}, j=1,...,n_i; i=1,2$ 生成两个样本，其中 $\boldsymbol{\mu}_2=\boldsymbol{\mu}_1+\delta\mathbf{h}, \boldsymbol{\Sigma}=\sigma^2[(1-\rho)\mathbf{I}_p+\rho\mathbf{J}_p]$, $\mathbf{I}_p$ 是 p 维单位矩阵， $\mathbf{J}_p$ 是 p 维的元素全为 1 的矩阵， $\mathbf{z}_{ij}=(z_{ij1},...,z_{ijp})^T, j=1,...,n_i; i=1,2$ 是来自以下三个模型的独立随机变量：

$$Model\ 1.\quad z_{ijt}, t=1,...,p, \overset{iid}{\sim} N(0,1). \tag{3.1}$$

$$Model\ 2.\quad z_{ijt}=w_{ijt}/\sqrt{2}, t=1,...,p,\quad w_{ijt}\overset{iid}{\sim} t_4. \tag{3.2}$$

$$Model\ 3.\quad z_{ijt}=(w_{ijt}-1)/\sqrt{2}, t=1,...,p,\quad w_{ijt}\overset{iid}{\sim}\chi^2_1. \tag{3.3}$$

参数 $\delta,\mathbf{h},\rho$ 控制着向量差 $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$ 的大小，其中 $\boldsymbol{\mu}_1=\mathbf{0}, \mathbf{h}=\boldsymbol{\mu}/\|\boldsymbol{\mu}\|, \sigma_2=1, \boldsymbol{\mu}=(1,...,p)^T$ 。基于 $Model\ 1$ 产生的是高维正态分布数据，基于 $Model\ 2$ 产生的是 t 分布数据，基于 $Model\ 3$ 产生的是卡方分布数据。

为了比较两种不同检验方法的检验效果，在数据模拟中将对维数 p、样本量 (n_1,n_2) 分别考虑三种情况：$p=50,500,1000$；$(n_1,n_2)=(30,50)$，$(120,200)$，$(240,400)$，即每个 (n_1,n_2) 会对应维数 p 的三种情况分别生成数据，并且改变 ρ 的值分别为 $0.1,0.5,0.9$，在经验水平基于 $\alpha=0.05$ 的时候，重复 10000 次模拟运算，得出结果，再加以分析来比较两种检验方法的功效和犯第一类错误的概率，整理

模拟结果如表 1、2 所示，图 1、2、3 为 ρ 的值分别为 $0.1, 0.5, 0.9$ 时两种方法的犯第一类错误的概率对比图，为更好对比我们从所有数据中分情况挑选了九组数据来画散点图，大致分布在 0.05 左右的表示该方法第一类错误控制得很好。

上述模型中的协方差矩阵在不同维度上的方差相同，即协方差矩阵内部结构为 $\sigma_{11} = \sigma_{22} = ... = \sigma_{pp}$ 的情况，基于这种模型产生的数据在实际生活中并不是非常常见，绝大多数情况下待检验样本的协方差矩阵在不同维度上的方差并不会太过整齐，总会存在或大或小的差异。

我们的检验方法正是为了处理上述情况而设计的，所以下面我们将基于样本数据多个分量的方差不等的情况生成模拟数据，即协方差矩阵中 $\sigma_{ii} \neq \sigma_{jj} \neq ... \neq \sigma_{kk}$, $(1 \leq i, j, ..., k \leq p)$ 的情况，分别将两种检验方法应用于该模拟数据集，再分析假设检验结果来进一步比较检验方法 L^2new 与前文提到的 $L^2 - norm$ 检验方法的检验效果。

3.1.2 协方差矩阵不同维度上的方差不同情况下的统计模拟

在上述基础上，我们仍然采用初始模型中基于 $Model$ 1、$Model$ 2、$Model$ 3 产生的多元正态分布数据、t 分布数据和卡方分布数据，均值向量、数据结构、调节参数的设置同上，只对样本协方差矩阵做出改变，即改变 $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}^{1/2}\mathbf{z}_{ij}, j = 1, ..., n_i; i = 1, 2$ 中的 $\boldsymbol{\Sigma}$，让 $\boldsymbol{\Sigma}$ 同时左乘右乘元素为 0.1 到 5 上的均匀分布随机数的对角阵，以改变协方差矩阵不同维度上的方差：

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= \begin{pmatrix} \mathbf{U}_{11}(0.1,5) & & & \\ & \mathbf{U}_{12}(0.1,5) & & \\ & & \vdots & \\ & & & \mathbf{U}_{1p}(0.1,5) \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} \mathbf{U}_{21}(0.1,5) & & & \\ & \mathbf{U}_{22}(0.1,5) & & \\ & & \vdots & \\ & & & \mathbf{U}_{2p}(0.1,5) \end{pmatrix}, \\ &= \begin{pmatrix} \mathbf{U}_{11}\mathbf{U}_{21}\sigma_{11} & \mathbf{U}_{11}\mathbf{U}_{22}\sigma_{12} & \mathbf{U}_{11}\mathbf{U}_{23}\sigma_{13} & \cdots & \mathbf{U}_{11}\mathbf{U}_{2p}\sigma_{1p} \\ \mathbf{U}_{12}\mathbf{U}_{21}\sigma_{21} & \mathbf{U}_{12}\mathbf{U}_{22}\sigma_{22} & \mathbf{U}_{12}\mathbf{U}_{23}\sigma_{23} & \cdots & \mathbf{U}_{12}\mathbf{U}_{2p}\sigma_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{U}_{1p}\mathbf{U}_{21}\sigma_{p1} & \mathbf{U}_{1p}\mathbf{U}_{22}\sigma_{p2} & \mathbf{U}_{1p}\mathbf{U}_{23}\sigma_{p3} & \cdots & \mathbf{U}_{1p}\mathbf{U}_{2p}\sigma_{pp} \end{pmatrix}. \end{aligned} \tag{3.4}$$

其中 $U(0.1, 5)$ 为 0.1 到 5 上的均匀分布随机数，为了比较检验方法的效果，对维数 p、样本量 (n_1, n_2) 分别考虑三种情况：$p = 50, 500, 1000$; $(n_1, n_2) = (30, 50)$, $(120, 200), (240, 400)$ ，并改变 ρ 的值分别为 $0.1, 0.5, 0.9$ ，在经验水平基于 $\alpha = 0.05$ 的时候，重复 10000 次模拟运算，得出结果并进行分析以比较两种检验方法的功

效和犯第一类错误的概率，整理模拟结果如表 3、4 所示，图 4、5、6 为 ρ 的值分别为 $0.1, 0.5, 0.9$ 时两种方法的犯第一类错误的概率对比图，为更好对比我们从所有数据中分情况挑选了九组数据来画散点图，大致分布在 0.05 左右的表示该方法第一类错误控制得很好。

由于 $U(0.1, 5)$ 为 0.1 到 5 上的均匀分布随机数是随机产生的，所以相同的概率极低，近乎为零，能很好地保证改变后的 Σ 矩阵不同维度上的方差不完全相同，所以能很好地模拟多个分量上的方差不同的情况下的数据，且数据产生模型仍采用之前的三种不同分布类型的，能进一步检验我们的检验方法 L^2new 对于多种数据特点的适应性的强弱。

表 1: 统计量的检验功效对比

				ρ=0.1		ρ=0.5		ρ=0.9	
Model	p	(n_1, n_2)	δ	L^2norm	L^2new	L^2norm	L^2new	L^2norm	L^2new
1	50	(30,50)	1.0	36.89	40.09	13.5	14.3	10.05	10.19
		(120,200)	0.6	50.63	51.59	17.19	17.5	11.1	11.11
		(240,400)	0.4	45.04	45.48	15.4	15.53	10.94	10.96
	500	(30,50)	3.2	52.75	57.44	14.62	15.25	9.96	10.02
		(120,200)	1.5	47.51	48.66	12.46	12.62	9.26	9.27
		(240,400)	1.2	58.28	58.82	14.37	14.43	9.53	9.54
	1000	(30,50)	4.5	54.06	58.42	13.9	14.6	10.11	10.21
		(120,200)	2.0	44.28	45.3	11.61	11.79	8.19	8.22
		(240,400)	1.5	47.79	48.3	12.41	12.53	9.15	9.15
2	50	(30,50)	1.0	35.84	41.65	13.62	14.32	9.98	10.04
		(120,200)	0.6	51.48	53.54	16.92	17.09	11.03	11.06
		(240,400)	0.4	45.43	46.62	15.39	15.46	9.7	9.72
	500	(30,50)	3.2	52.62	57.3	14.5	15.09	9.77	9.89
		(120,200)	1.5	47.02	47.99	11.87	11.98	9.42	9.45
		(240,400)	1.2	57.03	57.37	14.61	14.62	9.76	9.76
	1000	(30,50)	4.5	52.65	56.58	13.87	14.55	10.11	10.19
		(120,200)	2.0	43.84	44.41	11.91	12	8.39	8.4
		(240,400)	1.5	47.91	48.24	12.81	12.91	8.58	8.58
3	50	(30,50)	1.0	35.85	54.79	13.52	15.92	10.05	10.38
		(120,200)	0.6	50.24	59	16.32	17.68	11.28	11.45
		(240,400)	0.4	46.41	52.49	16.45	17.22	10.32	10.4
	500	(30,50)	3.2	52.88	68.92	13.55	15.51	10.03	10.24
		(120,200)	1.5	47.39	56.81	12.48	13.53	8.81	8.85
		(240,400)	1.2	57.13	63.56	14.73	15.54	10.77	10.83
	1000	(30,50)	4.5	53.18	69.65	14.74	16.89	10.43	10.57
		(120,200)	2.0	43.35	52.7	12.01	12.83	8.59	8.61
		(240,400)	1.5	47.57	54.17	12.38	12.95	8.76	8.81

表 2: 统计量的犯第一类错误的概率对比

			ρ=0.1		ρ=0.5		ρ=0.9	
Model	p	(n_1, n_2)	L^2norm	L^2new	L^2norm	L^2new	L^2norm	L^2new
1	50	(30,50)	5.59	6.99	5.55	5.99	5.82	5.87
		(120,200)	5.15	5.53	5.22	5.34	5.49	5.51
		(240,400)	4.87	5.16	5.36	5.43	4.79	4.8
	500	(30,50)	6.4	7.98	5.82	6.11	5.32	5.35
		(120,200)	6.19	6.58	5.64	5.69	5.05	5.06
		(240,400)	5.71	5.8	5.48	5.51	4.87	4.88
	1000	(30,50)	7.04	8.63	5.64	6.04	5.7	5.75
		(120,200)	6.37	6.63	5.2	5.23	5.08	5.09
		(240,400)	6.09	6.16	5.17	5.23	5.28	5.28
2	50	(30,50)	5.01	6.13	5.99	6.24	6.16	6.26
		(120,200)	5.16	5.19	5.18	5.25	4.78	4.81
		(240,400)	4.74	4.66	5.13	5.18	4.78	4.81
	500	(30,50)	6.16	7.39	6.06	6.42	5.39	5.46
		(120,200)	6.17	6.36	5.25	5.31	5.36	5.38
		(240,400)	5.81	5.99	5.7	5.72	5.07	5.07
	1000	(30,50)	6.65	7.76	5.76	6.1	5.62	5.67
		(120,200)	6.37	6.45	5.62	5.67	5.05	5.05
		(240,400)	6.14	6.25	5.11	5.13	4.87	4.88
3	50	(30,50)	5	6.46	5.82	6.03	5.35	5.38
		(120,200)	5.28	5.87	5.47	5.49	4.84	4.85
		(240,400)	4.99	5.39	4.99	5.06	4.8	4.81
	500	(30,50)	6.48	8.66	5.79	6.02	5.96	6.01
		(120,200)	6.38	7.01	5.56	5.63	5.18	5.23
		(240,400)	6.17	6.29	5.3	5.31	4.92	4.95
	1000	(30,50)	6.57	9.04	5.89	6.29	5.43	5.47
		(120,200)	6.2	6.98	5.32	5.29	5.24	5.28
		(240,400)	6.07	6.56	5.39	5.49	5.38	5.37

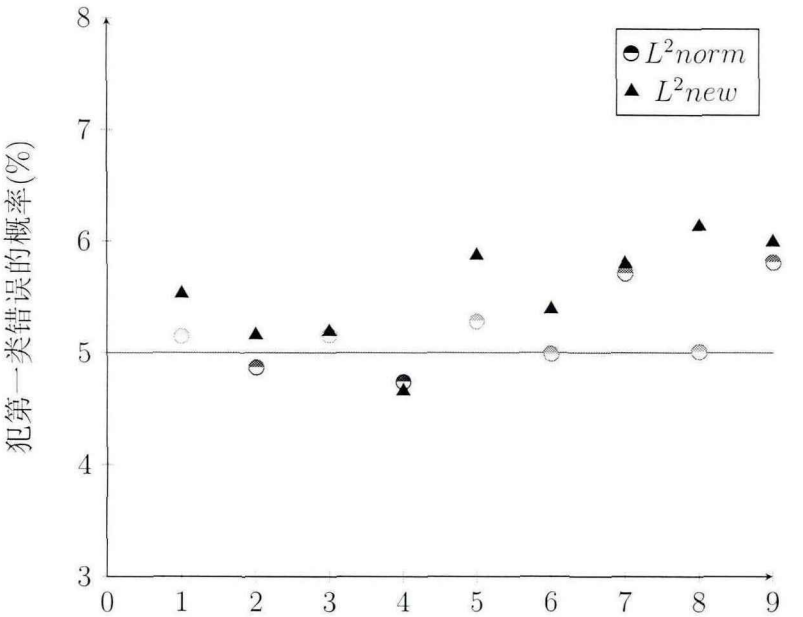

图 1: ρ=0.1时犯第一类错误的概率对比

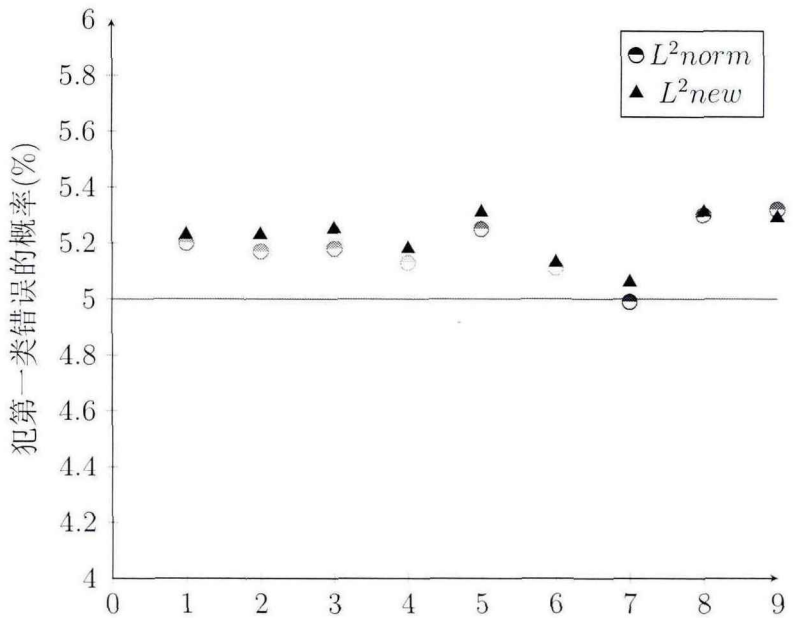

图 2: ρ=0.5时犯第一类错误的概率对比

表 3: Σ_1 结构下的统计量的检验功效对比

				ρ=0.1		ρ=0.5		ρ=0.9	
Model	p	(n_1, n_2)	δ	L^2norm	L^2new	L^2norm	L^2new	L^2norm	L^2new
1	50	(30,50)	1.0	8.02	73.74	6.19	18.72	5.82	9.98
		(120,200)	0.6	8.54	73.49	6.31	92.87	5.45	9.26
		(240,400)	0.4	8.1	32.8	6.46	11.86	5.59	11.87
	500	(30,50)	3.2	10.51	100	6.73	28.48	6.13	9.01
		(120,200)	1.5	9.13	88.82	6.19	11.2	5.5	8.85
		(240,400)	1.2	10.36	99.68	6.74	20.42	5.48	8.84
	1000	(30,50)	4.5	10.61	100	6.5	18.69	5.68	12.84
		(120,200)	2.0	9.35	99.96	5.8	13.85	5.4	8.82
		(240,400)	1.5	9.72	100	5.93	16.04	5.19	10.94
2	50	(30,50)	1.0	7.62	72.61	6.59	11.47	5.9	14.64
		(120,200)	0.6	8.36	33.68	6.09	17.54	5.81	63.94
		(240,400)	0.4	8.75	57.32	6.26	76.41	6	9.98
	500	(30,50)	3.2	9.37	99.4	7.16	20.28	6.3	10.56
		(120,200)	1.5	9.51	100	6.08	13.49	5.35	10.17
		(240,400)	1.2	10.21	100	6.32	16.31	5.55	11.14
	1000	(30,50)	4.5	10.87	100	6.31	13.95	6.49	11.12
		(120,200)	2.0	9.47	97.76	5.87	15.48	5.43	10.5
		(240,400)	1.5	9.6	99.99	5.76	14.65	5.02	9.11
3	50	(30,50)	1.0	5.66	98.29	6.1	87.89	6.13	11.63
		(120,200)	0.6	7.94	78.37	5.99	24.37	6	8.41
		(240,400)	0.4	7.86	21.23	6.29	14.53	5.66	17.6
	500	(30,50)	3.2	9.43	100	6.91	27.05	6.09	12.26
		(120,200)	1.5	9.43	99.82	5.55	21.47	5.51	8.68
		(240,400)	1.2	10.27	100	5.97	22.77	5.47	11.59
	1000	(30,50)	4.5	10.53	99.84	6.59	16.51	6.12	15.34
		(120,200)	2.0	9.92	99.29	5.89	13.43	5.91	9.26
		(240,400)	1.5	9.21	99.48	5.89	20.18	5.09	9.3

表 4: Σ_1 结构下的统计量的犯第一类错误的概率对比

			ρ=0.1		ρ=0.5		ρ=0.9	
Model	p	(n_1, n_2)	L^2norm	L^2new	L^2norm	L^2new	L^2norm	L^2new
1	50	(30,50)	5.2	6.79	6.06	6.43	5.87	5.95
		(120,200)	5.06	5.3	5.05	5.25	5.23	5.1
		(240,400)	5.06	5.39	5.12	5.41	4.99	4.97
	500	(30,50)	6.38	8.42	6.1	6.54	5.72	5.81
		(120,200)	5.7	6.25	5.75	5.87	5.28	5.28
		(240,400)	5.59	6.14	5.43	5.41	4.9	4.89
	1000	(30,50)	6.53	8.5	5.75	5.98	5.75	5.79
		(120,200)	6	6.47	5.28	5.31	5.14	5.15
		(240,400)	6.33	6.5	5.29	5.42	5.14	5.13
2	50	(30,50)	4.8	6.05	6.01	6.59	5.29	5.23
		(120,200)	5.25	5.49	5.21	5.31	4.95	4.94
		(240,400)	5.45	5.39	5.28	5.29	4.94	4.93
	500	(30,50)	6.11	7.46	5.79	6.07	5.59	5.65
		(120,200)	5.97	6.32	5.49	5.56	4.91	4.99
		(240,400)	5.95	6.13	5.51	5.47	4.88	4.87
	1000	(30,50)	6.67	7.83	5.57	5.87	5.96	6.04
		(120,200)	6.38	6.6	5.38	5.44	5.23	5.23
		(240,400)	6.19	6.21	5.35	5.42	5.1	5.07
3	50	(30,50)	5.19	6.51	5.45	5.73	5.79	5.8
		(120,200)	5.02	5.34	5.45	5.43	5.08	5.16
		(240,400)	5.2	5.12	5.28	5.43	4.77	4.86
	500	(30,50)	6.01	7.7	5.17	5.54	5.21	5.27
		(120,200)	5.72	6.98	5.61	5.88	5.05	5
		(240,400)	5.91	6.58	5.63	5.64	5.02	5.03
	1000	(30,50)	6.33	7.56	5.96	6.3	5.56	5.68
		(120,200)	6.31	6.99	5.66	5.69	4.98	4.99
		(240,400)	6.26	6.87	5.24	5.34	5.07	5.07

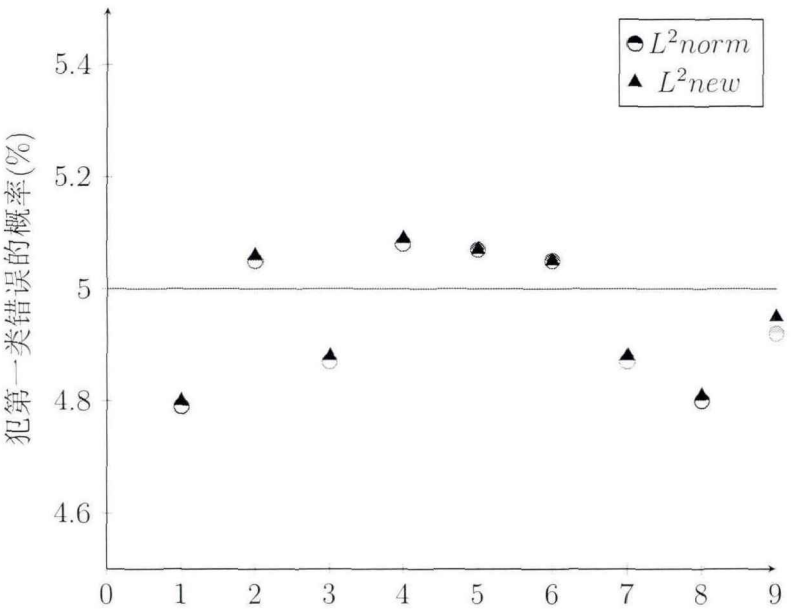

图 3: $\rho = 0.9$ 时犯第一类错误的概率对比

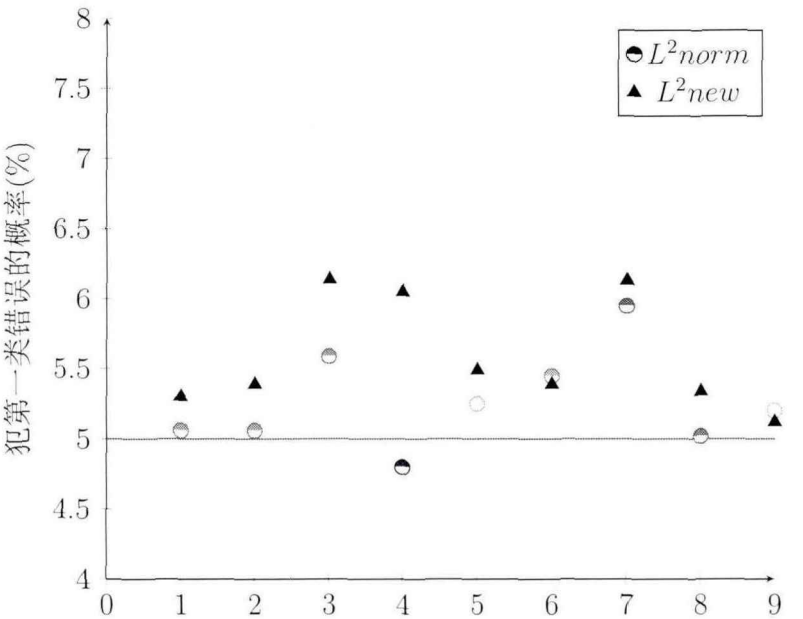

图 4: Σ_1 结构下 $\rho = 0.1$ 时犯第一类错误的概率对比

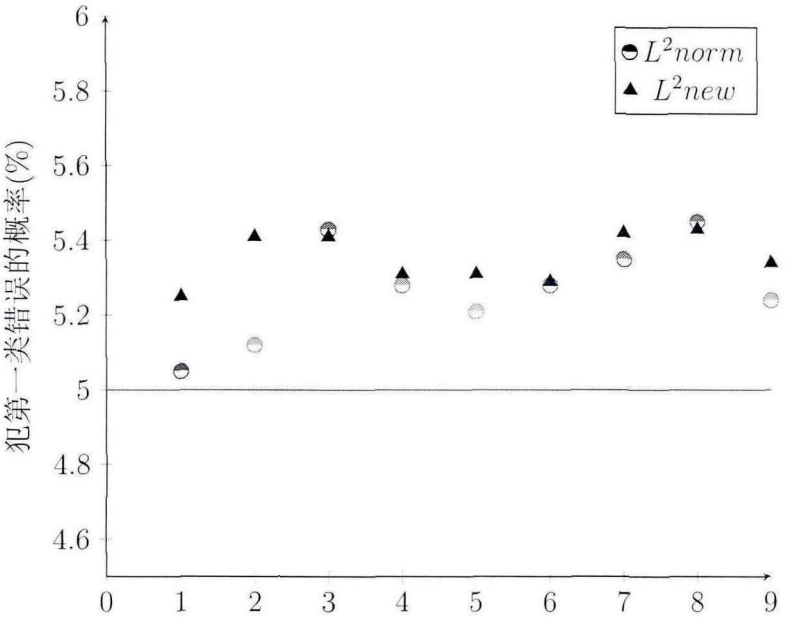

图 5: $\mathbf{\Sigma}_1$ 结构下 $\rho = 0.5$ 时犯第一类错误的概率对比

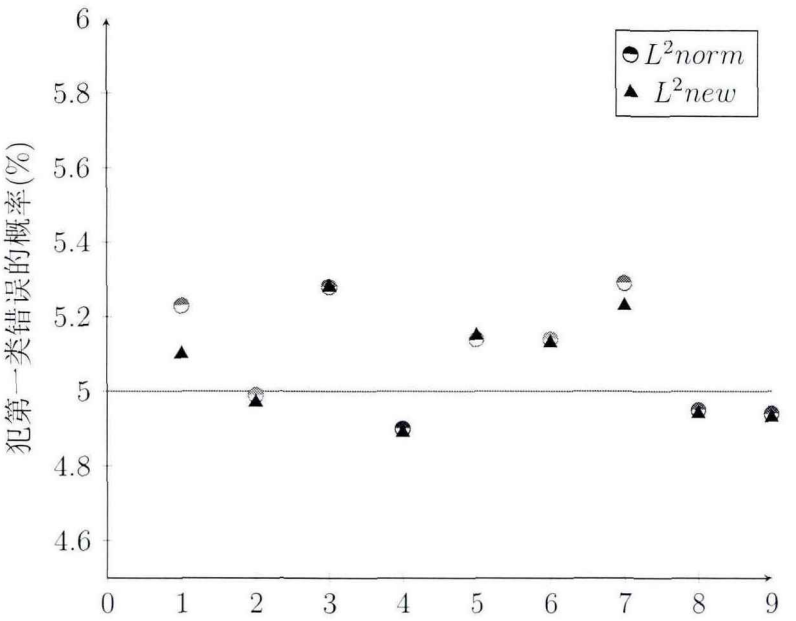

图 6: $\mathbf{\Sigma}_1$ 结构下 $\rho = 0.9$ 时犯第一类错误的概率对比

3.2 结果分析

在这一节我们将对以上的数据从多个角度进行具体的分析。

在数据模拟阶段我们采用了三种模型，分别有正态分布数据、t 分布数据和卡方分布数据，结合上述表格和对比图片来看，可以得出检验方法 L^2new 对于这三种数据模型均有较好的检验效果，这说明检验方法 L^2new 能较好地适应多种数据特点。且随着维数 p 的增加，即使维数 p 远大于样本量 (n_1, n_2)，检验方法 L^2new 也能有较好的检验功效和较低的犯第一类错误的概率，所以在解决高维数据的问题时，可以采取我们的方法进行计算。

从表 2 和图 1、2、3 可以看出在协方差矩阵不同维度上的方差相同的情况下检验方法 L^2new 和 $L^2 - norm$ 方法犯第一类错误的概率近似相等，但根据表 1 可知此时检验方法 L^2new 的功效略高于检验方法 $L^2 - norm$。而在多个分量上的方差不同的情况下，随机向量内部相关性较复杂时，由表 3 可知检验方法 L^2new 的检验功效明显高于检验方法 $L^2 - norm$，且二者犯第一类错误的概率并没有明显的差距，从表 4 和图 4、5、6 中可以直观看出大致都分布在 0.05 左右，均控制在较好范围内。

基于以上的分析可以看出，我们提出的检验方法 L^2new 可以较好地适应不同数据类型的情况，而且在针对高维数据时有着不错的检验效果，且对于多个分量方差不等的数据有着较好的检验效果，所以对于处理高维的内部相关性复杂的数据可以采用我们的检验方法，其具有较强的适用性。

但该方法仍有一些问题尚未解决: 该方法是基于两样本数据的协方差矩阵相同的情况下得出的，虽然解决了协方差矩阵不同维度上的方差不同的情况下的两样本均值检验问题，但是现实生活中我们会经常遇到两样本数据的协方差矩阵不同的情况，这个时候我们的检验方法就不再适用了。所以后续的研究方向将致力于解决该问题，也即解决协方差矩阵不同时的两样本数据均值检验问题。

第四章　实际数据分析

4.1　数据来源以及背景

这一部分是基于结肠数据集进行的实际数据分析，结肠癌是人类最常见的恶性肿瘤之一，在人群中的发病率越来越高，对其发病机理的研究有着重要的实际意义。现有的一系列生物学研究已经证实了它的发生和发展是多基因协同作用的结果，比如 CIP2A、EphA2、MDM2 和 CDC25B 等基因经研究证实与结肠癌的发生和发展有密切的关系，所以基因检测在诊断结肠癌方面具有重要作用。我们所用到的数据集是 Alon、Barkai、Gish、Levine、Mack、Notterman 和 Ybarra 于 1999 年所发表的用寡核苷酸芯片检测正常和肿瘤结肠组织基因表达谱的研究中的数据，在 $http://genomics-pubs.princeton.edu/oncology/affydata/index.html$ 网站上可以找到相应数据，这里引用该数据集的目的就是比较正常结肠组织与肿瘤结肠组织在特定基因表达载体上的基因表达是否存在显著差异。

该数据集来自于 62 个数据采集对象，其中 22 个未患结肠肿瘤的人士，40 个结肠肿瘤患者，即正常结肠组织的观测样本有 22 个 ($n_1 = 22$)，肿瘤结肠组织的观测样本有 40 个 ($n_2 = 40$)，基因表达的比较基于从结肠组织上挑选出来的 2000 个基因表达载体 ($p = 2000$)。

4.2　结果分析

基于该数据集，首先对样本进行正态性检验，原假设为两组观测样本的均值相等，利用 R 程序计算两种检验方法的检验 p 值，分析结果见表 4.1。

表 5: 基于结肠数据集进行的实际数据分析

方法	统计量	p值	$\hat{\beta}$	$\hat{d}$
L2N	1.34×10^9	0.53×10^{-3}	5.38×10^7	6.52
T_{new}	8.30×10^4	0.28×10^{-1}	6.89×10^3	4.69

由表 4.1 可知，检验方法 $L^2 - norm$ 和 L^2new 均拒绝原假设，即认为正常结肠组织与肿瘤结肠组织的基因表达存在显著差异，这是正确的。但是 $L^2 - norm$ 方法的检验 p 值与 0.05 的差距过大，不太可靠，相比之下我们的方法可信度较高，所以在这个例子中，使用 L^2new 方法会更适合一些。

第五章　总结与展望

本文研究的是高维数据中两样本总体均值的假设检验问题，因为高维数据不能保证维度远大于样本量，所以经典的检验方法在使用上适用性不是很强，在此基础上本文提出了一种新的检验方法 L^2new 来检验高维两样本均值向量差异的问题，该方法能较好的处理高维数据。并且文中利用计算机生成模拟数据，对检验方法 L^2new 和 L^2-norm 方法进行了模拟运算，并通过模拟的结果进行了检验效果的比较，印证了我们提出的检验方法可以处理多种类型的数据，且有着不错的检验效果。

本文基于高维数据的特点，采用了三种不同的数据类型，从模拟的结果可以看出，检验方法 L^2new 和 L^2-norm 在这三种不同的情况下均有着不错的表现，犯第一类错误的概率基本上都能控制在 0.05 左右，且均有着较高的功效，说明检验方法 L^2new 和 L^2-norm 均能较好地适应多种数据特点。而在处理多个分量方差不等的数据时，检验方法 L^2new 和 L^2-norm 的检验 p 值基本上都能控制在较好范围，但是检验方法 L^2new 比 L^2-norm 有着更高的功效，说明其能更好的处理内部相关性较复杂的数据。

但运用检验方法 L^2new 的前提是两样本数据的协方差矩阵相同，它虽然能解决多个分量上的方差不同的两样本均值检验问题，却不适用于两样本数据协方差矩阵不同的情况，而这种情况在现实生活中是比较常见的。所以我们后续的研究将致力于解决两样本数据协方差矩阵不同的均值检验问题，此时，对于 $tr(\mathbf{\Sigma})$、$tr^2(\mathbf{\Sigma})$、$tr(\mathbf{\Sigma}^2)$ 将要采用新的估计方法来计算。所以在接下来的工作中，我们的研究方向是找出两样本数据协方差矩阵不同时估计 $tr(\mathbf{\Sigma})$、$tr^2(\mathbf{\Sigma})$、$tr(\mathbf{\Sigma}^2)$ 的方法，从而对我们提出的检验方法进一步进行优化，增强它的适用性。

参考文献

[1] Jinting Zhang,Jia Guo,Bu Zhou,Mingyen Cheng. A Simple Two-Sample Test in High Dimensions Based on L2-Norm[J]. Journal of the American Statistical Association,2020,115(530):1011-1027.

[2] Songxi Chen,Yingli Qin. A Two-Sample Test for High-Dimensional Data With Applications to Gene-Set Testing[J]. The Annals of Statistics, 2010, 38(2):808-835.

[3] Zhengbang Li,Fuxiang Liu,Luanjie Zeng,Guoxin Zuo. A stationary bootstrap test about two mean vectors comparison with somewhat dense differences and fewer sample size than dimension[J]. Computational Statistics, 36(2021):941 - 960.

[4] Muni S. Srivastava,Meng Du. A Test for the Mean Vector With Fewer Observations Than the Dimension[J]. Journal of Multivariate Analysis, 2008, 99(3):386 - 402.

[5] B. L. Welch. The Generalization of Student' s Problem When Several Different Population Variances Are Involved[J]. Biometrika, 1947, 34(1-2):28-35.

[6] Gregory Karl Bruce,Carroll Raymond J,Baladandayuthapani Veerabhadran,Lahiri Soumendra N. A two-sample test for equality of means in high dimension[J]. Journal of the American Statistical Association, 2015, 110(510):837-849.

[7] Anderson T. W. An Introduction to Multivariate Statistical Analysis. [J]. Biometrics, 1985, 41(3).

[8] Gongjun Xu, Lifeng Lin, Peng Wei, Wei Pan. An adaptive two-sample test for high-dimensional means[J]. Biometrika, 2016, 103(3):609-624.

[9] Jinting Zhang. Approximate and Asymptotic Distributions of Chi-Squared-Type Mixtures With Applications[J]. Journal of the American statistical association, 2005, 100(469):273-285.

[10] Katayama Shota, Kano Yutaka,Muni S. Srivastava. Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension[J]. Journal of Multivariate Analysis, 2013, 116(2):410-421.

[11] Zhidong Bai, Hewa Saranadasa. Effect of high dimension: By an example of a two sample problem[J]. Statistica Sinica, 1996, 6(2):311-329.

[12] Muni S. Srivastava, Yasunori Fujikoshi. Multivariate analysis of variance with fewer observations than the dimension[J]. Journal of Multivariate Analysis, 2008, 97(9):1927-1940.

[13] Tony Cai, Weidong Liu, Yin Xia. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings[J]. Journal of the Amercian Statstical Association, 2013, 108(501):265-277.

[14] A. P. Dempster. A High Dimensional Two Sample Significance Test[J]. Annals of Mathematical Statistics, 1958, 29(4):995-1010.

[15] E. S. Pearson. Note on an Approximation to the Distribution of Non-Central χ^2 [J]. Biometrika, 1959, 46(3/4):364-364.

[16] James R. Schott. Matrix analysis for statistics[J]. Wiley, New York, 1997.

[17] E. L. Lehmann. Testing statistical hypotheses[J]. Wiley, New York, 1959.

[18] James R. Schott. Some high-dimensional tests for a one-way MANOVA[J]. Journal of Multivariate Analysis, 2006, 98(9).

致　谢

我在华中师范大学三年的研究生生活即将结束，这段时间在华师的学习、生活等各方面的经历，都是我人生中宝贵的财富。桂子山的图书馆让我在期末备考和寻找论文素材时有了一个可以专心思考的地方，桂子山的食堂也让我在陌生的城市吃到了心仪的食物，满山的桂花更是让人心旷神怡，每天都能以阳光的心情去学习和生活，感谢华师给我带来的这一切。而这篇论文的完成更是离不开老师和同学们的无私帮助，在这里我要真诚地向他们表示我的谢意。

首先，我要感谢我的导师李正帮老师。在我创作论文陷入困境时，及时引导我找到新的研究思路，在我遇到难题时，不厌其烦地耐心讲解，每次在我研究方向出现偏差时，都有老师帮助和引导我回到正确的研究方向上的身影，非常感谢在华师的这三年中老师对我的帮助。

然后，我还要感谢我同窗的小伙伴们，她们一直非常热心地解答我在学习上的困惑，并且当我在生活中遇到困难时及时伸出援手，在此感谢她们一直以来对我的支持和鼓励。

最后我还要感谢我的父母，在关注、支持我学业的同时，并不会盲目地给我施加压力，我会继续努力前行，不让他们失望。

胡玉婷

2022 年 4 月