分类号_____ 密级_____

UDC_____ 编号_____

華中師範大學

硕士学位论文

一种新的高维两样本均值检验

学位申请人姓名：胡 浩

申请学位学生类别：全日制硕士

申请学位学科专业：概率论与数理统计

指导教师姓名：李正帮 副教授

硕士学位论文

一种新的高维两样本均值检验

论文作者: 胡 浩

指导教师: 李正帮 副教授

学科专业: 数学

研究方向: 概率论与数理统计

华中师范大学数学与统计学学院

2021 年 5 月

硕士学位论文
MASTER'S THESIS

A new high-dimensional two-sample mean test

A Thesis

Submitted in Partial Fulfillment of the Requirements

For the M.S.Degree in Mathematics

By

Hao Hu

Postgraduate Program

School of Mathematics and Statistics

Central China Normal University

Supervisor : **Zhengbang Li**

Academic Title **Professor** **Signature:** 李正邦

Approved

May 2021

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：胡浩　　日期：2021 年 5 月 31 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：胡浩　　导师签名：李正帮

日期：2021 年 5 月 31 日　　日期：2021 年 6 月 1 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：胡浩　　导师签名：李正帮

日期：2021 年 5 月 31 日　　日期：2021 年 6 月 1 日

摘要

在当今时代, 随着时代和科技的发展, 伴随着基因问题的兴起, 一种新的数据——高维数据, 出现在各个领域, 给传统的统计多元统计理论以及相关的方法带来了非常巨大的挑战. 在高维数据带来的假设检验的问题中, 一方面, 经典的检验统计量在高维数据的情况下毫无意义, 无法被我们使用; 另一方面, 当样本容量 n 远远大于维数 p 时, 经典的极限理论也就不再适用. 创造出适合高维数据的极限理论,改进原本的统计量是如今非常热门的研究领域. 本文从总体均值的假设检验的问题出发,提出了新的思路以及新的方法. 在基于 stationary bootstrap 方法下, 通过对原有统计量的改进, 提出在一定条件下最优的统计量. 与此同时, 通过 R 语言进行数据模拟, 和以往的一些方法进行比较, 结果发现我们提出的统计量在一定的条件下经验功效和经验水平都有着比较好的表现, 此外运用 stationary bootstrap 的方法也避免了理论上计算的困难.

关键词：高维数据; 两样本均值检验; stationary bootstrap; 协方差估计.

Abstract

In today's era, with the rise of genetic problems, a new kind of data, high-dimensional data, appears in various fields, which brings great challenges to the traditional statistical multivariate statistical theory and related methods. In the hypothesis testing brought by high-dimensional data, on the one hand, the classical test statistics are meaningless in the case of high-dimensional data; On the other hand, when the sample size n is far larger than the dimension p, the limit theory is no longer applicable. Creating limit theory suitable for high-dimensional data and improving the original statistics are popular areas of research nowadays. Based on the hypothesis test of the population mean, this article puts forward new ideas and new methods. Based on the statistical bootstrap method, the optimal statistics under certain conditions are proposed by improving the original statistics. At the same time, through the data simulation in R language, it is found that compared with previous methods, the statistics we put forward are superior in power and p-value under certain conditions, and the statistical bootstrap method we use also avoids the difficulty of theoretical calculation.

Keywords: high-dimensional data ; two sample mean test; stationary bootstrap; covariance estimation.

目录

第一章 绪论

1.1 研究背景

随着时代的发展, 如今已经步入了经济和信息全球化的时代, 科学技术也随之飞速发展, 随着计算机技术的兴起, 人们现在所能采用的技术也越来越发达, 所解决的问题在内容上也更加高深, 在思想上也更加复杂, 在数据上相比之前也更加庞大, 在实际工作中收集的数据的类型, 收集收集的方法和分和以前相比都有了本质上的区别. 这种情况和改变就给统计学的工作带来了挑战, 对于统计工作者而言, 针对不同的问题寻找与之相应的解法就成了切实之需. 例如, 在复杂系统的统计建模, 经济金融中的预测和决策, 高维, 定性和不彻底数据的统计分析等都是目前需要解决的问题, 虽然目前已经提出了一些方法,但是在问题的解决上还不够彻底, 也不能够很好地解决其他类型的问题. 因此, 统计学正面临着前所未有的巨大挑战和机遇. 同时还存在一些新的问题需要我们去探究. 例如, 在基因问题中, 我们获得的样本量往往很少, 但是每个样本量的基因信息很多, 这种数据就是我们常说的高维数据, 在许多时候这些数据彼此之间还可能是相互联系的, 不断变化的, 有时还会存在数据收集不完整的问题; 还例如通过计算机可以将图象和信号的信息转化为高维数据的信息进行处理, 而这种信号所转化数据由于其本身的特性一般来说都是前后相互联系的; 还有一些高维数据和原来的形式不同, (例如 DNA 和基因数据), 这种新型数据相对其维数而言, 样本量反而相当小, 但是在许多小样本问题中收集的数据常常含有没有收集彻底的数据等等. 对于以上这些存在的问题, 统计学工作者还有一些在实际领域中从事数据分析的工作者都进行了相应的研究, 在有些方面取得了一定的成绩, 已经形成了一些系统的且比较行之有效的方式, 但是在很多其他的方面统计研究还处在探索阶段, 在目前来说, 并没有一些系统的或者有效的方式,更严重的是在该领域甚至还缺乏完整的系统理论. 作为目前研究的重点. 近些年, 特别是在发达国家中, 有许多统计学家深入到实际领域, 或与实际领域的专家合作, 探索解决这些挑战性问题的途径.

在研究中, 我们所说的高维数据的一个共同特点就是数据的维度往往高于样本的大小, 这就是我们所熟知的"大 p ,小 n ". 在高维数据的情况下, 我们原来所应用的一些经典的极限理论已经不实用了. 如点估计的偏差在这种情况下就会特别大, 有些理论就算可以使用, 其效率也会非常低, 我们需要在原有的基础上提出新的统计量来对高维数据矩阵进行估计. 下面我们通过一个例子来说明这种高维数据对统计学产生的影响.

例: 设样本 $x_1, x_2, \cdots, x_n$ 是 n 个服从正态分布 $N_p \sim (\mu, I)$ 的 p 维随机变量,

且其均值为:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

记 $\overline{x}_i$ 和 μ_i 分别为 $\overline{x}$ 和 μ 的第 i 个分量,那么

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

即

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

从而

$$\|\overline{X} - \mu\|^2 = \sum_{i=1}^{p}(\overline{x}_i - \mu_i)^2 = \sum_{i=1}^{p} O_p(\frac{1}{n}) = p \cdot O_p(\frac{1}{n})$$

在上述分析中. 看到样本容量 n 保持不变时而维数 p 逐渐增加时, 样本均值 $\bar{X}$ 作为均值 μ 的估计的时候就会变得很差. 为了使 $||\bar{X} - \mu||^2 \xrightarrow{p} 0$ 能够在 p 增加的情况下仍然成立, 样本容量 n 必须以与 p 相同或者更快的速度增加, 这就是我们经常说的"维数诅咒"的一个典型的例子.

在目前来说,在解决高维数据统计推断的问题的时候, 我们经常采取的一个比较主流的思想, 是利用新的数学工具对原来的经典理论进行修正, 让原来的数学工具适应新的理论, 建立起高维框架下新的极限理论. Bai [1] 还有一些人利用大维随机矩阵谱理论对多元统计中很多经典的方法都进行了修正, 建立起了高维框架下统计推断的方法, 在他们的工作北背景下, 我们可以利用他们的方法对原来的统计检验方法进行修正, 提出新的统计检验方法, 让新的统计检验方法能够很好地应用在高维数据的情况下.

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列均值向量是否相等, 为了说明上的方便和叙述上的简洁明了, 在这里我们规定一下记号:假设随机向量 $X = (X_1, X_2..., X_p)$ 分布的均值向量为 $\mu_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1p})$, 协方差矩阵为 $\Sigma_1 = (\sigma_{1:q,s})_{q,s=1,...,p}$.假设随机向量 $Y = (Y_1, Y_2..., Y_p)$ 分布的均值向量为 $\mu_2 = (\mu_{21}, \mu_{22}, ..., \mu_{2p})$, 协方差矩阵为 $\Sigma_2 = (\sigma_{2:q,s})_{q,s=1,...,p}$.

我们考虑如下的假设检验问题:

$$H_0 : \mu_1 = \mu_2 \qquad vs \qquad H_1 : \mu_1 \neq \mu_2$$

假设 m 个来自 X 的独立样本 $X^{(1)} = (X_{11}, ..., X_{1p})^T, ...,^{(m)} = (X_{m1}, ..., X_{mp})^T$,

n 个来自 Y 的独立样本 $Y^{(1)}=(Y_{11},...,Y_{1p})^T,...,Y^{(m)}=(Y_{m1},...,Y_{mp})^T$.记:

$$\bar{X}=(\bar{X}_1,...,\bar{X}_p)^T=\frac{1}{m}\sum_{i=1}^{m}X^i,\hat{\Sigma_1}=(\hat{\sigma}_{1:q,s})_{q,s=1,...,p}=\frac{\sum_{i=1}^{m}(X^i-\bar{X})(X^i-\bar{X})^T}{m-1}$$

$$\bar{Y}=(\bar{Y}_1,...,\bar{Y}_p)^T=\frac{1}{n}\sum_{i=1}^{n}Y^i,\hat{\Sigma_2}=(\hat{\sigma_{2:q,s}})_{q.s=1,...,p}=\frac{\sum_{i=1}^{n}(Y^i-\bar{Y})(Y^i-\bar{Y})^T}{n-1}$$

当 m 和 n 大于 p 时, 传统的 Hotelling T^2 检验可以很好地解决问题: 当 m 和 n 小于 p 时, 传统的 Hotelling T^2 [2] 检验就不能应用了, 因为在这种情况下, 协方差矩阵 $\hat{\Sigma_1}$ 和 $\hat{\Sigma_2}$ 是奇异的, 此时的矩阵是不可逆的. 针对这个问题, 统计检验一般主要分为两种类型. 一种类型是基于两个样本均值向量的一些二次形式 $\bar{X}-\bar{Y}$ 的差异, 例如 Bai 和 Saranadasa[1] , Chen 和 Qin[3] , Srivastava (2013)[4], Gregory 等2015[5]这种类型的检验方法对于检测 $\mu_1-\mu_2$ 中的非零分量的稠密差异效果比较好; 另一种类型的检验是基于两个样本均值 $\bar{X}-\bar{Y}$ 的最大差异, 例如例如 Cai 等(2013, 2014)[6], Chang 等(2017)[7] 这个测试类型对于检验 $\mu_1-\mu_2$ 中非零向量较少时效果很好.

众所周知, 以上所有检验都不能在所有的情况下都保持一个好的经验功效. 在本文中, 我们重点介绍基于关于两个平均向量差 $\bar{X}-\bar{Y}$ 的一些二次形式. 当列向量的方差几乎相同的时候, 我们一般会采用统计量 $\bar{X}_1-\bar{Y}_1,...$ 和 $\bar{X}_p-\bar{Y}_p$ (Bai 和 Saranadasa ,1996[1]; Chen 和 Qin ,2010[3]) 来进行检验. 当列方差不同时,我们会采用有关总体均值的统计量 $\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n}}},...,\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n}}}$ (Gregory2015[5]). 为了进行统计推断, 一般需要估算统计量 $tr\Sigma_1$, $tr\Sigma_2$, $tr\Sigma_1^2$, $tr\Sigma_2^2$, $tr\Sigma_1\Sigma_2$ (Bai 和 Saranadasa 1996[1]; Chen 和 Qin 2010[3]);同时还需要估算 trR_1 , trR_2 , trR_1^2 , trR_2^2 , trR_1R_2 这几个统计量, 其中 R_1 和 R_2 分别是 Σ_1 和 Σ_2 的对应相关矩阵.(Gregory 等2015[5])采用了一种截断技术来估计非标准化下的列统计量的平方和, 虽然这些检验在某些情况下是可行的, 但在有些其他的情况下它们仍然无法达到理想的效果. 在本文中我们将采用 bootstrap 算法来计算 p 值, 从而避免了在理论上计算的繁琐和困难.

Efron (1979)[8]提出了 bootstrap 算法, 该算法的优势在于即使是在数据不独立时也能进行计算, 所以受到了统计学家的大量关注. Kunsch [9] 在 1989 年基于 bootstrap 的基础上提出了一种 moving block 的算法, 该算法一般应用在弱相关的数据上; Liu 和 Singh [10]在1992年也提出了一种 moving block 的算法. Politis 和

Romano (1994)[11]进一步提出了 stationary bootstrap 算法. 借用以上算法, 我们可以通过计算机编程来进行计算, 从而简化我们的工作.

1.2 本文思路

在前面的介绍中, 我们可以发现两样本均值检验统计量在形式上具有相同的部分, 在分子的部分是两样本均值向量差 $\bar{X}-\bar{Y}$ 的平方, 在分母部分则是对他们的方差的估计, 基于这种形式, 在本文中, 我们将提出一个带有权重的统计量, 统计量的分子也是均值向量的平方, 统计量的分母我们将以权重的形式给出, 也就是说分母是不确定的量, 需要我们通过计算去得出最优的估计. 在这里, 我们将两样本均值向量差看做一个整体, 具体的统计量记为如下的形式:

$$T_\omega=\frac{\bar{X}_1^2}{\omega_1^2}+\frac{\bar{X}_2^2}{\omega_2^2}+\cdots+\frac{\bar{X}_p^2}{\omega_p^2}$$

并找出统计量的权重的最优值. 其中 $\bar{X}_i(i=1,2...p)$ 就是列均值向量, $\omega_i^2(i=1,2...p)$ 就是未知的权重. 在这里如果 $\sum_{i=1}^p\omega_i^2$ 是已知的话 T_ω 就是一个统计量, 改变 $\omega_i^2(i=1,2...p)$ 为不同的值, 统计量也就会相应的变成不同的形式. 在后续的讨论中, 我们将通过计算得出权重的最优值. 在本节接下来的部分, 我们来说一说本文的主要思路安排, 如下:

- 第一节内容主要是介绍一些相关的预备知识还有在本文中将将要应用到的基本定理来方便我们后面的证明.
- 第二节内容主要是将基于我们提出的统计检验的方法, 来分析统计量的性质, 包括计算统计量的期望和方差以及功效函数, 最后我们通过不等式求出统计量权重的最优值. 在本节我们也会介绍其他的几种统计检验方法, 并在之后与我们的方法进行比较. 在这一节我们将介绍 stationary bootstrap 算法下我们检验方法的具体原理.
- 第三节内容主要是我们会进行大量的模拟来计算经验 p 值, 我们将会通过模拟产生正态分布的数据, 拉普拉斯分布的数据, 混合分布的数据这三种数据在两种不同的协方差结构下的模拟来观察我们提出的统计量的性能.
- 第四节内容主要是我们会用我们提出的检验方法对实际数据进行分析, 通过比较得出我们的统计方法是否具有比较好的性能.
- 第五节内容主要是我们将总结我们的统计量的优缺点, 以及提出后续工作的改进方向.

1.3 预备知识

在这一节, 我们主要介绍一下在本文中应用到的相关符号和相关定义.

定义1.1. *协方差矩阵的概念如下:*

设 $X=(X_1,X_2,\cdots,X_n)'$ 为 n 维随机变量, 称矩阵

$$C=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为 n 维随机向量X的协方差矩阵, 也记为$D(X)$,其中

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

为 X 的分量 X_i 和 X_j 的协方差*(在此假设它们都存在)*.

定义1.2. *多元正态分布的定义如下:*

设 $x=(x_1,\cdots,x_n)'$ 是一个 n 维随机向量, 且 $E(x)=\mu$ *(实向量)*, $Var(x)=\Sigma$ *(正定阵或非负定阵)*, 假如它的特征函数为

$$f_x(t)=exp\{it'\mu-\frac{1}{2}t'\Sigma t\}$$

则称 x 为 n 元正态随机向量, 其分布称为 n 元正态分布,记为 $x\sim N_n(\mu,\Sigma)$,而矩阵 Σ 的秩 $Rank(\Sigma)=r$ 称为这个分布的秩.

定义1.3. *无偏估计的定义如下:*

设 $(\mathscr{X},\mathscr{B},\{P_\theta,\theta\in\Theta\})$ 为可控参数统计结构, $g(\theta)$ 是未知参数, $x=(x_1,x_2,\cdots,x_n)$ 是来自该统计结构的一个样本, 若用 $\widehat{g}(x)$ 估计 $g(\theta)$,且

$$E_\theta(\widehat{g}(x))=g(\theta),\ \forall\,\theta\in\Theta$$

则称 $\widehat{g}(x)$ 为 $g(\theta)$ 的无偏估计.

定义1.4. *原假设和备择假设:*

在统计中常把要检验的假设称为原假设, 为 H_0 ;在拒绝时所接受的假设称为备择假设, 记为 H_1 .

双边假设　　$H_0:\mu=\mu_0$　　$H_1:\mu\neq\mu_0$

定义1.5. 拒绝域和接受域:

使原假设被拒绝的样本观察值所组成的区域称为拒绝域; 保留原假设样本观察值所组成的区域称为接受域.

定义1.6. 第一类错误和第二类错误:

第一类错误(弃真错误)——原假设 H_0 为真,而检验结果为拒绝 H_0 ,记其概率为 α ,即

P (拒绝 $H_0|H_0$ 为真)$=\alpha$.

第二类错误(受伪错误)——原假设 H_0 不符合实际,而检验结果为接受 H_0 ,记其概率为 β ,即

P (接受 $H_0|H_0$ 为假)$=\beta$.

定义1.7. p值:

在一个假设检验问题中,拒绝假设 H_0 的最小显著性水平称为 p 值,利用 p 值和给定的显著性水平 α 可以建立如下的判断法则:

若 $\alpha \geq p$ 值,则拒绝原假设 H_0 ;

若 $\alpha \leq p$ 值,则保留原假设 H_0 .

定义1.8. 单个正态总体的假设检验:

设总体 $X \sim N(\mu, \sigma^2), X_1, X_2, ..., X_n$ 是取自总体的样本, $\bar{X}$ 和 S^2 为样本均值和方差, α 为显著性水平.

当 σ^2 已知时,均值 μ 的检验(U 检验):

双边检验:若检验的假设是: $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$ 取检验统计量 $U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

当 H_0 成立时, $U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

单边检验:若检验的假设是: $H_0 : \mu = \mu_0 \quad H_1 : \mu > \mu_0$ 取检验统计量 $U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

当 H_0 成立时, $U = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

当 σ^2 未知时,均值 μ 的检验(t 检验):

当 σ^2 已知时,均值 μ 的检验(U 检验):

双边检验:若检验的假设是: $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ 取检验统计量 $U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

当 H_0 成立时, $U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

单边检验:若检验的假设是: $H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ 取检验统计量 $U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

当 H_0 成立时, $U = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

第二章　新的统计检验方法

2.1　现有的其他的检验方法及研究动机

在本节, 我们将介绍两种已经有的检验方法. Srivastava (2013)[4]和其他一些人在2013年基于两样本均值统计量 $\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n}}},...,\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n}}}$ 提出了 SR 方法. 记:

$$A_{m,n,p}=(\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n}}})^2+...+(\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n}}})^2-p$$

当 $m,n,p\to\infty$ 时, Srivastava (2013)[4]已经证明了以下结论的正确性:

引理2.1.1.

$$SR=\frac{A_{m,n,p}}{\sqrt{Var(\hat{A}_{m,n,p})}}\xrightarrow{L}N(0,1)$$

在一些常规条件和0假设下, 我们有:

$Var(\hat{A}_{m,n,p})=\frac{2trR^2}{p}-\frac{2}{p(m-1)m^2}(tr\hat{D^{-1}}\hat{\Sigma_1})^2-\frac{2}{p(n-1)n^2}(tr\hat{D^{-1}}\hat{\Sigma_2})^2$, 其中 $R=\hat{D^{-1/2}}(\frac{\hat{\Sigma_1}}{m}+\frac{\hat{\Sigma_2}}{n})\hat{D^{-1/2}}$, $\hat{D}=\frac{\hat{D}_1}{m}+\frac{\hat{D}_2}{n}$, $\hat{D}_i=diag(\hat{\sigma_{i:1,1}},...,\hat{\sigma_{i:p,p}})(i=1,2)$. 在这个方法中, 对 $Var(\hat{A}_{m,n,p})$ 的估计是 SR 方法中非常重要的一部分, 但是 $Var(\hat{A}_{m,n,p})$ 的估计并不是容易得到的.

Gregory (2015)[5]和一些人在2015年提出了 GCT 的检验方法, 同样也是基于平方和统计量 $\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n}}},...,\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n}}}$,记:

$$B_{m,m,p}=(\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n}}})^2+...+(\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n}}})^2-p\hat{\xi}_{m,n,p}$$

其中 $\hat{\xi}_{m,n,p}$ 在 p 不是很大时, 即 $(p=O((n+m)^2)$ 时为1,在 p 很大时, 即 $(p=O((n+m)^6)$ 取值为其它数, 同样地, 当 $m,n,p\to\infty$ 时, 在0假设的情况下, Gregory (2015)[5]已经证明了以下结论成立:

引理2.1.2.

$$GCT = \frac{B_{m,n,p}}{\sqrt{Var(\hat{B}_{m,n,p})}} \xrightarrow{L} N(0,1)$$

其中 $Var(\hat{A}_{m,n,p}) = \sqrt{p}\hat{\xi}^2_{m,n,p}$, $\sqrt{p}\hat{\xi}^2_{m,n,p} = \sum_{|k|<L}\omega(k/L)\hat{\gamma}(k)$, $\hat{\gamma}(k) = (p-k)^{-1}\sum_{j=1}^{p-k}(t^2_{s,j} - Q_{m,n,p})(t^2_{s,(j+k)} - Q_{m,n,p})$, $t_{s,1} = \frac{\bar{X}_1 - \bar{Y}_1}{\sqrt{\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n}}}$,

$Q_{m,n,p} = \frac{1}{p}\sum_{q=1}^{p}(\frac{\bar{X}_p - \bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m} + \frac{\hat{\sigma_{2:p,p}}}{n}}})$, $\omega(x)$ 是 x 的阶梯偶函数, 满足: $\omega(0) = 1$,对于所有的 x , $|\omega(x)| \leq 1$,还有当 $|X| > 1$ 时, $\omega(x) = 0$, L 是用户选择的滞后窗口的大小. 同样的, 对 $Var(\hat{B}_{m,n,p})$ 的估计是 GCT 方法中非常重要的一部分, 但是 $Var(\hat{B}_{m,n,p})$ 的估计并不是容易得到的. Gregory (2015)[5]采用了 Brockwell (1991)[12], Politis 和 Romano (1995)[13]的想法来估计 $Var(\hat{B}_{m,n,p})$. 由于 $Var(\hat{A}_{m,n,p})$ 和 $Var(\hat{B}_{m,n,p})$ 的值并不一样, 所以 SR 方法和 GCT 方法在功效上的表现也不一样.

2.2 新的统计检验方法

2.2.1 加权均值统计量

我们在前面介绍了在两样本的情况下之前已经提出来的检验方法, 可以看到前面两种检验方法的统计量在形式上具有相近的形式, 在本节中, 我们将基于前两种统计量的一般形式提出了新的检验方法 T^{NEW} ,检验统计量和之前的两种方法不同, 在接下来, 我们将定义一种新的加权均值统计量的形式. 不妨令 $x_1, x_2, \cdots, x_n$ 是 n 个独立同分布的 p 维随机向量, 设 $X = (X_1, X_2, ..., X_n)$ 为 n 维随机变量, 称矩阵

$$X_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n\times p} \tag{2.1}$$

为 n 维随机向量 X 的数据矩阵, 记为 $X_{n\times p}$,其中 $c_{ij} = Cov(X_i, X_j)$.

我们再规定以下记号:

$$E(\bar{x}_i) = \mu_i,\ D(\bar{x}_i) = \sigma^2_{ii}\ ,\ E(\bar{x}^4_i) = \sigma^4_{ii}\gamma_{ii},$$

$$X = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)^T \ . \ \mu = (\mu_1, \mu_2, \ldots, \mu_p)^T.$$

权重的矩阵记为:

$$\omega = \begin{pmatrix} \frac{1}{\omega_{11}^2} & & & \\ & \frac{1}{\omega_{22}^2} & & \\ & & \vdots & \\ & & & \frac{1}{\omega_{pp}^2} \end{pmatrix} \tag{2.2}$$

在这里,我们从最特殊的情况入手. 将协方差矩阵取为对角矩阵并记为:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & & & \\ & \sigma_{22}^2 & & \\ & & \vdots & \\ & & & \sigma_{pp}^2 \end{pmatrix} \tag{2.3}$$

构造如下的统计量:

$$T_\omega = \frac{\bar{X}_1^2}{\omega_1^2} + \frac{\bar{X}_2^2}{\omega_2^2} + \cdots + \frac{\bar{X}_p^2}{\omega_p^2}$$

写成向量的形式可以写为 $T_\omega = X^T \omega X$.

我们构造的统计量和在第一节中给出的基于平方和检验的统计量 $\frac{\bar{X}_p - \bar{Y}_p}{\sqrt{\frac{\hat{\sigma_{1:p,p}}}{m} + \frac{\hat{\sigma_{2:p,p}}}{n}}}$ 在形式上是一致的, 只不过为了讨论的方便, 我们将原来的统计量进行了归纳, 并提出了新的统计量:

$$T_\omega = \frac{\bar{X}_1^2}{\omega_1^2} + \frac{\bar{X}_2^2}{\omega_2^2} + \cdots + \frac{\bar{X}_p^2}{\omega_p^2}$$

其中 $\bar{X}_i (i = 1, 2...p)$ 就是列均值向量, $\omega_i^2 (i = 1, 2...p)$ 就是未知的权重. 我们称这种统计量为加权均值统计量, 加权统计量不同于之前的统计量, 我们定义加权均值统计量的目的是为了找到更好的权重, 在下一节, 我们将探究具有这种形式的加权统计量的性质, 通过计算解出在一定条件下的最优权重最优权重. 我们将探究该统计量的一般性质.

2.2.2 在权重已知时, T_ω 的性质

在前面，我们将统计量记为$T_\omega = X^T \omega X$

利用该统计量的向量形式我们可以很轻松的计算出该统计量的期望:

推论2.1.

$$E(T_\omega) = tr\omega\Sigma + \mu^T\omega\mu$$

证明：

$$E(T_\omega) = E(X^T\omega X) = E[tr(X^T\omega X)] = tr\omega\Sigma + \mu^T\omega\mu$$

□

如果要写成分量的形式,有以下的推论:

推论2.2.

$$E(T_\omega) = \sum_{i=1}^{p} \frac{\mu_p^2}{\omega_p^2}$$

证明：

$$E(T_\omega) = E(\frac{\bar{X}_1^2}{\omega_1^2} + \frac{\bar{X}_2^2}{\omega_2^2} + \cdots + \frac{\bar{X}_p^2}{\omega_p^2}) = \sum_{i=1}^{p} E(\frac{\bar{X}_p^2}{\omega_p^2}) = \sum_{i=1}^{p} \frac{\mu_p^2}{\omega_p^2}$$

□

分量的计算过程相对向量的计算过程相对繁琐一些, 但是分量的形式在方差的计算中较为清晰和直观, 我们将在第六节里给出详细的计算过程. 方差的计算较期望在计算过程上较为繁琐, 在这里我们给出以下的定理:

推论2.3.

$$D(T_\omega) = \Sigma_{i=1}^{p} \frac{\sigma_{ii}^4\gamma_{ii}}{\omega_i^4} - \Sigma_{i=1}^{p} \frac{(\sigma_{ii}^2 + \mu_i^2)^2}{\omega_i^4}$$

证明：

因为 $T_\omega^2 = (\frac{\bar{X}_1^2}{\omega_1^2} + \frac{\bar{X}_2^2}{\omega_2^2} + \cdots + \frac{\bar{X}_p^2}{\omega_p^2})^2 = \Sigma_{i=1}^{p} \frac{\bar{X}_i^4}{\omega_i^4} + \Sigma_{i\neq j}^{p} \frac{\bar{X}_i^2 X_j^2}{\omega_i^2\omega_j^2}$

所以 $E(T_\omega^2) = E(\Sigma_{i=1}^{p} \frac{\bar{X}_i^4}{\omega_i^4}) + E(\Sigma_{i\neq j}^{p} \frac{\bar{X}_i^2 X_j^2}{\omega_i^2\omega_j^2})$

又因为 $E^2(T_\omega) = (\sum_{i=1}^{p} \frac{\mu_p^2}{\omega_p^2})^2$

所以 $D(T_\omega) = \Sigma_{i=1}^{p} \frac{\sigma_{ii}^4\gamma_{ii}}{\omega_i^4} - \Sigma_{i=1}^{p} \frac{(\sigma_{ii}^2 + \mu_i^2)^2}{\omega_i^4}$

写成向量的形式为: $D(T_\omega) = (\gamma - 1)tr(\omega\Sigma^2\omega) - (\mu^T\omega\mu) + 2(\mu^T\omega\Sigma\omega\mu)$ □

引理 2.1. *设 X_n 是一列独立同分布(i.i.d)的随机变量序列, 具有公共的数学期望 μ 和方差 σ^2 ,则*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

由引理 2.1 可知, 如果说给出的高维数据矩阵中的元素是独立同分布的, 那么矩阵的列平均值 $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p$ 也服从正态分布, 在这种情况下那么有:

$$\gamma_{11} = \gamma_{22} = \cdots = \gamma_{pp} = \gamma$$

结合上述的条件, 我们刚刚得到的式子可以进行化简:

推论2.4.

$$D(T_\omega) = (\gamma - 1)\sum_{i=1}^{p}\frac{\sigma_{ii}^4}{\omega_i^4} + 2\sum_{i=1}^{p}\frac{\sigma_{ii}^2\mu_i^2}{\omega_i^4} - \sum_{i=1}^{p}\frac{\mu_i^4}{\omega_i^4}$$

证明: 当 $\mu = 0$ 时,带入上述式子有: $D(T_\omega) = (\gamma - 1)\sum_{i=1}^{p}\frac{\sigma_{ii}^4}{\omega_i^4}$,

转换为向量形式: $D(T_\omega) = (\gamma - 1)tr(\omega\Sigma^2\omega) - (\mu^T\omega\mu) + 2(\mu^T\omega\Sigma\omega\mu)$,

当 $\mu = 0$ 时, $D(T_\omega) = (\gamma - 1)tr(\omega\Sigma^2\omega)$. □

2.2.3 在权重已知时, T_ω 假设检验和功效函数的计算

在预备知识中, 我们介绍了假设检验的方法和步骤, 在数理统计中, 我们需要用样本去判断正确与否的命题成为一个假设. 根据研究的目的提出的假设称为原假设, 记为 H_0 ; 而其对立面假设称为备择假设, 记为 H_1 .

假设检验是统计检验中强有力的数学工具, 运用"假设检验"的思想来解决统计检验的问题, 能够避免自己的主观印象对结果的影响, 保证了判断结果的客观公正、符合实际. 合理运用假设检验的统计方法,分析现场的实际数据, 对结果做出了合理的判断, 可以为正确的决策提供科学的依据, 也可以很好地推动各行各业的深入发展.

假设检验的基本思想是小概率反证法思想. 假设检验可以分为正态分布检验、正态总体均值分布检验、非参数检验三类. 在本文中, 我们应用到的是正态总体均值检验的双边检验.

基于我们的统计量, 我们提出的假设检验如下:

$$H_0 : \mu = 0 \qquad H_1 : \mu \neq 0$$

拒绝域为:

$$Z > Z_{\alpha/2} \text{ 或 } Z < Z_{-\alpha/2}$$

由此假设,可以计算出来功效函数记为:

推论2.5.

$$Power = 1 + P(\frac{(T_\omega - tr\omega\Sigma)}{\sqrt{(\gamma-1)tr(\omega\Sigma^2\omega)}} > Z_{\alpha/2}) - P(\frac{(T_\omega - tr\omega\Sigma)}{\sqrt{(\gamma-1)tr(\omega\Sigma^2\omega)}} > Z_{-\alpha/2})$$

以上的记号过于繁琐, 为了书写的方便, 我们将

$$\frac{\sqrt{((\gamma-1)tr(\omega\Sigma^2\omega) - \mu^T\omega\mu + 2\mu^T\omega\Sigma\omega\mu)}}{\sqrt{(\gamma-1)tr(\omega\Sigma^2\omega)}}$$

记为 λ .

在此记号下, 原来的功效函数可以表示为:

推论2.6.

$$Power = 1 + P(\frac{(T_\omega - tr\omega\Sigma - \mu^T\omega\mu) + \mu^T\omega\mu}{\sqrt{((\gamma-1)tr(\omega\Sigma^2\omega) - \mu^T\omega\mu + 2\mu^T\omega\Sigma\omega\mu)}} * \lambda > Z_{\alpha/2})$$

$$-P(\frac{(T_\omega - tr\omega\Sigma - \mu^T\omega\mu) + \mu^T\omega\mu}{\sqrt{((\gamma-1)tr(\omega\Sigma^2\omega) - \mu^T\omega\mu + 2\mu^T\omega\Sigma\omega\mu)}} * \lambda > Z_{-\alpha/2})$$

$$= 1 + \Phi(\frac{Z_{\alpha/2}}{\lambda} - \frac{(T_\omega - tr\omega\Sigma - \mu^T\omega\mu) + \mu^T\omega\mu}{\sqrt{((\gamma-1)tr(\omega\Sigma^2\omega) - \mu^T\omega\mu + 2\mu^T\omega\Sigma\omega\mu)}})$$

$$-\Phi(\frac{Z_{-\alpha/2}}{\lambda} - \frac{(T_\omega - tr\omega\Sigma - \mu^T\omega\mu) + \mu^T\omega\mu}{\sqrt{((\gamma-1)tr(\omega\Sigma^2\omega) - \mu^T\omega\mu + 2\mu^T\omega\Sigma\omega\mu)}})$$

由于第一项的值很小, 在一般的情况下可以忽略不计. 并且在之前的假设条件下, 我们有以下的两个条件成立:

引理 2.2.

$$\mu^T\omega\mu = o(n^{-1}tr(\Sigma^2))$$

引理 2.3.

$$\mu^T\Sigma\mu = o(n^{-1}tr(\Sigma^2))$$

应用上面的两个条件可以将我们的功效函数进一步化简, 写成如下的形式:

推论2.7.

$$Power = 1 - \Phi(\frac{Z_{-\alpha/2}\sqrt{2tr(\omega\Sigma)^2} - \mu^T\omega\mu}{\sqrt{(\gamma-1)tr(\omega\Sigma^2\omega)}})$$

写成分量的形式有:

$$Power = 1 - \Phi(\frac{Z_{-\alpha/2}\sqrt{(\gamma-1)\sum_{i=1}^{p}\frac{\sigma_{ii}^4}{\omega_i^4}} - \sum_{i=1}^{p}\frac{\mu_i^2}{\omega_i^2}}{\sqrt{(\gamma-1)\sum_{i=1}^{p}\frac{\sigma_{ii}^4}{\omega_{ii}^4}}})$$

证明： 将上述式子整理得:

$$Power = 1 - \Phi(Z_{-\alpha/2} - \frac{\sum_{i=1}^{p}\frac{\mu_i^2}{\omega_i^2}}{\sqrt{(\gamma-1)\sum_{i=1}^{p}\sigma_{ii}^4}\sqrt{\sum_{i=1}^{p}\frac{1}{\omega_i^4}}})$$

$$Power = 1 - \Phi(Z_{-\alpha/2} - \frac{\sum_{i=1}^{p}\frac{\mu_i^2}{\omega_i^2}}{\sqrt{(\gamma-1)\sum_{i=1}^{p}\sigma_{ii}^4}\sqrt{\sum_{i=1}^{p}\frac{1}{\omega_i^4}}})$$

□

在上述式子的形式中, 我们可以看出当 $\sum_{i=1}^{p}\frac{\mu_i^2}{\omega_i^2}/\sqrt{\sum_{i=1}^{p}\frac{1}{\omega_i^4}}$ 越大时, 功效函数越大.

在这里, 我们令: $\alpha_i = \frac{\frac{1}{\omega_i^2}}{\sqrt{\sum_{i=1}^{p}\frac{1}{\omega_i^4}}}$, 由于权重的和为 1 , 于是有: $\alpha_1^2+\alpha_2^2+\cdots+\alpha_p^2 = 1$. 记: $f = \sum_{i=1}^{p}\frac{\mu_i^2}{\omega_i^2}/\sqrt{\frac{1}{\omega_i^4}}$, 则 $f = \sum_{i=1}^{p}\mu_i^2\alpha_i$.

引理 2.4. （*Cauchy*不等式）

$$(a_1^2 + a_2^2 + \cdots + a_p^2)\cdot(b_1^2 + b_2^2 + \cdots + b_p^2) \geq (a_1b_1 + a_2b_2 + \cdots + a_pb_p)^2$$

等号当且仅当 $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_p}{b_p}$ 时成立.

为了能够更加直观的应用该定理, 我们不妨记:

$\frac{\alpha_i}{\mu_i^2} = c$,则有 $\alpha_i = c\mu_i^2$. 可以计算出 $c = \frac{1}{\sqrt{\sum_{i=1}^{p}\mu_i^4}}$,

又由于 $\frac{\frac{1}{\omega_1^2}}{\mu_1^2} = \frac{\frac{1}{\omega_2^2}}{\mu_2^2} = \cdots = \frac{\frac{1}{\omega_p^2}}{\mu_p^2}$,

所以有 $\frac{1}{\omega_i^2} = \frac{\mu_i^2}{\sum_{i=1}^p \mu_i^2}$,

所以最优统计量为:

$$T_\omega = \frac{\bar{X}_1^2\mu_1^2}{\sum_{i=1}^p \mu_i^4} + \frac{\bar{X}_2^2\mu_2^2}{\sum_{i=1}^p \mu_i^4} + \cdots + \frac{\bar{X}_p^2\mu_p^2}{\sum_{i=1}^p \mu_i^4}$$

将上述结论应用在两样本的情况下, 统计量的形式变为:

$$T_\omega = \frac{(\bar{X}_1 - \bar{Y}_1)^2(\mu_{1,1} - \mu_{2,1})^2}{\sum_{i=1}^p (\mu_{1,i} - \mu_{2,i})^4} + \frac{(\bar{X}_2 - \bar{Y}_2)^2(\mu_{1,2} - \mu_{2,2})^2}{\sum_{i=1}^p (\mu_{1,i} - \mu_{2,i})^4} + \cdots + \frac{(\bar{X}_p - \bar{Y}_p)^2(\mu_{1,p} - \mu_{2,p})^2}{\sum_{i=1}^p (\mu_{1,i} - \mu_{2,i})^4}$$

通过上面的计算. 我们可以发现, 在我们给出的条件下, 当两样本的总体均值是四阶的时候, 得出来的经验功效会比为二阶的时候要好, 在这里, 我们给出 $\sum_{i=1}^p (\mu_{1,i} - \mu_{2,i})^4$ 统计量的估计量 $(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2$, 所以我们统计量的最终形式为:

$$T_\omega = \frac{(\bar{X}_1 - \bar{Y}_1)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2} + \frac{(\bar{X}_2 - \bar{Y}_2)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2} + \cdots + \frac{(\bar{X}_p - \bar{Y}_p)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2}$$

由中心极限定理可知:

$$T^{NEW} = \frac{T_\omega - E(T_\omega)}{\sqrt{Var(\hat{T_\omega})}} \xrightarrow{L} N(0,1)$$

2.3 Bootstrap 算法原理

在上文中,我们给出了统计检验方法 T^{NEW} , 统计量为:

$$T_\omega = \frac{(\bar{X}_1 - \bar{Y}_1)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2} + \frac{(\bar{X}_2 - \bar{Y}_2)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2} + \cdots + \frac{(\bar{X}_p - \bar{Y}_p)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m} + \frac{\hat{\sigma_{2:1,1}}}{n})^2}$$

我们知道在四阶的情况下, 新方法中的关于统计量的协方差 $Var(\hat{T_\omega})$ 的计算非常复杂, 因为协方差已经是八阶的统计量了, 在理论上的计算已经非常困难, 在这里我们不选择直接证明去求统计量的具体形式, 而是采用 stationary bootstrap 的算法重抽样的方式来计算协方差 $Var(\hat{T_\omega})$ 的数值. 在这一节中,我们将介绍 Bootstrap 算法具体情况. 以及如何利用 stationary bootstrap 的算法来计算.

Bootstrap 方法是非常有用的一种统计学上的估计方法, 是斯坦福统计系的教授 Bradley Efron [8]在总结、归纳前人研究成果的基础上提出一种新的非参数统计方法. Bootstrap 抽样的基本思想是在全部样本未知的情况下, 或者说目前只有部分样本的情况下, 我们借助当前已有的部分样本, 将这部分样本有放回地多次抽样, 通过这种方法就可以构建某个估计的置信区间. 简单地说, 就是如果直接使用当前的样本得到的估计并没有充分利用样本中有用的信息, 但是 Bootstrap 利用重抽样的方法, 相当于把样本量进行了扩大, 样本中的信息被充分的使用, 样本的剩余价值也就发挥在构建置信区间上.

在本文中, 我们将采用 stationary bootstrap 的算法来计算, stationary bootstrap 算法是 Politis 和 Romano [11]在1994年提出的方法. 关于 Block Bootstrap 中参数 block size 的最优估计可以参考 Politis 和 White (2004)[14]. 该重抽样方法可以采用 R 软件中的“tsbootstrap”函数进行实现运用这个方法, 我们可以通过计算机来完成统计量的计算过程.

2.4 Stationary bootstrap 算法

在本节, 我们将重点介绍 stationary bootstrap 算法的应用, 计算出统计量的 p 值. 假设 $t_{s,1}=\dfrac{(\bar{X}_1-\bar{Y}_1)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n})^2},...,t_{s,p}=\dfrac{(\bar{X}_p-\bar{Y}_p)^4}{(\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n})^2}$ 是渐进严平稳弱相关时间序列, 其均值在 0 假设下都渐近等于 3 , 其中 $\bar{X}_1,...,\bar{X}_p$ 为观察值 $X^i(i=1,...,m)$ 各维的均值, $\bar{Y}_1,...,\bar{Y}_p$ 为观察值 $Y^i(i=1,...,m)$ 各维的均值. 则基于 stationary bootstrap 算法的检验过程如下:

步骤1: 利用模拟的数据计算统计量中给出的 p 个学生化列统计量的值 $t_{s,1}=\dfrac{(\bar{X}_1-\bar{Y}_1)^4}{(\frac{\hat{\sigma_{1:1,1}}}{m}+\frac{\hat{\sigma_{2:1,1}}}{n})^2},...,t_{s,p}=\dfrac{(\bar{X}_p-\bar{Y}_p)^4}{(\frac{\hat{\sigma_{1:p,p}}}{m}+\frac{\hat{\sigma_{2:p,p}}}{n})^2}$ 并计算统计量的值, 记为:

$$T_0^{NEW}=\frac{p(\bar{t_s}-3)^2}{\bar{v_s}},$$

其中 $\bar{t_s}=\dfrac{1}{p}\sum_{i=1}^{p}t_{s,i}$, $\bar{v_s}=\dfrac{1}{p-1}\sum_{i=1}^{p}(t_{s,i}-\bar{t_s})^2$.

步骤2: 从**步骤1**中利用 stationary bootstrap 方法抽出 p 个样本, 我们把它们记为 $t_{s,1}^*,...,t_{s,p}^*$,并使用抽出来的样本计算 bootstrap 均值 $(\bar{t_{s,(1)}^*})$ 和方差 $(\bar{v_{s,(1)}^*})$,接着计算 bootstrap 统计量的值, 记为:

$$T_1^{NEW*}=\frac{p(\bar{t_{s,(1)}^*}-\bar{t_s})^2}{\bar{v_{s,(1)}^*}}.$$

步骤3: 重复**步骤2** B 次, 得到 B 个 bootstrap 统计量值, 记为T_1^{NEW*},T_2^{NEW*},T_B^{NEW*}.

步骤4: 计算 p 值.

$$P_{SB} = \frac{\sum_{b=1}^{B} I_{(T_B^{NEW*} > T_0^{NEW*})}}{B}.$$

第三章　统计模拟以及 p 值的计算

3.1　数据模拟

在这一节, 我们将对我们的理论部分进行模拟, 观察运算的结果,并对结果进行分析. 我们会将我们提出的统计量检验 T^{NEW} 与前文提到的 SR 方法以及 GCT 方法进行比较, 我们知道 GCT 是非常有效的, Gregory 等人为 GCT 提供了 R 包: "highD2pop" . 我们直接使用 GCT 的 R 包 "highD2pop" .

我们分别从 p 维多元正态分布和 Laplace 分布生成 m 个独立样本 $(X_{i1},\cdots,X_{ip})^T$ $(i=1\cdots m)$, 其中均值为 $(\mu_{1,1},\cdots,\mu_{1,p})^T$ $(i=1\cdots m)$, 协方差矩阵 Σ_1 . 同样的, 我们产生了 n 个独立的 p 维多元正态分布和拉普拉斯分布的样本 $(Y_{i1},\cdots,Y_{ip})^T(i=1\cdots n)$ 分别为 $(\mu_{2,1},\cdots,\mu_{2,p})^T(i=1\cdots m)$ 和协方差矩阵 Σ_2 .我们一般选择令 $\mu_1=(\mu_{1,1},\cdots,\mu_{1,p})^T=\mu_2=(\mu_{2,1},\cdots,\mu_{2,p})^T=(0,\cdots,0)$. 我们令 $D=(d_{q,s})q,s=1,...,p$ 为对角矩阵, 其对角元素产生于均匀分布(1,3). 在基因数据分析中, 相关系数矩阵往往都是稀疏的, 因此我们考虑如下的两种致密的协方差结构:

结构1: 致密结构, 协方差矩阵为 $\Sigma_1=D^{1/2}R_1D^{1/2}$, $\Sigma_2=D^{1/2}R_2D^{1/2}$.其中 $R_1=(\rho_1^{|q-s|})q,s=1,...,p$, $R_2=(\rho_2^{|q-s|})q,s=1,...,p$,且 $\rho_1,\rho_2=0,0.1,0.2,...,0.5$;

结构2: 致密结构, 协方差矩阵为 $\Sigma_1=D^{1/2}R_1^{-1}D^{1/2}$, $\Sigma_2=D^{1/2}R_2^{-1}D^{1/2}$,其中 $R_1=(\rho_1^{|q-s|})q,s=1,...,p$, $R_2=(\rho_2^{|q-s|})q,s=1,...,p$,且 $\rho_1,\rho_2=0,0.1,0.2,...,0.5$.

在高维数据的背景下, 总体均值往往都是稀疏的, 也就是说总体均值的很多分量都是 0 , 只有很少的一部分不是 0 . 为了观察检验的效果, 我们一般选取令 $\mu_1=(\mu_{1,1},\cdots,\mu_{1,p})^T=\mu_2=(\mu_{2,1},\cdots,\mu_{2,p})^T=(0,\cdots,0)$.

3.2　实验数据

我们先令 $m=60$, $n=80$,在经验水平基于 $\alpha=0.05$ 的时候, ρ_1,ρ_2 分别取 0 和 0.5 时, 并同时改变维数 p 的值分别为 300 , 500 , 800 .在基于结构 1 和结构 2 的正态分布下重复 2000 次运算, 由于 GCT 的性能比较差, 在这里我们主要比较 SR 和 T^{NEW} 两种方法的经验水平. 如下表所示, 可以看到在维数变化时, 两种方法都能保持经验水平在一个比较合理的范围内.

表 1: 基于结构1在正态分布下统计量犯第一类错误的概率

ρ_1	ρ_2	维数p	SR	T^{NEW}
0.5	0	300	0.0350	0.0595
		500	0.0420	0.0480
		700	0.0365	0.0400
0	0.5	300	0.0410	0.0610
		500	0.0420	0.0455
		700	0.0415	0.0455

表 2: 基于结构2在正态分布下统计量犯第一类错误的概率

ρ_1	ρ_2	维数p	SR	T^{NEW}
0.5	0	300	0.0425	0.0450
		500	0.0445	0.0400
		700	0.0450	0.0445
0	0.5	300	0.040	0.0475
		500	0.038	0.055
		700	0.0425	0.045

通过表格中的数据我们可以看到, 在维数增加的时候, 我们新提出的检验方法仍然能比较好的保持经验水平在 0.05 左右, 基本上没有受到维数增大的影响, 反而是在结构 1 的情况下, 当维数增加时, 经验水平控制得越好. 总的来说, T^{NEW} 检验方法和 SR 检验方法在样本时正态分布的情况下都能够比较好控制经验水平.

在接下来我们增加模拟的次数, 这里我们控制样本容量 $m = 60$, $n = 80$,数据维度 $p = 300$, 并且重复循环 $s = 2000$ 次. 让 ρ_1、ρ_2 分别取值 0.1 , 0.2 , 0.3 , 0.4 , 0.5.我们这里采用了三种不同的数据类型: 正态分布, 拉普拉斯分布, 混合分布. 从以下六张表的结果可以看出, 在基于这两种协方差结构的情况下, 三种分布下犯第一类错误的概率一直控制在 0.05 左右. 在大多数情况下, 我们提出的检验方法 T^{NEW} 都比 GCT 要控制得更好. 综上所述, 新提出的检验方法 T^{NEW} 具有良好的性能, 与其他两种同类的统计量相比, 它的经验水平控制的不错. 在下一章, 我们准备选用实际数据来进行模拟, 观察 T^{NEW} 在实际情况的应用下的效果.

表 3: 在两种协方差结构下统计量的经验功效

	ρ_1	ρ_2	SR	T	GCT10	GCT15	GCT20
正态分布 协方差结构1	0	0.5	0.934	0.905	0.8825	0.8835	0.8855
	0.5	0	0.8985	0.8875	0.8645	0.85	0.853
正态分布 协方差结构2	0	0.5	0.72	0.635	0.61	0.6145	0.6186
	0.5	0	0.681	0.595	0.573	0.5745	0.579
拉普拉斯 协方差结构1	0	0.5	0.9305	0.8965	0.8865	0.888	0.8895
	0.5	0	0.9395	0.9085	0.8865	0.889	0.8865
拉普拉斯 协方差结构2	0	0.5	0.7885	0.6965	0.657	0.6625	0.656
	0.5	0	0.7145	0.6245	0.5995	0.603	0.6035
混合分布 协方差结构1	0	0.5	0.914	0.8765	0.8555	0.86	0.863
	0.5	0	0.9055	0.865	0.8485	0.848	0.851
混合分布 协方差结构2	0	0.5	0.8285	0.7445	0.723	0.72	0.72
	0.5	0	0.711	0.6235	0.604	0.6105	0.6115

表 4: 基于协方差结构1在正态分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.041	0.0545	0.068	0.073	0.076
0.1	0.2	0.0435	0.0420	0.0645	0.071	0.075
0.1	0.3	0.0415	0.0520	0.065	0.0675	0.067
0.1	0.4	0.0425	0.0480	0.0635	0.0685	0.074
0.1	0.5	0.036	0.0515	0.065	0.069	0.0725
0.2	0.1	0.0405	0.0535	0.0625	0.065	0.0695
0.2	0.2	0.05	0.0400	0.057	0.062	0.0685
0.2	0.3	0.0435	0.0505	0.071	0.0715	0.074
0.2	0.4	0.0415	0.0044	0.072	0.0745	0.0775
0.2	0.5	0.045	0.0545	0.0615	0.0685	0.0705
0.3	0.1	0.0465	0.0500	0.061	0.063	0.0675
0.3	0.2	0.0355	0.0560	0.066	0.0715	0.074
0.3	0.3	0.0425	0.0570	0.066	0.068	0.074
0.3	0.4	0.034	0.0570	0.068	0.069	0.0755
0.3	0.5	0.035	0.0555	0.0635	0.0645	0.066
0.4	0.1	0.041	0.0515	0.0665	0.0705	0.0735
0.4	0.2	0.039	0.057	0.0755	0.0775	0.081
0.4	0.3	0.0435	0.0485	0.0635	0.0805	0.085
0.4	0.4	0.0385	0.0525	0.0685	0.069	0.073
0.4	0.5	0.0405	0.0505	0.0685	0.07	0.0735
0.5	0.1	0.042	0.0485	0.0795	0.08	0.0825
0.5	0.2	0.0385	0.0565	0.0725	0.074	0.0805
0.5	0.3	0.0385	0.0525	0.0695	0.073	0.075
0.5	0.4	0.0395	0.0515	0.0655	0.0655	0.066
0.5	0.5	0.037	0.0480	0.076	0.0765	0.079

表 5: 基于协方差结构2在正态分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.036	0.0435	0.059	0.063	0.0665
0.1	0.2	0.04	0.05	0.058	0.0655	0.07
0.1	0.3	0.0495	0.0455	0.0775	0.0795	0.0835
0.1	0.4	0.042	0.054	0.0755	0.0775	0.0825
0.1	0.5	0.045	0.0405	0.065	0.073	0.0765
0.2	0.1	0.04	0.056	0.0655	0.067	0.0705
0.2	0.2	0.0395	0.051	0.0615	0.065	0.07
0.2	0.3	0.0405	0.052	0.061	0.063	0.068
0.2	0.4	0.044	0.054	0.0665	0.0735	0.0745
0.2	0.5	0.0345	0.0475	0.071	0.072	0.073
0.3	0.1	0.0455	0.047	0.0555	0.0585	0.0625
0.3	0.2	0.0405	0.0425	0.06	0.0635	0.071
0.3	0.3	0.044	0.054	0.0715	0.0745	0.077
0.3	0.4	0.0435	0.049	0.0645	0.067	0.068
0.3	0.5	0.042	0.0515	0.081	0.084	0.0885
0.4	0.1	0.0435	0.0535	0.0695	0.075	0.077
0.4	0.2	0.039	0.051	0.0585	0.0625	0.0655
0.4	0.3	0.0445	0.0425	0.062	0.066	0.0695
0.4	0.4	0.043	0.049	0.0775	0.0815	0.0845
0.4	0.5	0.0415	0.0525	0.0675	0.072	0.0755
0.5	0.1	0.044	0.0475	0.065	0.069	0.071
0.5	0.2	0.033	0.051	0.067	0.0685	0.0705
0.5	0.3	0.044	0.0505	0.074	0.0765	0.082
0.5	0.4	0.0405	0.054	0.0715	0.075	0.078
0.5	0.5	0.037	0.0425	0.0735	0.0755	0.076

表 6: 基于协方差结构1在拉普拉斯分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.0465	0.056	0.068	0.0725	0.075
0.1	0.2	0.04	0.0565	0.0595	0.066	0.068
0.1	0.3	0.038	0.049	0.063	0.066	0.0705
0.1	0.4	0.036	0.0545	0.0675	0.073	0.0785
0.1	0.5	0.0405	0.0655	0.0665	0.066	0.0705
0.2	0.1	0.0365	0.058	0.081	0.0855	0.087
0.2	0.2	0.0395	0.056	0.072	0.073	0.077
0.2	0.3	0.0405	0.063	0.0605	0.0645	0.07
0.2	0.4	0.043	0.059	0.077	0.084	0.089
0.2	0.5	0.042	0.072	0.066	0.0675	0.072
0.3	0.1	0.0435	0.054	0.0715	0.0755	0.08
0.3	0.2	0.0445	0.055	0.0705	0.074	0.0735
0.3	0.3	0.0305	0.0575	0.068	0.07	0.075
0.3	0.4	0.0435	0.0695	0.069	0.0705	0.0735
0.3	0.5	0.044	0.068	0.0705	0.0775	0.0825
0.4	0.1	0.041	0.058	0.069	0.072	0.0755
0.4	0.2	0.037	0.0645	0.075	0.076	0.0785
0.4	0.3	0.046	0.0625	0.07	0.075	0.081
0.4	0.4	0.0365	0.066	0.067	0.0695	0.0725
0.4	0.5	0.043	0.077	0.065	0.0685	0.073
0.5	0.1	0.035	0.063	0.0685	0.0715	0.0735
0.5	0.2	0.041	0.063	0.0715	0.0685	0.0685
0.5	0.3	0.042	0.061	0.0745	0.081	0.0845
0.5	0.4	0.035	0.059	0.065	0.068	0.072
0.5	0.5	0.0395	0.074	0.0635	0.066	0.0695

表 7: 基于协方差结构2在拉普拉斯分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.045	0.06	0.07	0.0715	0.0745
0.1	0.2	0.04	0.048	0.06	0.066	0.0685
0.1	0.3	0.035	0.0535	0.07	0.0755	0.08
0.1	0.4	0.04	0.057	0.0755	0.08	0.086
0.1	0.5	0.038	0.065	0.07	0.073	0.077
0.2	0.1	0.0375	0.0515	0.0645	0.0675	0.0695
0.2	0.2	0.0455	0.047	0.059	0.064	0.069
0.2	0.3	0.0355	0.058	0.06	0.0635	0.067
0.2	0.4	0.047	0.0595	0.0695	0.0745	0.081
0.2	0.5	0.033	0.0585	0.067	0.0695	0.0745
0.3	0.1	0.0365	0.0615	0.072	0.0785	0.08
0.3	0.2	0.036	0.052	0.084	0.0865	0.09
0.3	0.3	0.0515	0.057	0.067	0.069	0.073
0.3	0.4	0.034	0.0525	0.0675	0.071	0.0735
0.3	0.5	0.0485	0.0695	0.0755	0.076	0.079
0.4	0.1	0.035	0.056	0.0695	0.0725	0.077
0.4	0.2	0.039	0.062	0.064	0.0665	0.074
0.4	0.3	0.037	0.0565	0.0635	0.067	0.0705
0.4	0.4	0.04	0.049	0.0745	0.079	0.0845
0.4	0.5	0.037	0.0625	0.0775	0.0795	0.079
0.5	0.1	0.033	0.0545	0.071	0.074	0.076
0.5	0.2	0.0395	0.061	0.071	0.0735	0.0765
0.5	0.3	0.0425	0.0635	0.0725	0.074	0.0765
0.5	0.4	0.046	0.0565	0.0705	0.0705	0.073
0.5	0.5	0.043	0.065	0.0725	0.0735	0.0775

表 8: 基于协方差结构1在混合分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.044	0.0425	0.061	0.065	0.0695
0.1	0.2	0.036	0.0465	0.0725	0.077	0.079
0.1	0.3	0.039	0.051	0.0525	0.055	0.0585
0.1	0.4	0.037	0.0565	0.068	0.072	0.0755
0.1	0.5	0.0375	0.0595	0.0645	0.0635	0.067
0.2	0.1	0.046	0.056	0.081	0.0865	0.0895
0.2	0.2	0.041	0.0495	0.072	0.072	0.077
0.2	0.3	0.042	0.059	0.07	0.075	0.0775
0.2	0.4	0.0435	0.061	0.066	0.0695	0.0765
0.2	0.5	0.0365	0.053	0.079	0.078	0.082
0.3	0.1	0.0405	0.0545	0.077	0.079	0.083
0.3	0.2	0.029	0.05	0.0685	0.0685	0.072
0.3	0.3	0.0415	0.0615	0.0645	0.069	0.0715
0.3	0.4	0.0495	0.0595	0.086	0.0915	0.095
0.3	0.5	0.0415	0.0645	0.075	0.08	0.0865
0.4	0.1	0.036	0.067	0.065	0.0665	0.071
0.4	0.2	0.0395	0.063	0.067	0.0695	0.074
0.4	0.3	0.04	0.05	0.0715	0.078	0.0705
0.4	0.4	0.0465	0.0725	0.068	0.071	0.077
0.4	0.5	0.0375	0.057	0.0755	0.079	0.0835
0.5	0.1	0.0505	0.061	0.065	0.069	0.076
0.5	0.2	0.043	0.0565	0.0755	0.0785	0.083
0.5	0.3	0.0455	0.0605	0.075	0.075	0.0815
0.5	0.4	0.0375	0.0695	0.0745	0.076	0.079
0.5	0.5	0.039	0.0735	0.081	0.083	0.09

表 9: 基于协方差结构2在混合分布下统计量犯第一类错误的概率

ρ_1	ρ_2	SR	T^{NEW}	GCT10	GCT15	GCT20
0.1	0.1	0.04	0.048	0.0645	0.0655	0.068
0.1	0.2	0.035	0.0535	0.0635	0.0645	0.0675
0.1	0.3	0.04	0.057	0.0665	0.067	0.07
0.1	0.4	0.038	0.065	0.07	0.0725	0.0805
0.1	0.5	0.0375	0.0515	0.067	0.0707	0.073
0.2	0.1	0.0455	0.047	0.06	0.0645	0.0665
0.2	0.2	0.0355	0.058	0.063	0.069	0.0745
0.2	0.3	0.047	0.0595	0.0685	0.0695	0.0775
0.2	0.4	0.033	0.0585	0.0725	0.0745	0.0795
0.2	0.5	0.0365	0.0615	0.0675	0.071	0.0755
0.3	0.1	0.036	0.052	0.074	0.0775	0.083
0.3	0.2	0.0515	0.057	0.0645	0.0695	0.073
0.3	0.3	0.034	0.0525	0.0755	0.077	0.0825
0.3	0.4	0.0485	0.0695	0.0705	0.075	0.079
0.3	0.5	0.035	0.056	0.074	0.076	0.0795
0.4	0.1	0.039	0.062	0.074	0.0775	0.0815
0.4	0.2	0.037	0.0565	0.069	0.066	0.0695
0.4	0.3	0.04	0.049	0.072	0.0755	0.082
0.4	0.4	0.037	0.0625	0.0715	0.0775	0.0815
0.4	0.5	0.033	0.0545	0.0795	0.083	0.088
0.5	0.1	0.0395	0.061	0.0665	0.0715	0.0785
0.5	0.2	0.0425	0.0545	0.0675	0.067	0.074
0.5	0.3	0.044	0.0635	0.07	0.0705	0.076
0.5	0.4	0.046	0.0565	0.0695	0.0745	0.0765
0.5	0.5	0.043	0.065	0.065	0.071	0.0745

3.3 结果分析

在这一节我们将对以上的数据从多个角度进行更加具体的分析.

(1)从协方差结构上来看: 当 ρ_1 和 ρ_2 取的值比较小的时候, 可以看出检验方法 T^{NEW} 在第一种协方差结构下犯第一类错误的概率一般会小于在第二种协方差结构下犯第一类错误的概率, 但是随着 ρ_1 和 ρ_2 取值的增大, 会发现检验方法 T^{NEW} 在第二种协方差结构下犯第一类错误的概率一般会小于在第一种协方差结构下犯第一类错误的概率. 这是由于协方差结构上的差异.

(2)从数据结构上来看: 检验方法 T^{NEW} 在数据是正态分布的情况下, 无论是在第一种协方差结构还是在第二种协方差结构下都控制的比较好, 波动也非常小; 在数据是混合分布时, 控制得也比较好, 波动也不是很大; 但是在数据是拉普拉斯分布的情况下, 只能在部分时候控制得比较好, 数据的波动也比较大, 这也是我们检验方法 T^{NEW} 的一个缺点, 不能很好地适应所有的数据特点.

(3)从维数上来看: 随着维数的增加, SR 方法明显会不断变大, 波动水平比较大, 检验方法 T^{NEW} 在维数变化的时候结果比较稳定, 这说明检验方法 T^{NEW} 能够很好地适应高维的情况.

基于以上的分析我们可以看出,我们提出的统计检验方法 T^{NEW} 虽然在数据类型不同的情况下会产生一些波动,但是犯第一类错误的概率都保持在一个不错的范围区间内,而且该检验方法在针对高维数据时有着不错的稳定性, 维数对其影响比较小, 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.

该方法也有着一些问题需要我们去解决: 该方法并不能很好地适应所有的数据, 对于 stationary bootstrap 来说, 数据要求是严平稳的弱相关的时间序列, 但是并不是所有的数据都能满足这种条件;其次在上文也提到, 这种计算方法的缺点就是在计算上耗费的时间比较多, 因为在计算的过程中需要对原来的样本进行重抽样, 这样势必会影响计算的效率, 但是好处就是能够消灭一部分数据计算中的误差,使实验的结果更加准确.

第四章　实际数据模拟

4.1　数据来源以及背景

在本章中, 我们将采用里面的数据集进行真实数据模拟工作. R程序软件安装包 "datamicroarray" 提供了一组数据集. 我们下载了该包中关于乳腺癌的数据集———— Gravier et al (2010)[15]研究的数据集. 女性乳腺癌近来已超过肺癌, 成为全球发病率第一癌症, 2020 年全球新发乳腺癌病例高达 230 万例, 占所有癌症发病患者的 11.7% . 而在我国, 乳腺癌的发病率也逐年增高, 带来严重健康问题. 特别是近年来, 乳腺癌的发病年龄逐渐年轻化, 虽然全球范围内乳腺癌的死亡率逐年降低, 但是目前国内乳腺癌死亡率下降的趋势并不明显. 到目前为止, 尽管科学家并不能准确地分析出乳腺癌的发病机制, 但现在已经发现了许多与乳腺癌发病、预后相关的因素. 在本节的数据集中, 该数据集认为基因组信息可以预测淋巴结阴性的 T 1 、 T 2 期浸润性导管癌(T 1, T 2 , N 0)的早期转移. 他们使用比较基因组杂交技术对 168 名患者进行了为期 5 年的研究, 其中的 111 例患者在诊断后没有出现病状情况的为良性, 57 例早期转移的患者为恶性, 他们的 2905 组基因的表达水平的数据已经通过 log2 转移归一化.

4.2　模拟结果

我们将 2905 个基因变量分为 6 个组, 前 5 组每组 500 个, 最后一组 405 个. 利用上述三种方法计算出来的 p 值如下表所示:

表 10: 统计量的 p 值

组号	个数	SR	T^{NEW}	GCT10	GCT15	GCT20
1	500	1.25e-01	1.35e-01	3.53e-05	2.94e-04	8.89e-04
2	500	2.52e-04	3.52e-02	4.67e-07	1.83e-05	1.27e-04
3	500	8.55e-12	2.89e-02	2.00e-07	1.25e-05	1.08e-04
4	500	1.85e-01	4.65e-03	1.52e-03	3.77e-03	5.65e-03
5	500	1.00e-01	2.56e-02	2.36e-03	7.29e-03	1.31e-02
6	405	8.24e-02	3.97e-02	3.87e-05	4.38e-04	1.60e-03

可以看到我们的检验统计量 T^{NEW} 可以较好地识别差异并分析数据, 检验的结果也是比较敏锐的.

第五章 总结与展望

本文研究的是基于高维数据的情况下, 两样本总体均值的假设检验的问题. 因为高维数据和传统数据类型的不同, 随之给经典的检验方法带来了很大的困难, 传统的统计检验方法在高维数据的背景下不再适用. 在本文中, 提出了一种新的方法 T^{NEW} 来检验高维两样本均值向量差异的问题. 该方法在一定程度上能够解决高维数据带来的困难, 与此同时, 由于计算机技术的飞速发展, 在本文的检验统计量的计算中我们采用了 stationary bootstrap 这一新方法. 同时利用计算机生成模拟数据, 对我们的检验方法以及其他两种检验方法进行了模拟, 通过模拟的结果, 印证了我们的检验方法在许多情况下是可以使用的, 且有着不错的效果.

本文为了研究检验方法 T^{NEW} 在一般的协方差结构下也能有效的使用, 基于高维数据的特点, 在协方差结构上采用了两种不同的协方差结构. 在数据的类型上, 采用了三种不同的数据类型. 由此产生的模拟结果应该是可靠的. 从模拟的结果可以看出, T^{NEW} 在这六种不同的情况下均有着不错的表现, p 值基本上都控制在 0.05 左右.

通过大量的数据模拟也可以看出三种方法各自不同的特点, SR 方法的经验水平基本上都有一定的保守性, 由于其估计的保守, 在模拟的数据中, 性能也是三种方法之中最好的; 而 GCT 方法在有些时候会失去控制, 性能也不够稳定; T^{NEW} 方法犯第一类错误的概率都基本上在合理的范围内, 整体性能表现也不错. 从理论上来看, 这三种方法的协方差结构都比较复杂, 特别是我们新提出来的方法 T^{NEW}, 而对协方差估计的好坏就会直接影响整体的经验水平, 本文的创新之处是采用了 stationary bootstrap 算法避免了从理论上对协方差的估计, 而是直接通过重抽样的方式直接计算, 带来的是在理论证明上的简单. 但是由于算法的循环比较多, 在数据模拟的时候, 大量的计算带来的弊端就是计算机在运行程序时, 第三种方法的速度就比较慢,所需时间就比较长.

在接下来的工作中, 我们需要改进的地方是可以增加理论上的证明, 加强本文提出的方法的可行性以及实用性, 同时对我们提出的检验方法进行优化, 让它的稳定性和适应性更强, 这也是我们进一步工作中将要研究的问题.

参考文献

[1] Bai Z,Saranadasa H. Effect of high dimension: By an example of a two sample problem.Statist.Sinica,1996:311-329.MR1399305.

[2] Hotelling H.The generalization of Student's ratio[M].Breakthroughs in statistics.Springer,New York,NY,1992:54-65.

[3] Chen S X, Qin Y L. A two-sample test for high-dimensional data with applications to gene-set testing. Ann.Statist.38(2010) 808-835.

[4] Srivastava M S, Fujikoshi. Multivariate analysis of variance with fewer observations than the dimension[J].Multivariate Anal, 97(2006)1927-1940.

[5] Gregory K B,Carroll R J,Baladayuthapani V,et al. A two-sample test for equality of means in high dimensional and sparse[J]. Journal of the American Statistical Association, 2015, 110(510):837-849.

[6] Cai T, Liu W, Xia Y. 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.

[7] Chang J, Zheng C, Zhou W X, et al. Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity[J]. Biometrics, 2017, 73(4): 1300-1310.

[8] Efron B. Bootstrap methods: another look at the jackknife[M]. New York: Breakthroughs in Statistics. Springer New York, 1992: 569-593.

[9] Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations.The Annals of Statistics, Vol. 17(3), 1217 - 1241.

[10] Liu, R. Y. and K. Singh (1992). Moving blocks jackknife and bootstrap capture weak dependence. In Exploring the Limits of Bootstrap (R. Lepage and L. Billard, eds.), 225 - 248. Wiley, New York.

[11] Dimitris N. Politis, Joseph P. Romano. The stationary bootstrap[J]. Journal of the American Statistical Association, 1994, 89(428): 1303-1313.

[12] Brockwell P J, Davis R A, Fienberg S E. Time series: theory and methods: theory and methods[M]. Springer Science and Business Media, 1991.

[13] Dimitris N, Politis, et al. Bias-corrected nonparametric spectral estimation[J]. Journal of Time Series Analysis, 1995, 16(1): 67-103.

[14] Politis, D. N. and White, H. (2004). Automatic block-length selection for the dependent bootstrap. Econometric Reviews, 23(1), 53-70.

[15] Gravier E, Pierron G, Vincent-Salomon A,et al.A prognostic DNA signature for T1T2 node-negative breast cancer patients[J]. Genes,chrpmosomes and cancer, 2010, 49(12): 1125-1134.

致　谢

时光飞逝，转眼已经到了要毕业的时候了，怀念华师的食堂，怀念华师的风景，更怀念在华师的点点滴滴，都是我人生中精彩的回忆。首先，我最要感谢我的导师李正帮老师。在论文创作过程中，我的导师一直时时跟进我的进度，虽然因为找工作还有疫情的原因耽误了论文的写作，但是我的导师一直都在热心帮助我，指导我，理解我，让我能够直面论文撰写过程中的困难，也非常感谢这三年来导师对我的培养，不仅教会了我许多知识，还教会了我很多人生的哲理，在此我表示对他衷心的感谢，感谢导师一直以来对我的帮助。

对于论文中使用R语言编写程序的部分，十分感谢贾婉茹以及周雅诗同学对我的支持，一直在帮助我解决编程中的困难。因为疫情的原因，大家的交流不是很方便，但是他们依然给予了我非常大的帮助，在此感谢她们，谢谢他们的支持和鼓励。

此外我还要感谢在我的大学期间对我生活帮助很大的宋老师，作为宋老师的助管，老师一直非常热心，不仅关心我的学业，还帮助我解决生活中的困难。一直给我提供很好的学习环境，谢谢宋老师一直以来的帮助和关心。

最后我还要感谢我的父母，是他们做好了我的后勤保障工作，为我写论文提供了良好的环境，在我的求学生涯中，不仅在我生活上，更在精神上无微不至地关心我，感谢他们的无私付出，我才能取得今天的成绩。

谨以此文献给所有无私帮助过我，关心我，支持我的人！

胡浩

2021 年 5 月