

电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS



论文题目 基于置换检验对高维数据两样本均值的假设
检验

学科专业 统计学

学 号 201821110327

作者姓名 杨 静

指导教师 孔 婀 芳 教 授

分类号 _____ 密级 _____

UDC ^{注1} _____

学位论文

基于置换检验对高维数据两样本均值的 假设检验

(题名和副题名)

杨 静

(作者姓名)

指导教师 孔 婀 芳 教 授

电子科技大学 成 都

(姓名、职称、单位名称)

申请学位级别 硕 士 学科专业 统 计 学

提交论文日期 2021.4.13 论文答辩日期 2021.5.24

学位授予单位和日期 电子科技大学 2021.06

答辩委员会主席 陈绍刚

评阅人 夏应存、高晴

注 1: 注明《国际十进分类法 UDC》的类号。

Permutation Tests for Two-Sample Means of High-dimensional Data

**A Master Thesis Submitted to
University of Electronic Science and Technology of China**

Discipline: Statistics

Author: Jing Yang

Supervisor: Prof. Efang Kong

School: School of Mathematical Sciences UESTC

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名： 杨静 日期： 2021 年 6 月 24 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名： 杨静 导师签名： 孔珂

日期： 2021 年 6 月 24 日

摘 要

高维数据均值向量假设检验是近来统计学者研究的热点及难点问题，很多实际问题都会涉及到高维均值的检验，例如，基因表达是否有差异，两种治疗方案疗效是否有显著差异等等。前人对此也做出较多研究，一方面，经典的极限理论是在维数固定而样本容量趋于无穷的假定下推导的，许多经典的多元统计方法将不再适用于高维数据；另一方面，大部分高维数据均值向量的检验方法都需明确总体分布或假定总体分布为正态分布，在实际情况中，总体分布经常是未知或不易确定的，因此这些方法便不再适用。本文提出了基于边缘标准化统计量的置换检验方法对高维两样本均值进行检验，重点对该方法的有效性条件以及检验效能进行研究。

本文提出基于边缘标准化的检验统计量，该检验统计量只需计算协方差矩阵的对角线元素，形式简单，应用于置换检验时大大减轻了计算负担。并且证实了基于边缘标准化统计量的置换检验一致有效性条件为 $\Sigma_x = \Sigma_y$ 或者 $c = 1/2$ ，即两总体方差相等或两样本数相同。本文利用 Bootstrap 方法的原理即生成的数据与原始观测数据具有相同的属性，从样本数较多的样本中抽取伪样本，构造两样本数 $m = n$ ，满足一致有效性条件。将伪样本带入到检验统计量中，形成经验分布函数，我们发现现在宽松条件下基于边缘标准化统计量的置换检验是一致有效的。

最后，本文通过模拟实验，比较新检验方法与现存的两种方法犯第一类错误的经验概率以及经验检验势能，验证了新检验方法可以更好的控制犯第一类错误的概率以及具有更高的检验势能，在大多数 (n, p) 组合下都具有较大的优势。在实际案例中，分析我国 A 股 6 个高维数据行业在两个不同时间段的平均月收益率是否有差异，发现有 5 个行业具有明显的“五月卖出”效应，这 5 个行业的股票平均月收益率在 5 月到 10 月和上一年 11 月到 4 月这两个时间段是不相等的。

关键词：置换检验，高维数据，样本均值，Bootstrap 方法，一致有效性

ABSTRACT

Mean vector hypothesis testing of high-dimensional data is a hot and difficult problem recently. Many practical problems will involve the test of two-sample means with high dimensional data, such as whether there is a difference in gene expression, whether there is a significant difference in the efficacy of two treatment regimens, and so on. On the one hand, the classical limit theory is derived under the assumption that the dimension is fixed and the sample size tends to be infinite. So many classical multivariate statistical methods will no longer be suitable for high-dimensional data; On the other hand, most methods of two-sample means with high dimensional data need to clarify the population distribution or assume that the population distribution is normal distribution, which is not suitable. In practice, the population distribution is unknown or uncertain generally, so these methods are no longer applicable. In this paper, we propose a permutation test based on marginal standardized statistics to test the hypothesis of two-sample means with high dimensional data. And this paper focuses on the validity condition and test efficiency of the method.

In this paper, we propose a test statistic based on marginal normalization. The test statistic only needs to calculate the diagonal elements of the covariance matrix, and its form is simple. When it is applied to permutation test, the computational burden is greatly reduced. In order to use the test statistic, we need to satisfy the uniform validity condition of $\Sigma_X = \Sigma_Y$ or $c = 1/2$, that is, the two populations have the same variance or the same number of samples. Using the principle of Bootstrap methods that the generated data has the same properties as the original observation data, pseudo samples are extracted from the samples with a large number, which can make the two samples have same numbers. We find that permutation test based on marginal standardized statistics is asymptotically consistent under mild conditions by introducing pseudo samples into test statistics to form the empirical distribution function.

Finally, the empirical sizes and empirical test potential energy of the new test method and the existing two methods are compared through some simulation experiments. It is verified that the new test method can better control the probability of making the first type of error and has higher test potential energy, which has greater advantages in most (n, p) combinations. In the actual case, we collect and analyze the stock prices of six

ABSTRACT

high-dimensional data industries in China's A-share market from May 2001 to October 2017. There are five industries with significant "sell in May" effect, which means the average monthly return of stocks from May to October is not equal to that from November to April of the previous year.

Keywords: permutation tests, high-dimansionality, test of mean-difference, Bootstrap methods, consistency of test

目 录

第一章 引言	1
1.1 研究工作的背景与意义.....	1
1.2 高维数据两样本均值检验方法的国内外研究现状与概述.....	2
1.2.1 传统的 Hotelling 检验方法	3
1.2.2 SD 检验方法	4
1.2.3 BS 检验方法	6
1.2.4 CQ 检验方法.....	7
1.2.5 其他检验方法	9
1.3 本文的主要贡献与创新.....	9
1.4 研究内容和结构安排.....	10
第二章 基于置换检验对两样本均值的假设检验	12
2.1 置换检验.....	12
2.1.1 置换检验的定义	12
2.1.2 置换检验的关键点	12
2.1.3 置换检验的优缺点	13
2.2 基于置换检验对一维数据两样本均值的假设检验.....	14
2.2.1 一维数据两样本均值的置换检验	14
2.2.2 一维数据两样本均值的其他检验方法	14
2.3 基于置换检验对多维数据两样本均值的假设检验.....	15
2.3.1 多维数据两样本均值的置换检验	15
2.3.2 多维数据两样本均值的其他检验方法	15
2.4 基于置换检验对高维数据两样本均值的假设检验.....	16
2.4.1 基于样本之差统计量的置换检验	17
2.4.1.1 检验统计量	17
2.4.1.2 置换检验的有效性条件	18
2.4.2 基于极大值统计量的置换检验	18
2.4.2.1 检验统计量	18
2.4.2.2 置换检验的有效性条件	19
2.4.3 基于改进的 Hotelling's T^2 统计量的置换检验	19
2.4.3.1 检验统计量	19

2.4.3.2 置换检验的有效性条件	20
2.4.4 基于边缘标准化统计量的置换检验	20
2.4.4.1 检验统计量	20
2.4.4.2 构造两样本数相等	21
2.4.4.3 检验的优缺点	21
2.5 本章小结	22
第三章 BOOTSTRAP 方法构造置换检验的一致有效性条件	23
3.1 BOOTSTRAP 方法	23
3.1.1 Bootstrap 方法的定义	23
3.1.2 Bootstrap 方法的原理	24
3.2 基于置换检验利用 BOOTSTRAP 方法对高维数据两样本均值假设检验	25
3.2.1 检验统计量	25
3.2.2 利用 Bootstrap 方法构造检验的一致有效性	26
3.3 本章小结	28
第四章 数值模拟分析	29
4.1 前期准备	29
4.1.1 数据来源	29
4.1.2 样本数量及维数选择	30
4.1.3 确定置换次数	30
4.2 模拟犯第一类错误的经验概率	30
4.3 模拟经验检验势能	33
4.4 实例分析	36
4.5 本章小结	38
第五章 全文总结与展望	39
5.1 全文总结	39
5.2 展望与不足	40
致 谢	41
参考文献	42
攻读硕士学位期间取得的成果	45

第一章 引言

1.1 研究工作的背景与意义

统计学中总体的一些性质通常是利用收集到的样本数据来推断, 其中假设检验是一种常用且重要的推断方法^[1]。对于均值的假设检验更是吸引了众多学者的研究与讨论, 其中对高维数据的研究更是不断深入。现实中很多问题都会关联到高维均值的检验, 例如, 基因表现是否有不同, 两种医疗方案效果是否有差异等等。现存的文献可以大致分为两类。一类是基于均值差异的平方和, 这类方法更适用于稠密的差异, 其中研究包括 Bai and Saranadasa(1996)^[2] 和 Chen and Qin(2010)^[3]。另一类是基于均值差异的绝对值, 这类方法更适用于稀疏的差异, 其中研究包括 Cai et al. (2014)^[4], Xu et al. (2016)^[5] 和 Chang et al. (2017)^[6]。

高维数据的出现对传统的多元统计理论是一种冲击, 经典的极限理论在高维情形下无法推导, 多数传统的多元统计方法变得不再适用^[23]。现存的大部分多元统计方法应用于高维数据时, 检验的经验概率可能会接近 1, 就算可以应用, 表现也会较差, 见 Hall 等(2005)^[7], Donoho 和 Tanner(2005)^[8]。当检验的数据总体分布为正态时, 一维常用的方法有 t 检验、U 检验、F 检验^[9], 多维有 Hotelling T^2 检验, Wilks Λ 检验^[10]。当总体分布不是正态分布或不易确定时, 检验统计量在 H_0 下的分布是什么成为统计学者研究的关键。相应的检验方法随之产生, R.A.Fisher(1935)^[11] 和 E.Pitman(1937,1938)^{[12][13]} 提出了置换检验方法, F.Wilcoxon(1945)^[14] 在已有方法的基础上提出了秩和检验, Efron(1979)^[15] 随后提出了自助检验。

置换检验是一种常用的检验方法, 随着运算速度的提高, 置换检验也越来越受到人们的关注^[16]。一些经典的假设检验方法需要明确在原假设成立下的渐进分布, 但当总体分布未知或总体分布不是正态分布时, 该渐进分布不易找到。置换检验利用一群算子对样本数据进行变换, 形成新的样本数据, 通过新样本数据计算统计量值来构成的经验分布函数, 用经验分布函数代替原假设成立下的渐进分布进行检验。如果经验分布和原假设成立下的渐进分布相同, 则说明该置换检验是一致有效的, 即可在有限的样本内控制犯第一类错误的概率。

假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立同服从 F_1 分布, 样本 $\{Y_j, j=1, \dots, n\}$ 相互独立同服从 F_2 分布, 并且都是 P 维的随机变量。

$$EX_i = \mu^X, \text{cov}(X_i) = \Sigma_i^X, EY_j = \mu^Y, \text{cov}(Y_j) = \Sigma_j^Y, N = m + n. \quad (1-1)$$

当维数 $P > 1$ 时,

$$X_i = (x_{i,1}, \dots, x_{i,p}) + \mu^X, \text{ 和 } Y_j = (y_{j,1}, \dots, y_{j,p}) + \mu^Y. \quad (1-2)$$

其中 μ^X 和 μ^Y 是 R^p 空间中的两个向量, 对任意的 $i=1, \dots, m, j=1, \dots, n$, $\{x_{i,k}, k=1, 2, \dots\}$ 和 $\{y_{j,k}, k=1, 2, \dots\}$ 是如下的时间序列:

$$x_{i,k+1} = a_i x_{i,k} + \xi_k, y_{j,k+1} = b_j y_{j,k} + \eta_k, k=1, 2, \dots, \quad (1-3)$$

其中 ξ_k 和 η_k 是随机误差项, $\{a_i, i=1, \dots, m\}, \{b_j, j=1, \dots, n\}$, 是超参数, 既可以是固定的也可以是随机的。

为了定义置换过程中的检验统计量, 首先说明

$$\begin{aligned} \bar{X}_m &\equiv (\bar{X}_{m,1}, \dots, \bar{X}_{m,p})^T = \frac{1}{m} \sum_{i=1}^m X_i, \bar{Y}_n \equiv (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,p})^T = \frac{1}{n} \sum_{j=1}^n Y_j, \\ \hat{\Sigma}_m^X &\equiv \hat{\Sigma}(X_1, \dots, X_m) = \frac{1}{m} \sum_i (X_i - \bar{X}_m)(X_i - \bar{X}_m)^T, \\ \hat{\Sigma}_n^Y &\equiv \hat{\Sigma}(Y_1, \dots, Y_n) = \frac{1}{n} \sum_j (Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T. \end{aligned} \quad (1-4)$$

定义 $Z^N = \{Z_1, \dots, Z_N\}$, 其中 $N = m + n$, $Z_i = X_i, 1 \leq i \leq m$; $Z_{m+j} = Y_j, 1 \leq j \leq n$ 。特别地, 假定 G_N 是样本的所有置换形式, 对任意的 $\pi \in G_N$, 令 Z_π^N 为样本 Z^N 根据置换形式 π 得到的新样本, 并且 $Z_{\pi(i)}^N (1 \leq i \leq N)$ 是 Z_π^N 的第 i 个部分。假设 S_N 为定义的检验统计量, $S_N(Z^N)$ 为根据原样本计算得到的统计量值。将置换后的样本带入 S_N , 重新计算检验统计量值 $S_N(Z_\pi^N)$, 并用 $\hat{R}_N^S(\cdot)$ 表示 $S_N(Z_\pi^N)$ 的经验分布函数, 定义为

$$\hat{R}_N^S(t) = \frac{1}{N!} \sum_{\pi \in G_N} \mathbf{I}\{S_N(Z_\pi^N) \leq t\}. \quad (1-5)$$

则相应置换检验拒绝 H_0 的条件是^[17]

$$\hat{R}_N^S(S_N(Z^N)) \geq 1 - \alpha. \quad (1-6)$$

此条件表明, 将 $S_N(Z^N)$ 这个初始统计量值带入 (1-5) 式, 若结果值大于等于 $1 - \alpha$, 则拒绝原假设。

1.2 高维数据两样本均值检验方法的国内外研究现状与概述

对于两样本高维数据均值的假设检验, 现存的文献里介绍了几种方法。这几种方法有相似也有差异, 都在不断尝试中做出改进。

1.2.1 传统的 Hotelling 检验方法

单变量的 t 统计量是大多数统计研究员所熟知的，应用也十分广泛。如果具有 n 个观测值的随机样本取自均值 μ 和方差 δ^2 的正态分布，则检验统计量形式如下：

$$t = \frac{\bar{x} - \mu}{S_0 / \sqrt{n}}, \quad (1-7)$$

其中 \bar{x} 为样本均值， S_0 为样本标准差，且分布服从自由度为 $(n-1)$ 的 t 分布。将检验统计量平方，得到

$$t^2 = n(\bar{x} - \mu)' (S_0^2)^{-1} (\bar{x} - \mu). \quad (1-8)$$

该值可定义为 \bar{x} 和 μ 之间的平方欧氏距离。

多变量的 t 统计量也是很容易构造的，假定 n 个随机样本 (X_1, \dots, X_n) 来源于 p 元正态分布，正态分布的均值向量为 μ^X ，且 μ^X 是 R^p 空间中的向量，协方差矩阵为 Σ 。公式 (1-7) 中 t 统计量的多元版本由下式给出：

$$T^2 = n(\bar{X} - \mu^X)' S^{-1} (\bar{X} - \mu^X), \quad (1-9)$$

其中样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ，样本协方差 $S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' / (n-1)$ 。公式 (1-

9) 中的统计量及其抽样分布最早由 Harold Hotelling(1931)^[19] 提出，通常称为 Hotelling's T^2 统计量。它首次被应用于检验原假设 $H_0: \mu = \mu_0$ ，即多元正态分布的均值向量 μ 等于一个特定向量 μ_0 ，备择假设为 $H_A: \mu \neq \mu_0$ 。因此 Hotelling's T^2 统计量可以作为多元单样本假设检验的统计量， T^2 统计量在原假设成立下的渐进分布可以由中心化的 F 分布给出

$$T^2 \sim \frac{(n-1)p}{n-p} F_{(p, n-p)}, \quad (1-10)$$

其中 $F_{(p, n-p)}$ 服从自由度 p 和 $(n-p)$ 的 F 分布，并且该分布的极限形式是卡方分布，卡方分布的自由度为 p 。

对于两样本均值的假设检验，在原假设成立的情况下，当两个总体都是 p 元正态分布，并且协方差矩阵满足 $\Sigma_X = \Sigma_Y$ ，Hotelling 检验为：

$$T^2 = \frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)' S^{-1} (\bar{X}_m - \bar{Y}_n), \quad (1-11)$$

其中样本均值 \bar{X}_m ， \bar{Y}_n 和样本协方差矩阵 S 定义如下：

$$\bar{X}_m \equiv (\bar{X}_{m,1}, \dots, \bar{X}_{m,p})^T = \frac{1}{m} \sum_{i=1}^m X_i, \bar{Y}_n \equiv (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,p})^T = \frac{1}{n} \sum_{j=1}^n Y_j,$$

$$S = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X}_m)(X_i - \bar{X}_m)^T + \sum_{j=1}^n (Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T \right). \quad (1-12)$$

同样的， T^2 统计量在原假设成立下的渐进分布可以由中心化的 F 分布给出

$$T^2 \sim \frac{(m+n-p-1)}{(m+n-2)p} F_{(p, m+n-p-1)}, \text{ 其中 } F_{(p, m+n-p-1)} \text{ 服从自由度 } p \text{ 和 } (m+n-p-1) \text{ 的 } F \text{ 分布。}$$

在“固定 p ，大 n ”的经典背景下，传统的 Hotelling's T^2 统计量有很多优良的性质。首先，它在所有的非奇异线性变化下即 $Z = AX + b$ 都是不变的。其中 X 是多元正态分布的抽样观测值， b 是已知的常数向量， A 是已知值构成的非奇异矩阵。这种不变性意味着 T^2 统计量的计算结果值与数据尺度大小和原点变化无关，因此基于 Z 的 T^2 统计量计算值与基于 X 的相同。由此可以得出结论：假设检验 $H_0: \mu_Z = \mu_{Z_0}$ 等同于检验 $H_0: \mu_X = \mu_{X_0}$ 。其次，基于 Hotelling's T^2 统计量的假设检验是一致最有效不变检验，这意味着在所有的不变检验中，没有一种检验比 Hotelling's T^2 统计量更能检测出真实的零假设。最后，Hotelling's T^2 统计量等价于似然比统计量，因此似然比检验的最优和分布性质完全适用于 Hotelling's T^2 统计量。

而且除了在多元分析中的应用外，该检验统计量在多元统计过程控制中也得到了广泛的应用。在 1947 年 Harold Hotelling^[20] 使用该方法解决了有关轰炸瞄准器的问题，而且随着计算能力的提升，Hotelling's T^2 统计量已经成为监测多变量过程中最流行的图表统计量之一。

当 $p > m+n-2$ ，即数据维数大于样本容量时，样本协方差矩阵 S_p 是不可以求逆的，因此检验统计量 T^2 没有定义，Hotelling 检验方法不可行^[23]。所以对于高维数据的均值检验问题，传统的 Hotelling 检验并不适用，需要寻找更加有效的方法。

1.2.2 SD 检验方法

Hotelling 检验统计量在高维数据的情形下是不可行的，为了使得样本协方差矩阵可逆，Srivastava 和 Du(2008)^[21] 想到了一个方法即用样本协方差矩阵 S 的对角线元素构成的对角阵来代替检验统计量中的 S 。假定两总体都服从 P 元正态分布，每个总体中的样本相互独立同分布，其中 s_{11}, \dots, s_{pp} 是样本协方差矩阵的对角线元素，定义如下：

$$D_s = \text{diag}(s_{11}, \dots, s_{pp}), \quad (1-13)$$

$$R = D_s^{-\frac{1}{2}} S D_s^{-\frac{1}{2}}, \quad (1-14)$$

R 是样本的相关系数矩阵, 定义修正系数 $c_{p,k}$ 如下:

$$k=m+n-2, \quad (1-15)$$

$$c_{p,k} = 1 + \frac{\text{tr} R^2}{p^{3/2}}, \quad (1-16)$$

在原假设成立的情况下, 给出了三个检验统计量:

$$T_1 = \frac{\frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)^T D_s^{-1} (\bar{X}_m - \bar{Y}_n) - \frac{kp}{k-2}}{\left[2 \left(\text{tr} R^2 - \frac{p^2}{k} \right) c_{p,k} \right]^{\frac{1}{2}}}, \quad (1-17)$$

$$T_2 = \frac{\frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)^T (\bar{X}_m - \bar{Y}_n)}{\text{tr} S}, \quad (1-18)$$

$$T_3 = \frac{\frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)^T (\bar{X}_m - \bar{Y}_n) - \text{tr} S}{\left[\frac{2k(k+1)}{(k-1)(k-2)} \left(\text{tr} S^2 - \frac{(\text{tr} S)^2}{k} \right) \right]^{\frac{1}{2}}}, \quad (1-19)$$

在如下备择假设

$$\mu_X - \mu_Y = \left(\frac{m+n}{kmn} \right)^{\frac{1}{2}} \delta, \quad (1-20)$$

其中 δ 为常数向量且满足条件 $\frac{\delta' D_s^{-1} \delta}{n} \leq M$, 对 $\forall p$ 且常数 M 并不依赖于维度 p 。三

种检验统计量的势能函数依次为

$$\beta(T_1 | \mu_X, \mu_Y) \approx \Phi \left(-Z_{1-\alpha} + \frac{mn}{m+n} \frac{(\mu_Y - \mu_X)' D_s^{-1} (\mu_Y - \mu_X)}{\sqrt{2 \text{tr} R^2}} \right), \quad (1-21)$$

$$\begin{aligned} \beta(T_2 | \mu_X, \mu_Y) &\approx \beta(T_3 | \mu_X, \mu_Y) \\ &\approx \Phi \left(-Z_{1-\alpha} + \frac{mn}{m+n} \frac{(\mu_Y - \mu_X)' (\mu_Y - \mu_X)}{\sqrt{2 \text{tr} \Sigma^2}} \right). \end{aligned} \quad (1-22)$$

可以根据势能函数计算出检验统计量的势能, 判断哪种检验统计量更优。

对于样本协方差矩阵不可逆的问题, SD 检验方法虽然是有效的, 但当变量之间具有相关性该方法可能就不再适用。在 Bickel 和 Levina (2004) [22] 中, 作者提到在变量间不具有相关性的情形下, SD 检验方法这样的处理是合理且有效的。但

在实际问题中，变量之间往往是具有相关关系，完全独立是不现实的。赵俊光^[23]提出了改进后的 SD 检验方法，该方法表明，如果变量之间存在相关性，合理有效的利用好相关性信息，不仅可以增加检验的功效，还可以使检验结果更加贴合实际。

而且 SD 检验方法需要假定总体的分布为正态分布，而在很多情况下，并不知道总体的分布或者总体分布不是正态分布，这个时候该方法便不再适用了。

1.2.3 BS 检验方法

基于上述方法的不足点，Bai 和 Saranadasa 提出了新的检验方法。该方法针对高维数据是有效的，并且不需要假定总体的分布是正态分布。一些模拟结果表明，即使在中等维度大小和样本量的情况下，BS 检验方法比 Hotelling 检验更有效。检验统计量如下所示：

$$M = (\bar{X}_m - \bar{Y}_n)^T (\bar{X}_m - \bar{Y}_n) - \tau \text{tr}(S), \quad (1-23)$$

其中

$$\begin{aligned} \bar{X}_m &\equiv (\bar{X}_{m,1}, \dots, \bar{X}_{m,p})^T = \frac{1}{m} \sum_{i=1}^m \mathbf{X}_i, \bar{Y}_n \equiv (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,p})^T = \frac{1}{n} \sum_{j=1}^n \mathbf{Y}_j, \\ S &= \frac{1}{m+n-2} \left(\sum_{i=1}^m (\mathbf{X}_i - \bar{X}_m)(\mathbf{X}_i - \bar{X}_m)^T + \sum_{j=1}^n (\mathbf{Y}_j - \bar{Y}_n)(\mathbf{Y}_j - \bar{Y}_n)^T \right). \end{aligned}$$

可以看到检验统计量 M 中有对两个样本均值向量之差的模的平方的估计，表示为 $\|\bar{X}_m - \bar{Y}_n\|^2$ 。在 $\|\bar{X}_m - \bar{Y}_n\|^2$ 的基础上剪掉一项 $\tau \text{tr}(S)$ ，使其满足估计的无偏性，在原假设成立下，有 $EM = 0$ 。

对总体作如下假设^[23]：

(I) $\mathbf{X}_i = \Gamma_i z_i + \mu_X, i = 1, \dots, m, \mathbf{Y}_j = \Gamma z_j + \mu_Y, j = 1, \dots, n$ 。这里的 Γ 是一个 $p \times l$ 矩阵 ($l \leq \infty$)，满足 $\Gamma \Gamma^T = \Sigma$ ， z_i 和 z_j 是两列随机序列，并且序列中包含 l 个 *iid* 的分量。并且其中各个相互独立的序列满足 $Ez_i = Ez_j = 0$ ， $\text{Var}(z_i) = \text{Var}(z_j) = I_m$ 。令 $z_i = (z_{i1}, \dots, z_{il})^T$ ， $z_j = (z_{j1}, \dots, z_{jl})^T$ ，假设 $Ez_{ik}^4 = Ez_{jk}^4 = 3 + \Delta < \infty$ 以及当 $v_1 + \dots + v_4 = 4$ 时，如果最少有一个 $v_k = 1$ (或有两个 v_k 值为 2)，则 $E \prod_{k=1}^l z_{ik}^{v_k} = E \prod_{k=1}^l z_{jk}^{v_k} = 0$ (或 1)；

(II) $p/N \rightarrow y > 0$ 以及 $m/N \rightarrow \kappa \in (0, 1)$ ，其中 $N = m + n$ ；

(III) $\mu^T \Sigma \mu = o(\tau \text{tr}(\Sigma^2))$ 和 $\lambda_{\max} = o(\sqrt{\text{tr}(\Sigma^2)})$ 是成立的，其中

$$\tau = (m+n)/(mn), \quad \mu = \mu_X - \mu_Y.$$

Bai 和 Saranadasa^[2]证明当总体满足假设 (I) - (III), 在原假设成立的情况下, 有

$$Z = \frac{M}{\tau \sqrt{2 \left(1 + \frac{1}{n} \text{tr}(\Sigma^2) \right)}} \rightarrow N(0, 1). \quad (1-24)$$

由于 $\text{tr}(\Sigma^2)$ 未知, 为了明确检验统计量, 需要寻找 $\text{tr}(\Sigma^2)$ 的比例相合估计, 并带入 Z 中。但和经典模式下的估计不同, 当维数很大时, $\text{tr}(S^2)$ 既不是 $\text{tr}(\Sigma^2)$ 的无偏估计, 也不是其比例相合估计。因此在满足三个假设 (I) - (III), 文章找到了

$$B^2 = \frac{N^2}{(N+2)(N-1)} \left[\text{tr}(S^2) - \frac{1}{N} (\text{tr}(S))^2 \right], \quad (1-25)$$

是 $\text{tr}(\Sigma^2)$ 的比例相互估计。把 B^2 带入公式 (1-24) 得 BS 方法的检验统计量:

$$T_{BS} = \frac{\frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)^T (\bar{X}_m - \bar{Y}_n) - \tau \text{tr}(S)}{\sqrt{\frac{2(N+1)}{N} B}} \rightarrow N(0, 1). \quad (1-26)$$

因此, 假设检验的判别条件为当 $T_{BS} > Z_\alpha$ 时拒绝原假设。其势能函数为

$$\beta_{BS}(\mu_X - \mu_Y) = \Phi \left(-Z_\alpha + \frac{N\kappa(1-\kappa) \|\mu_X - \mu_Y\|^2}{\sqrt{2 \text{tr}(\Sigma^2)}} \right) \rightarrow 0. \quad (1-27)$$

Bai 和 Saranadasa 检验方法也存在一些不足之处, 首先该检验虽然不需假定总体分布为正态分布, 但该检验依赖于四阶矩 $\|\bar{X}_m - \bar{Y}_n\|^2$ 的存在, 相对来说是不稳定的, 而且四阶矩计算起来也是比较复杂的。其次为了寻找已知可计算的检验统计量, 需要明确 $\text{tr}(\Sigma^2)$ 的比例相合估计, 但对于高维数据, 很多经典的估计也不再适用, 因此不能用传统的思想去思考, 要用新的理论去证明, 这也是一大难点。并且最后得到的检验统计量形式较为复杂, 计算负担大。

1.2.4 CQ 检验方法

随着对高维数据两样本均值研究的不断深入, Chen 和 Qin(2010)^[3]提出了一种新的方法, 该方法对解决此类高维小样本的数据显得尤为重要。在原假设成立的情况下, 且总体满足如下条件:

- (I) $X_i = \Gamma z_i + \mu_X, i = 1, \dots, m, Y_j = \Gamma z_j + \mu_Y, j = 1, \dots, n$ 。这里的 Γ 是一个 $p \times l$ 矩阵 ($l \leq \infty$), 满足 $\Gamma \Gamma^T = \Sigma$, z_i 和 z_j 是两列含有 l 个 iid 分量的随机序列, 并且各

个相互独立的序列满足 $Ez_i = Ez_j = 0$, $\text{Var}(z_i) = \text{Var}(z_j) = I_m$ 。令 $z_i = (z_{i1}, \dots, z_{il})^T$, $z_j = (z_{j1}, \dots, z_{jl})^T$, 假设 $Ez_{ik}^4 = Ez_{jk}^4 = 3 + \Delta < \infty$ 以及当

$$E(z_{ih_1}^{\alpha_1} z_{ih_2}^{\alpha_2} \dots z_{ih_q}^{\alpha_q}) = E(z_{ih_1}^{\alpha_1}) E(z_{ih_2}^{\alpha_2}) \dots E(z_{ih_q}^{\alpha_q}),$$

$$E(z_{jh_1}^{\alpha_1} z_{jh_2}^{\alpha_2} \dots z_{jh_q}^{\alpha_q}) = E(z_{jh_1}^{\alpha_1}) E(z_{jh_2}^{\alpha_2}) \dots E(z_{jh_q}^{\alpha_q}),$$

q 为给定的正整数, 使得 $\sum_{r=1}^q \alpha_r \leq 8$ 且 $h_1 \neq h_2 \neq \dots \neq h_q$;

(II) $m/N \rightarrow \kappa \in (0,1)$, 其中 $N=m+n$;

(III) $\text{tr}(\Sigma_1 \Sigma_2 \Sigma_3 \Sigma_4) = o\left\{\text{tr}\left[(\Sigma_X + \Sigma_Y)^2\right]\right\}$, 其中 $\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$ 表示 Σ_X 或者 Σ_Y ;

(IV) $(\mu_X - \mu_Y)^T \Sigma_X (\mu_X - \mu_Y) = (\mu_X - \mu_Y)^T \Sigma_Y (\mu_X - \mu_Y) = o\left\{N^{-1} \text{tr}\left[(\Sigma_X + \Sigma_Y)^2\right]\right\}$ 。

Chen 和 Qin(2010)提出了下面的检验统计量

$$T_C = \frac{1}{m(m-1)} \sum_{i \neq j}^m X_i^T X_j + \frac{1}{n(n-1)} \sum_{i \neq j}^n Y_i^T Y_j - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n X_i^T Y_j. \quad (1-28)$$

可以看到 Bai 和 Saranadasa 检验统计量是在 $\|\bar{X}_m - \bar{Y}_n\|^2$ 的基础上减掉 $\tau \text{tr}(S)$, 而 Chen 和 Qin 检验统计量是在 $\|\bar{X}_m - \bar{Y}_n\|^2$ 的基础上减掉 $\left(\sum_{i=1}^m X_i^T X_i + \sum_{i=1}^n Y_i^T Y_i\right)$ 。

作者证明了当 $p \rightarrow \infty$, $n \rightarrow \infty$ 时, 该检验统计量渐进服从于正态分布

$$\frac{T_c}{\sqrt{\frac{2}{m(m-1)} \text{tr}(\Sigma_X^2) + \frac{2}{n(n-1)} \text{tr}(\Sigma_Y^2) + \frac{4}{mn} \text{tr}(\Sigma_X \Sigma_Y)}} \rightarrow N(0,1). \quad (1-29)$$

接着找到了 $\text{tr}(\Sigma_X^2)$, $\text{tr}(\Sigma_Y^2)$ 和 $\text{tr}(\Sigma_X \Sigma_Y)$ 的比例相合估计 $\text{tr}(\Sigma_X^2)$, $\text{tr}(\Sigma_Y^2)$ 和 $\text{tr}(\Sigma_X \Sigma_Y)$, 带入得到了已知可计算的检验统计量:

$$T_{Cq} = \frac{T_c}{\sqrt{\frac{2}{m(m-1)} \text{tr}(\Sigma_X^2) + \frac{2}{n(n-1)} \text{tr}(\Sigma_Y^2) + \frac{4}{mn} \text{tr}(\Sigma_X \Sigma_Y)}} \rightarrow N(0,1). \quad (1-30)$$

当 $T_{Cq} > z_\alpha$ 时, 拒绝原假设。功效函数为:

$$\beta_{Cq}(\|\mu_X - \mu_Y\|) = \Phi\left(-z_\alpha + \frac{N\kappa(1-\kappa)\|\mu_X - \mu_Y\|^2}{\sqrt{2 \text{tr}(\tilde{\Sigma}(\kappa)^2)}}\right) \rightarrow 0. \quad (1-31)$$

其中 $\tilde{\Sigma}(\kappa) = (1-\kappa)\Sigma_X + \kappa\Sigma_Y$ 。

首先 Chen 和 Qin 检验方法在满足一定的条件下解决了高维数据带来的问题,

且不需要假定总体分布为正态分布,这一优点使得该方法的应用面变广。并且检验统计量中也不含有四阶矩,较为稳定。

不足之处就是计算过程较为繁琐,在明确原假设成立下的渐进分布时,需要对 $\text{tr}(\Sigma_X^2)$, $\text{tr}(\Sigma_Y^2)$ 和 $\text{tr}(\Sigma_X \Sigma_Y)$ 都进行比例相合估计,这一步骤需花费一定的时间和计算空间,即不易确定原假设成立下的渐进分布。总的来说,在总体满足一定的条件下,CQ 检验方法对高维数据进行假设检验时实为一个不错的方法。

1.2.5 其他检验方法

近年来,一些统计研究员在高维数据场景中越来越频繁的运用经验似然方法。Hjort, Mckeague 和 Keilegom (2009)^[24]对经验似然比统计量求其渐进分布,在满足一定条件下,证实渐进分布为正态分布; Peng 和 Qi(2013)^[25]进一步研究讨论了经验似然方法受样本维数的影响有哪些; Li, Lin 和 Zhu (2012)^[26] 为了提高检验的功效,提出一系列方法改进经验似然方法; Peng, Qi 和 Wang(2014)^[27]在经验似然方法的基础上对不同的问题做出了解释说明。

上文提到的 Hotelling 检验统计量随着样本维数的增大,检验统计量 T^2 没有定义。因此对总体协方差矩阵的估计成了关键问题之一,前人也提出了很多在高维情形下对总体协方差矩阵的估计方法。在 Chen, Paula, Prenticea, Wang(2011)^[28], Shen, Lin 和 Zhu(2011)^[29]这些文献中,用的最多的是压缩估计,原理是利用总体协方差矩阵可以求逆的估计量替换 Hotelling 检验统计量中的样本协方差矩阵。

而且我们在现实研究中会遇到很多这样的情况, $\bar{X}_m - \bar{Y}_n$ 的很多分量都是 0,只有少数分量不为 0,即 $\bar{X}_m - \bar{Y}_n$ 具有稀疏性。Cai, Liu 和 Xia(2014)^[4]针对如上这样的稀疏性备择假设,提出了检验方法。

除了上面介绍的检验方法,还有一些其他的检验方法,在此不一一列举,见参考文献^[30-34]。

1.3 本文的主要贡献与创新

本论文以基于置换检验对高维数据两样本均值的假设检验新方法为重点研究内容,主要创新点与贡献如下:

- 1.不需要明确总体分布或假定总体分布为正态分布。对高维数据两样本均值的假设检验,前人也做出较多研究,但大部分都需明确总体分布或假定总体分布为正态分布。在实际情况中,总体分布经常是未知或不易确定的,因此这些方法便不再适用。本文利用置换检验对高维数据两样本均值进行假设检验,置换检验利用一群算子对样本数据进行变换,形成新的样本数据,通过新样本数据计算统计量值来构

成经验分布函数,用经验分布函数代替原假设成立下的渐进分布进行检验。故不需要明确总体的分布或假定总体分布为正态分布,应用面更广。

2.降低计算成本。置换检验基于传统的 Hotelling's T^2 检验、以及在 Chuang and Romano(2016)中提到的改进后的 Hotelling's T^2 检验,这些方法由于要计算高维协方差矩阵的逆,计算工作很是繁重,不好实施,并且应用于置换检验时,需多次重复置换计算高维逆矩阵,使得计算任务再次加重。基于边缘标准化统计量的置换检验,该检验统计量仅需计算协方差矩阵的对角线元素,这大大减少了计算量。

3.和其他检验方法相比较具有更好的 P 值和势能。将置换检验与其他现存的检验方法进行比较,发现本文提出的新检验方法具有更好的 P 值和势能。但这种方法要保证置换检验的一致有效性,两总体或样本需要满足条件: $\Sigma_X \propto \Sigma_Y$ 或 $m/n \rightarrow 1$ 。结合 Bootstrap 方法,本文通过新的构造方式,构造伪样本使其满足条件 $m = n$,代入伪样本计算检验统计量,证实基于边缘标准化统计量的置换检验是一致有效的。

1.4 研究内容和结构安排

本文一共分为五章,每一章的研究工作如下:

第一章为引言部分,首先论述了高维数据的发展和相关统计理论知识,对研究的背景和研究意义做出说明。其次介绍了传统的 Hotelling 检验方法、SD 检验方法、BS 检验方法、CQ 检验方法的国内外研究现状以及检验统计量的形式和检验的优缺点,并提出本文基于边缘标准化统计量的置换检验对高维数据两样本均值假设检验的研究方法和创新点。

第二章主要介绍了基于置换检验对两样本均值的假设检验问题。首先介绍了置换检验的相关定义、关键点以及优缺点;其次给出了基于置换检验对一维数据两样本均值、多维数据两样本均值、高维数据两样本均值的假设检验方法,包括检验的统计量及置换样本的方式,并与其他检验方法进行比较;然后重点介绍了对高维数据的研究,介绍了基于样本之差检验统计量、极大值检验统计量、改进的 Hotelling's T^2 统计量以及 Kong et al(2020)^[35]提出的边缘标准化统计量的置换检验方法及方法对应的一致有效性条件。

第三章是本文的重点研究内容,介绍了利用 Bootstrap 方法构造基于边缘标准化统计量置换检验的一致有效性条件。首先介绍了 Bootstrap 方法的定义及原理;其次给出了边缘标准化统计量的定义以及与改进后的 Hotelling's T^2 统计量相比较的优缺点;然后证明了基于边缘标准化统计量置换检验的一致有效性条件为两总体方差相等或两样本数相同;最后通过 Bootstrap 方法产生两个大小一样的伪样本,满足一致性条件,并将伪样本带入到检验统计量中形成置换分布。

第四章主要研究两个模拟实验以及一个实际案例。第一个模拟实验,计算本文提出的 $S_1^*(\cdot)$ 统计量的相关置换检验对第一类错误的控制,并与现存的 CQ(2010)、XLWP(2016)的检验方法进行比较;第二个模拟实验,计算 $S_1^*(\cdot)$ 统计量在备择假设成立下的经验势能,将其与 XLWP(2016)检验方法的经验势能进行比较,得出结论;最后在实际案例中对我国 A 股市场不同行业的股票月收盘价格进行研究,判断 6 个高维数据行业在两个不同时间段的平均月收益率是否有差异^[43]。

第五章对全文的内容进行了总结,对于不足之处,提出了下一步的研究计划和方向。

第二章 基于置换检验对两样本均值的假设检验

第二章主要介绍了基于置换检验对两样本均值的假设检验问题。首先介绍了置换检验的相关定义、关键点以及优缺点；其次给出了基于置换检验对一维数据两样本均值、多维数据两样本均值、高维数据两样本均值的假设检验方法，包括检验的统计量及样本置换方式，并与其他检验方法进行比较；然后重点介绍了对高维数据的研究，介绍了基于样本之差检验统计量、极大值检验统计量、改进的 Hotelling's T^2 统计量以及 Kong et al(2020)^[35]提出的边缘标准化统计量的置换检验方法及方法对应的一致有效性条件。

2.1 置换检验

2.1.1 置换检验的定义

置换检验是一种统计的检验方法，但之前由于计算量过大一直都没有得到广泛的关注。随着计算机的普及和运算效率的提高使得置换检验逐步引起统计研究人员的重视^[44]。基因表达这类高维数据的产生促使人们开始研究置换检验，这也使其应用越来越广泛。

该检验方法是利用一群算子对样本数据进行变换，形成新的样本数据，新的样本数据称为置换样本。通过新样本数据计算统计量值来构成的经验分布函数，也称为检验统计量的置换分布。最原始在置换前的样本记为初始样本，定义初始值为初始样本带入检验统计量中计算得到的值。初始值和经验分布函数共同构成了置换检验。

2.1.2 置换检验的关键点

在进行置换检验时，有几个关键点需要注意：

首先是置换的方式，在面对不同的情形，置换的方式也是不同的。对于单总体，设原假设 $H_0: \mu = \mu_0$ ，如果总体是对称分布，置换样本可以通过随意置换 $X - \mu_0$ 每个观测值正负号^[16]得到。对于两总体，当两个总体的方差相等时，设原假设为 $H_0: \mu_X = \mu_Y$ ，则可以将总体 X 的样本置换到总体 Y 的样本中，总体 Y 的样本置换到总体 X 的样本中。

其次是置换后的样本数量。置换 n 个观测值不为零的样本符号，可以得到 2^n 个不同的新样本。如果有两个样本，样本量分别为 m ， n ，置换后一共可以得到 C_{m+n}^m 个不同的置换样本。所以当置换后样本数过大时，可以抽取部分置换样本不用全部

抽取, 将其带入检验统计量中, 形成经验分布函数, 这叫做不完全样本的置换检验。在显著性水平 $\alpha = 0.05$ 下, Marco(2004)^[16]通过一系列的蒙特卡罗模拟实验证实, 在估计检验势能时, 只需大概 500 个随机样本就足够了。因此如果当置换样本数量很大时, 只需要从中随机抽取大约 500-1000 个样本即可。这样不仅可以节省计算的时间和空间, 还能模拟出置换分布。

然后, 给出置换检验的拒绝域。发现如果初始值在经验分布函数很极端的位置上, 那么有理由拒绝置换检验的原假设。证实了经验分布函数的分位数可当作临界值决定置换检验的拒绝域。置换检验有三种常用拒绝域^[16], 当显著性水平 $\alpha=0.05$ 时, 假定 S 为检验统计量, t 为 S 的初始值, 其中 S_α 满足 $P(S \leq S_\alpha) = \alpha$, $S_{1-\alpha}$ 满足 $P(S \geq S_{1-\alpha}) = \alpha$ (其中 $\alpha < 0.5$)。:

- (1) 右侧拒绝域: $\{S \geq S_{1-\alpha}\}$, 检验的 p 值为 $p = P(S \geq t)$
- (2) 左侧拒绝域: $\{S \leq S_\alpha\}$, 检验的 p 值为 $p = P(S \leq t)$
- (3) 双侧拒绝域: $\{S \leq S_{\alpha/2}\} \cup \{S \geq S_{1-\alpha/2}\}$, 检验的 p 值为

$$p = \begin{cases} 2P(S \geq t), & P(S \geq t) \leq 0.5 \\ 2P(S \leq t), & P(S \leq t) \leq 0.5 \end{cases} \quad (2-1)$$

2.1.3 置换检验的优缺点

已有的研究表明, 置换检验方法具有如下优点:

1.对总体分布自由。一些经典的假设检验方法需要明确在原假设成立下的渐进分布, 但当总体分布无法确定或者不是理想的正态分布时, 该渐进分布不易找到。而置换检验是用经验分布函数代替原假设成立下的渐进分布进行检验, 所以对总体分布自由且无需假定其正态性, 应用较为广泛。

2.具有一定的稳健性。由于置换检验对总体分布没有太多严苛的约束条件, 较贴合实际情况且不需要对统计相关理论做出过分理想化的假定。如果存在偏离很大的数据, 置换检验也不会太过于敏感, 即不会受到较大的影响, 具有一定的稳健性。

3.检验性能较高。置换检验充分利用样本数据的信息, 因此当样本数量较大时, 置换检验得到的结果与经典的参数检验近似; 当样本数量较少时, 检验的性能要优于参数检验。

但置换检验也有不足之处, 首先基于置换检验原理, 通过置换后样本数据计算统计量值来构成的经验分布函数, 用经验分布函数代替原假设成立下的渐进分布进行检验。如果经验分布和原假设成立下的渐进分布相同, 则说明该置换检验是一致有效的, 即可在有限的样本内控制犯第一类错误的概率。如果经验分布函数与原

假设成立下的渐进分布不同, 则不是一致有效的, 不能使用该检验方法。所以在进行检验之前, 需要判断总体与样本是否满足检验的一致性条件, 在不满足的情况下, 找到方法使其满足。例如对高维数据两样本均值的假设检验中, 两样本观测数分别为 m 和 n , 两总体方差分别为 Σ_X 和 Σ_Y , 置换检验的一致性条件为: $\Sigma_X \propto \Sigma_Y$ 或 $m/n \rightarrow 1$ 。其次, 由于置换后样本数变多, 再带入检验统计量, 这也增加了计算的时间与空间。

总的来说, 我们一般平时较为常用的检验都属于有参检验, 但其要求样本满足近似正态、无离群点、数据量大等条件。而实际问题中很难满足以上前提条件, 这时需使用非参检验。非参检验只需关注数据的秩, 但有时也无法处理一些样本量较少的情况, 这种情况便可以考虑置换检验, 例如高维数据的假设检验问题。

2.2 基于置换检验对一维数据两样本均值的假设检验

2.2.1 一维数据两样本均值的置换检验

假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立且服从 F_1 分布, 样本 $\{Y_j, j=1, \dots, n\}$ 相互独立且服从 F_2 分布, 并且都是一维的随机变量。 μ_X 和 μ_Y 分别是 X 和 Y 的均值, 检验两总体均值的假设为

$$H_0: \mu_X = \mu_Y \quad H_1: \mu_X \neq \mu_Y (\mu_X > \mu_Y, \mu_X < \mu_Y)$$

对一维数据两样本均值的检验, 其基本思想是, 满足两方差相等且在原假设成立的情况下, 可以将两个样本进行合并, 当作来自同一个总体。认为前 m 个样本来自 X 总体, 后 n 个样本来自 Y 总体, 对合样本中样本的位置进行置换得到新样本。重复置换这一步骤, 得到了 C_{m+n}^m 个新的置换样本。选择置换检验的检验统计量为

$$S = \bar{X} - \bar{Y}. \quad (2-2)$$

设样本均值 $\bar{X} = \sum_{i=1}^m X_i$, $\bar{Y} = \sum_{j=1}^n Y_j$ 。当然检验统计量的选择不唯一, 可以选择形式简单且符合实际情况的检验统计量。将置换后的新样本带入检验统计量中, 形成经验分布函数, 用经验分布函数代替原假设成立下的渐进分布。确定初始值以及拒绝域, 做出假设检验的判断。

2.2.2 一维数据两样本均值的其他检验方法

在前人的研究中, 存在其他对一维数据两样本均值的检验方法, 有参数检验方法和非参数检验方法。当总体分布为正态分布时, 传统的检验方法有 U 检验和 t 检验, 这两种方法最为常用。Welch 提出了 Welch t 统计量, 对解决两总体方差未知

并且不相等的检验问题有很大的帮助, 但该统计量在使用时也有限制条件, 在 $\min(m, n) < 6$ 时, 该检验方法并不适用。Xu(2009)^[16]对 Welch t 统计量做出了改进, 在样本数量固定且有限的情况下, 用 Welch t 统计量的功效渐进展开式进行检验。Wilcoxon 秩和检验具有稳健性, 不受离群点数据的影响, 是一种常用的检验方法。而自助检验方法当样本来自有相同位置参数的分布时, 是较为精确的, 但当样本量过小时, 也并不太适用。对以上四种检验方法进行比较, 可以得出^[16]:

(1) 置换检验不需要假定总体的分布, 通常情况下和 Welch t 检验和 Wilcoxon 秩和检验相比, 经验检验势能较高, 而且也能很好的控制犯第一类错误的概率;

(2) 自助检验方法虽然相比其他方法而言, 具有更高的检验势能, 但不能很好的控制犯第一类错误的概率, 比其他检验方法会犯更大的第一类错误;

(3) Welch t 检验和 Wilcoxon 秩和检验在势能方面的表现较为保守, 不好也不坏, 而且这两种方法对位置参数没那么敏感。

2.3 基于置换检验对多维数据两样本均值的假设检验

2.3.1 多维数据两样本均值的置换检验

多维两样本均值的置换检验类似于一维, 这里对多维的定义是样本维数 p 大于 1 小于两个总体的样本数 m 和 n 。假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立同服从 F_1 分布, 样本 $\{Y_j, j=1, \dots, n\}$ 相互独立同服从 F_2 分布, 并且都是 p 维的随机变量。 $EX_i = \mu^X$, $EY_j = \mu^Y$, $\text{Cov}(X_i) = \Sigma_i^X$, $\text{Cov}(Y_j) = \Sigma_i^Y$, $N = m + n$ 。提出两总体均值的检验假设

$$H_0: \mu_X = \mu_Y \quad H_1: \mu_X \neq \mu_Y (\mu_X > \mu_Y, \mu_X < \mu_Y)$$

多维两样本数据置换时, 应当注意置换时把 p 维向量看成一个整体。置换的方式还是选择将两个样本合并成一个样本, 总体 X 的样本置换到总体 Y 的样本中, 总体 Y 的样本置换到总体 X 的样本中。选择检验统计量为

$$S = \max_{\forall 1 \leq p \leq l (1 < l < m, n)} |\bar{X}_{mp} - \bar{Y}_{np}|. \quad (2-3)$$

其中 $\bar{X}_m \equiv (\bar{X}_{m,1}, \dots, \bar{X}_{m,p})^T = \frac{1}{m} \sum_{i=1}^m X_i$, $\bar{Y}_n \equiv (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,p})^T = \frac{1}{n} \sum_{j=1}^n Y_j$, 这里的检验统计

量选择也是不唯一的, 将置换后的新样本带入检验统计量, 形成经验分布函数, 确定拒绝域及初始值, 做出假设检验的判断。

2.3.2 多维数据两样本均值的其他检验方法

对于多维数据两样本均值的假设检验, 当总体分布为正态分布时, 经典的参数

检验方法有两协方差矩阵相等的 Hotelling's T^2 检验, 还有当两协方差矩阵不相等但两样本数 $m = n$ 时作为成对数据处理的方法; 还有当两样本数也不相等的处理方法。但最常用的还是置换检验和 Hotelling's T^2 检验, 因此将这两种方法进行比较:

Hotelling's T^2 检验方法在原假设成立的情况下, 当两个总体都是 p 元正态分布, 并且协方差矩阵满足 $\Sigma_X = \Sigma_Y$, Hotelling 检验统计量为:

$$T^2 = \frac{mn}{m+n} (\bar{X}_m - \bar{Y}_n)^T S^{-1} (\bar{X}_m - \bar{Y}_n)$$
 T^2 统计量在原假设成立下的渐进分布可以由中心化的 F 分布给出

$$T^2 \sim \frac{(m+n-p-1)}{(m+n-2)p} F_{(p, m+n-p-1)}$$

其中 $F_{(p, m+n-p-1)}$ 服从自由度 p 和 $(m+n-p-1)$ 的 F 分布, 且显著性水平为 α 的拒绝域为 $\{F \geq F_\alpha(p, m+n-p-1)\}$ 。由此可以看出, 在正态分布总体下, Hotelling's T^2 检验是精确检验, 而且检验流程清晰易懂。但当总体分布不满足正态性假定时, Hotelling's T^2 检验便不是最佳选择, 检验结果会出现偏差。当样本维数过大时, 即 $m, n > p$ 不成立时, Hotelling's T^2 检验统计量无意义, 该检验方法无法应用。

置换检验属于非参检验, 即不需要对总体进行正态性假定, 而且检验统计量可以有多种选择。所以在一些总体分布较为复杂且协方差矩阵不易计算或不相等时, 可以考虑使用置换检验。

2.4 基于置换检验对高维数据两样本均值的假设检验

高维数据的出现使得很多经典的多元统计方法变得不再适用, 也给传统的统计理论带来了新的挑战, 高维的定义是样本维数 p 远大于两个总体的样本数 m 和 n 。

并且一些经典的假设检验方法需要明确在原假设成立下的渐进分布, 但当总体分布未知或不是正态分布时, 该渐进分布不易找到。所有基于置换检验对高维数据两样本均值的假设检验运用而生, 部分统计研究者对其进行了深入的研究, 他们在原有经典方法的基础上提出了基于置换检验的新方法, 经验证这些方法对高维数据的假设检验是可行且一致有效的, 并且检验效果与其他现存的对高维数据均值检验方法相比, 效果更优。

但置换检验也有不足之处, 为了满足其有效性, 样本或总体需满足一定的条件。在 Chung and Romano(2016)^[18], 详细讨论了置换检验的有效性条件。由于置换检验的基本思想是用经验分布函数代替原假设成立下的渐进分布, 那置换检验是否有效即判断经验分布函数与渐进分布是否相等, $G(X)$ 假定代表置换样本带入检验统计量后形成的经验分布函数, $F(X)$ 代表原假设成立下的渐进分布函数, 有效性条

件为

$$\sup_{\forall X \in R} |F(X) - G(X)| \stackrel{P}{\rightarrow} 0. \quad (2-4)$$

公式(2-4)表示,对任意的 $X \in R$, $F(X)$ 和 $G(X)$ 的柯尔莫哥洛夫距离依概率趋于0,则说明该置换检验是一致有效的。

基于不同的检验统计量,该公式的表现形式即所需有效性条件会有所差异。下面介绍基于置换检验对高维数据两样本均值假设检验时,几种方法的检验统计量、检验有效性条件和优缺点等。

2.4.1 基于样本之差统计量的置换检验

2.4.1.1 检验统计量

假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立源自 X 总体,样本 $\{Y_j, j=1, \dots, n\}$ 相互独立源自 Y 总体,并且都是 p 维的随机变量。其中 X 总体的均值向量为 μ_X ,协方差矩阵为 Σ_X ;总体 Y 的均值向量为 μ_Y ,协方差矩阵为 Σ_Y 。令 $N=m+n$,定义 $Z=(Z_1, \dots, Z_N)=(X_1, \dots, X_m, Y_1, \dots, Y_n)$ 。设原假设为 $H_0: \mu_X = \mu_Y$,提出检验统计量如下所示:

$$T_{m,n} = (T_{m,n,1}, \dots, T_{m,n,p}) = m^{1/2} [\bar{X}_m - \bar{Y}_n]. \quad (2-5)$$

该检验统计量是基于两样本均值稠密差异提出的,均值稠密差异是指两样本均值差异向量中,有多个值不为0。根据Chung and Romano(2016)^[18]的证明,有定理2.1,即基于样本之差统计量的置换检验的经验分布函数服从 p 元正态分布,其中均值为0,协方差矩阵为 $\bar{\Sigma} = \frac{d}{1-d} \Sigma_X + \Sigma_Y$,其中 $d = \lim \frac{m}{m+n}$ 。基于检验统计量

(2-5)置换检验的经验分布函数

$$\hat{R}_{m,n}^T(t) = \frac{1}{N!} \sum_{\pi \in G_N} I\{T_{m,n}(Z_{\pi(1)}, \dots, Z_{\pi(N)}) \leq t\}. \quad (2-6)$$

其中 G_N 表示所有的置换方式集。

定理 2.1 假定 $E(X)=E(Y)$,并且 $0 < \text{Var}(X_i) < \infty$, $0 < \text{Var}(Y_j) < \infty$,令 $m \rightarrow \infty$, $n \rightarrow \infty$,设 $\bar{\Sigma}$ 为正定矩阵。 $T_{m,n}$ 的经验分布函数满足

$$\sup_{t \in R^d} |\hat{R}_{m,n}^T(t) - G(t)| \stackrel{P}{\rightarrow} 0, \quad (2-7)$$

其中 $G(t)$ 表示 p 元正态分布且均值为0,协方差矩阵为 $\bar{\Sigma} = \frac{d}{1-d} \Sigma_X + \Sigma_Y$ 。

2.4.1.2 置换检验的有效性条件

根据定理 2.1, 置换检验的经验分布函数 $G(t)$ 表示 p 元正态分布且均值为 0 , 协方差矩阵为 $\bar{\Sigma} = \frac{d}{1-d}\Sigma_X + \Sigma_Y$ 。而在原假设成立的情况下, 检验统计量 $T_{m,n}$ 的无条件抽样分布函数即渐进分布函数 $F(t)$, 服从 p 元正态分布且均值向量为 0 , 协方差矩阵为 $\bar{\bar{\Sigma}} = \Sigma_X + \frac{d}{1-d}\Sigma_Y$ 。

$$\begin{aligned} G(t) &\sim N\left(0, \frac{d}{1-d}\Sigma_X + \Sigma_Y\right), \\ F(t) &\sim N\left(0, \Sigma_X + \frac{d}{1-d}\Sigma_Y\right). \end{aligned} \quad (2-8)$$

比较这两个分布之间差异可知, 置换检验的经验分布函数 $G(t)$ 和原假设成立下的渐进分布函数 $F(t)$, 具有相同的均值向量且都为 0 , 协方差矩阵在表现形式上有所差异。要使得基于样本之差 $T_{m,n}$ 的置换检验是有效的, 则这两个分布应该完全相同, 协方差矩阵 $\bar{\Sigma} = \bar{\bar{\Sigma}}$, 即 $\frac{d}{1-d}\Sigma_X + \Sigma_Y = \Sigma_X + \frac{d}{1-d}\Sigma_Y$ 。等价于 $\Sigma_X \propto \Sigma_Y$ 或 $m/n \rightarrow 1$, 即为两总体方差相等或两样本数相等。因此, 在使用基于样本之差统计量的置换检验时, 要满足以上条件之一, 要么使其总体方差相等, 要么使其两样本数相等。如果这两个条件都无法满足, 则不能使用置换检验对两样本均值进行假设检验。

2.4.2 基于极大值统计量的置换检验

2.4.2.1 检验统计量

针对两样本均值稀疏差异提出了基于极大值检验统计量的置换方法, 均值稀疏差异是指两样本均值差异向量中, 有多个值为 0 。假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立源自 X 总体, 样本 $\{Y_j, j=1, \dots, n\}$ 相互独立源自 Y 总体, 并且都是 p 维的随机变量。令 $N=m+n$, 设原假设为 $H_0: \mu_X = \mu_Y$, 提出极大值检验统计量:

$$M_{m,n} = \max_{k=1, \dots, p} \left(|T_{m,n,k}| \right). \quad (2-9)$$

这里的 $T_{m,n,k}$ 表示 $T_{m,n}$ 中的第 k 个部分。基于检验统计量 (2-9) 置换检验的经验分布函数

$$\hat{R}_{m,n}^M(t) = \frac{1}{N!} \sum_{\pi \in G_N} \mathbf{I} \left\{ M_{m,n} \left(Z_{\pi(1)}, \dots, Z_{\pi(N)} \right) \leq t \right\}. \quad (2-10)$$

定理 2.2 Chung and Romano(2016)^[18] 证明了基于极大值统计量的置换检验的

经验分布函数满足

$$\sup_{t \in R} |\hat{R}_{m,n}^M(t) - M(t)| \xrightarrow{P} 0, \quad (2-11)$$

其中, $M(t)$ 是经验分布函数, 且满足

$$M(t) = \max_{k=1, \dots, p} (|G_1|, \dots, |G_p|), \quad (2-12)$$

这里的 (G_1, \dots, G_p) 服从多元正态分布, 和前文的 $G(t) \sim N\left(0, \frac{d}{1-d} \Sigma_X + \Sigma_Y\right)$ 分布相同。

2.4.2.2 置换检验的有效性条件

可以看到, 基于极大值统计量 $M_{m,n}$ 的经验分布函数, 依赖于协方差矩阵 $\bar{\Sigma}$, 没有去除对总体方差的影响。因此, 该方法无法有效的控制犯第一类错误的概率。对于稀疏均值差异的数据, Chuang 也提出的新的方法, 他通过构造新的检验统计量, 使其经验分布函数服从 $(0, 1)$ 上的均匀分布, 去除了方差的影响。

2.4.3 基于改进的 Hotelling's T^2 统计量的置换检验

之前提到了传统的 Hotelling's T^2 检验, 对于高维数据是不适用的, 并且还有总体正态性和方差齐性假定。因此, Chuang and Romano(2016)^[18]中提到了改进后的 Hotelling's T^2 检验。

2.4.3.1 检验统计量

当 $P > 1$ 时, 定义 $\delta_N = (\delta_{N,1}, \dots, \delta_{N,p})^T = N^{1/2}(\bar{X}_m - \bar{Y}_n)$, $\tilde{\Omega}_N$ 是 $\tilde{\Omega} = c^{-1}\Sigma_X + (1-c)^{-1}\Sigma_Y$ 的一致估计, 其中当 $m, n \rightarrow \infty$ 时, $m/N \rightarrow c$. 令 $h_N = \{\tilde{\Omega}_N\}^{-1/2} \delta_N$. 定义改进后 Hotelling's T^2 的检验统计量

$$H_N^2(Z^N) \equiv \|h_N\|_2. \quad (2-13)$$

其中, 对任意的向量 $v = (v_1, \dots, v_p) \in R^p$, $\|v\|_2 = (v_1^2 + \dots + v_p^2)^{1/2} / p^{1/2}$ 。基于检验统计量 (2-13) 置换检验的经验分布函数

$$\hat{R}_{m,n}^{H_N^2}(t) = \frac{1}{N!} \sum_{\pi \in G_N} \mathbf{I}\{H_N^2(Z_{\pi(1)}, \dots, Z_{\pi(N)}) \leq t\}. \quad (2-14)$$

其中 G_N 表示所有的置换方式集。定理证实 H_N^2 的经验分布函数满足

$$\sup_{t \in R} |\hat{R}_{m,n}^{H_N^2}(t) - \chi_k^2(t)| \xrightarrow{P} 0, \quad (2-15)$$

其中 $\chi_k^2(t)$ 表示自由度为 k 的卡方分布。

2.4.3.2 置换检验的有效性条件

改进后的 Hotelling's T^2 置换检验是一致有效的, 由于基于改进后 Hotelling's T^2 置换检验的经验分布函数服从自由度为 k 的卡方分布, 分布中不含有总体分布、样本数量的任何信息, 在对高维数据均值假设检验时, 此方法可以作为选择之一。

但该检验统计量也有不足之处, 可以看到检验统计量中含有 $h_N = \{\tilde{\Omega}_N\}^{-1/2} \delta_N$ 。在计算每一次检验统计量值时都需要计算高维协方差矩阵 $\tilde{\Omega}_N$ 的逆, 而且应用于置换检验, 随着置换样本的增多, 使得高维协方差矩阵也需要计算多次, 计算量繁重。花费较多的计算时间和计算成本。

2.4.4 基于边缘标准化统计量的置换检验

对于两样本高维数据的假设检验, 之前的文献提到了传统的 Hotelling's T^2 检验, 以及在 Chuang and Romano(2016)^[18]中提到的改进后的 Hotelling's T^2 检验。这些方法由于要计算高维协方差矩阵的逆, 计算工作很是繁重, 不好实施, 并且应用于置换检验时, 需多次重复置换计算高维逆矩阵, 使得计算任务再次加重。在 Kong et al(2020)中提出一种新的检验统计量, 该检验统计量仅需计算协方差矩阵的对角线元素, 这大大减少了计算量, 并且具有更好的 P 值和势能。但这种方法要保证置换检验的一致有效性, 两总体样本需要满足条件:

$$\Sigma_X \propto \Sigma_Y \text{ 或 } m/n \rightarrow 1. \quad (2-16)$$

因此作者提出通过构造伪样本的形式使得 m 和 n 相等, 利用伪样本数据计算统计量的值, 进行置换检验。而本文提出了一种新的构造样本数量 m 和 n 相等的方式, 发现检验效果较好。

2.4.4.1 检验统计量

该文提出的边缘标准化统计量是用协方差矩阵的对角矩阵 \tilde{D}_N 去代替 Hotelling's T^2 中的 $\tilde{\Omega}_N$, 这里 $\tilde{D}_N = \text{diag}(v_{N,k}^2, k=1, \dots, p)$,

$$v_{N,k}^2 = \frac{N}{m} \hat{\sigma}_{m,k}^2(X_1, \dots, X_m) + \frac{N}{n} \hat{s}_{n,k}^2(Y_1, \dots, Y_n). \quad (2-17)$$

而 $\hat{\sigma}_{m,k}^2(X_1, \dots, X_m)$, $\hat{s}_{n,k}^2(Y_1, \dots, Y_n)$ 分别是两样本协方差矩阵 $\hat{\sigma}_m^2(X_1, \dots, X_m)$, $\hat{s}_n^2(Y_1, \dots, Y_n)$ 的第 k 个对角线元素。对任意的 $\gamma > 0$, 给出边缘标准化统计量

$$S_N^\gamma(Z^N) \equiv S_N^\gamma(Z_1, \dots, Z_N) = \left\| \tilde{D}_N^{-1/2} \delta_N \right\|_\gamma = \left[p^{-1} \sum_{k=1}^p |\delta_{N,k} / \nu_{N,k}|^\gamma \right]^{1/\gamma}. \quad (2-18)$$

而基于该统计量的置换检验是有效的，需要满足 $\Sigma_X \propto \Sigma_Y$ 或 $m/n \rightarrow 1$ 。

2.4.4.2 构造两样本数相等

在 Kong et al(2020)中，通过构造伪样本数据，使得样本量 m 和 n 相等。构造方式如下：

假设 $m > n$ ，且对非负整数 K ， $k(0 \leq k \leq n)$ ，满足关系式 $m = K \times n + k$ 。定义 $X'_i = X_i - \mu^X$ ， $Y'_j = Y_j - \mu^Y$ ， $i = 1, \dots, m$ ， $j = 1, \dots, n$ 。在实际应用中，用样本均值 \bar{X}_m 去代替 μ^X 。如果 $k=0$ ，

$$X_i^* = \frac{n}{m} \sum_{j=(i-1)K+1}^{i \times K} X'_j, \quad i = 1, \dots, n; \quad (2-19)$$

如果 $k > 0$ ，将 X'_i 中剩余的部分加在 X_i^* 的第 K 项

$$X_i^* := X_i^* + \frac{n}{m} X'_{K \times n + i}, \quad i = 1, \dots, k. \quad (2-20)$$

将 $\{X_i^*, i = 1, \dots, m\}$ 和 $\{Y'_j, j = 1, \dots, n\}$ 定义为两个伪样本，因此后续的置换检验都是用伪样本去替代原始样本。

令 $Z^n = \{Z_1, \dots, Z_{2n}\}$ ，其中 $Z_i = X_i^*$ ， $1 \leq i \leq n$ ； $Z_{n+j} = Y'_j$ ， $1 \leq j \leq n$ ，和公式 (2-18) 类似，定义检验统计量

$$S_1^\gamma(Z^N) = \left[p^{-1} \sum_{k=1}^p |\delta_{n,k}^* / \nu_{n,k}^*|^\gamma \right]^{1/\gamma}. \quad (2-21)$$

这里的 $\delta_{n,k}^*$ 是 $\delta_n^* = n^{1/2} (\bar{X}^* - \bar{Y}^*)$ 的第 K 个元素，其中 $\bar{X}^* = \frac{1}{n} \sum_{i=1}^n X_i^*$ ， $\bar{Y}^* = \frac{1}{n} \sum_{i=1}^n Y'_i$ ， $\{\nu_{n,k}^*\}^2 = \hat{\delta}_{n,k}^2(X_1^*, \dots, X_n^*) + \hat{\delta}_{n,k}^2(Y'_1, \dots, Y'_n)$ 是两样本协方差矩阵的第 K 个对角线元素的和。重新排列 Z^n 的顺序 $(2n)!$ 次，依次带入 (2-21) 计算统计量的值并形成经验分布函数，基于新检验统计量的置换检验经证明是一致有效的。

2.4.4.3 检验的优缺点

由于高维数据的协方差矩阵本身较为复杂，基于边缘标准化统计量的置换检验，只需计算协方差矩阵的对角线元素，相比起改进后的 Hotelling's T^2 检验，形式简单，计算高效。应用于置换检验时，随着置换样本的增多，计算量也不会过分的增加，节省了计算时间和空间，更好的应用于高维数据的检验。

检验的有效性条件为：满足 $\Sigma_X \propto \Sigma_Y$ 或 $m/n \rightarrow 1$ 。也不需要限制总体的其他条件，若检验时，两总体样本数不相等，只需寻找方法将样本数变为相等，保证其一致有效性。Kong et al(2020)中将该方法与现存的其他检验方法进行比较，例如 Chen 和 Qin(2010)(CQ)，Cai. (2014) (CAI)，Xu.(2016)(XLWP)。比较后得出，基于边缘标准化统计量的置换检验无论在控制犯第一类错误的概率上还是检验的经验势能上，都比其他方法表现要好。

本文在 Kong et al(2020)的基础上，利用 Bootstrap 方法基于检验统计量 S_T^* 寻找了一种新的构造两样本数相等的方法。并将新的方法与现存的方法进行比较，分别比较控制犯第一类错误的概率和检验势能。

2.5 本章小结

本章节介绍了基于置换检验对两样本均值的假设检验，分为对一维数据两样本、多维数据两样本、高维数据两样本均值的检验。在对一维数据使用置换检验时，置换检验不需要假定总体的分布，而且和 Wilcoxon 秩和检验、Welch t 检验相比，检验势能更高，也能很好的控制犯第一类错误的概率。

对多维数据使用置换检验时，与 Hotelling's T^2 检验相比较，在正态分布总体下，Hotelling's T^2 检验是精确检验，而且检验流程清晰易懂。但当总体分布不满足正态性假定或当样本数 $m, n > p$ 不成立时，Hotelling's T^2 检验统计量无意义，而置换检验属于非参检验，即不需要对总体进行正态性假定，而且检验统计量可以有多种选择。

对高维数据使用置换检验时，必须满足其有效性条件，而基于不同的检验统计量，置换检验的有效性条件略有差异。本章节分别介绍了基于样本之差统计量、极大值统计量、改进后的 Hotelling's T^2 统计量、边缘标准化统计量的置换检验方法，以及对应的有效性条件，与其他检验方法相比较的优缺点等。

第三章 Bootstrap 方法构造置换检验的一致有效性条件

第三章是本文的重点研究内容，介绍了利用 Bootstrap 方法构造基于边缘标准化统计量置换检验的一致有效性条件。首先介绍了 Bootstrap 方法的定义及原理；其次给出了边缘标准化统计量的定义以及与改进后的 Hotelling's T^2 统计量相比较的优缺点；然后证明了基于边缘标准化统计量置换检验的一致有效性条件为两总体方差相等或两样本数相同；最后通过 Bootstrap 方法产生两个大小一样的伪样本，带入伪样本到检验统计量中形成置换分布进行检验。

3.1 Bootstrap 方法

3.1.1 Bootstrap 方法的定义

近年来在统计学研究领域应用较多的 Bootstrap 方法，最早由 Efron^[36]提出。Bootstrap 是一种非参数的检验方法，属于 Monte Carlo 方法中的一类，通过对样本的观测值再抽样，从而推断总体分布的相关特性。该方法能够从原始样本中充分挖掘出总体有用的信息，而且还不需要对讨论参数的真实分布作假设，也不需要增加样本的数量。可以这么说，Bootstrap 方法降低了对总体分布假定的要求，不需要提前告知总体的信息，对于总体分布中相关参数的估计具有优越性。

而且为了描述这种方法，研究员发明了一种自助语言，对近似分布所做的所有推断都是在 Bootstrap 世界中进行的。有放回地从原始观测样本数据中随机抽样，Bootstrap 样本生成的分布称为 Bootstrap 分布。Brown 等人（2013）^[37]介绍了自助性，并且描述了 Bootstrap 方法可以准确给出所讨论过程渐进特性的能力。

有关 Bootstrap 方法的直观概述，可见图 3-1。由于该方法充分利用了给定的样本观测信息，因此可以应用于多个领域，如生物统计、医学、通讯、金融、岩土工程等领域，文献^[38]提出了可靠性参数和特征量的 Bootstrap 区间估计方法，该方法可以对可靠性统计参数进行分析；将 Bootstrap 方法应用于岩土工程的可靠性分析与风险控制中^[39-40]，对最优 Copula 函数的识别有很大的帮助，以及可以解决有关小样本容量的岩土体参数最优边缘分布函数；在相关武器系统的精度评定方面，文献^[41]提出了在很小的子样条件下，Bootstrap 方法的导弹精度效果如何；文献^[42]提出了有关寿命分布的参数 Bootstrap 拟合优度检验方法，结果证明，在小样本的情况下，相较于传统的经验分布函数检验，该方法优势明显。

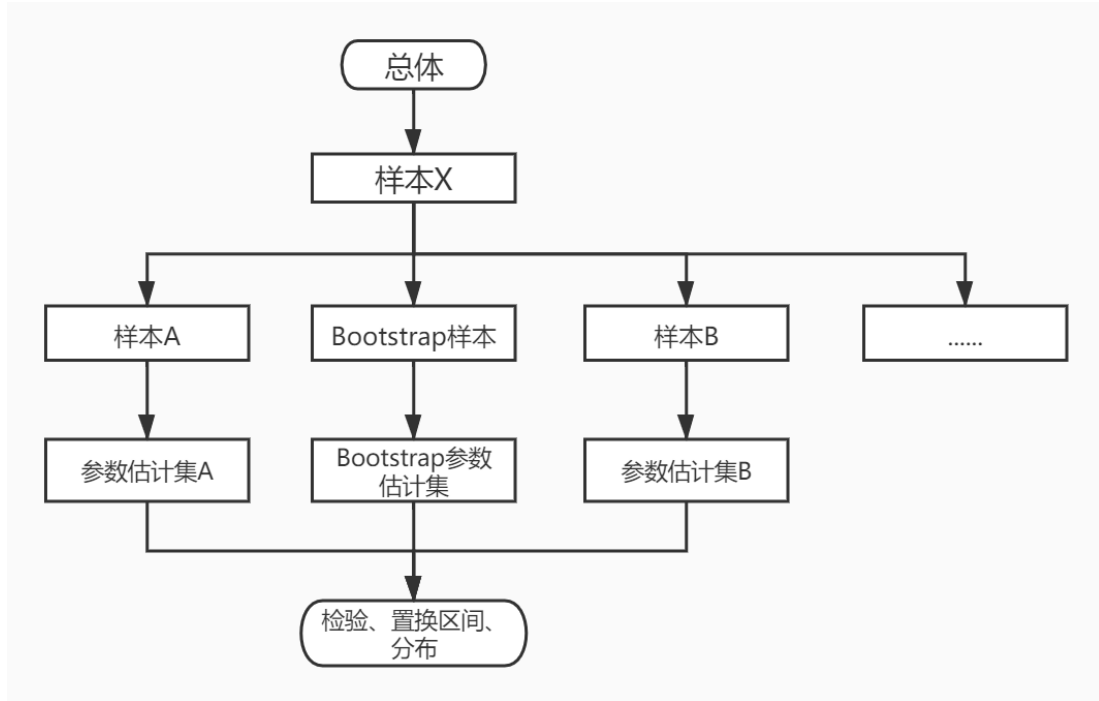


图 3-1 Bootstrap 方法直观概述

上文对 Bootstrap 方法的分析，可以看出该方法具有一定的优势，本文利用 Bootstrap 方法构造检验两样本的样本数相等，具体做法是从原始观测数据中抽取 Bootstrap 样本，使得样本观测值较多的总体和样本观测值较少的总体观测值相同，满足置换检验的有效性条件，从而可以使用边缘标准化统计量的置换检验，减轻计算负担。

3.1.2 Bootstrap 方法的原理

Bootstrap 方法的主要原理是：抽到的样本观测值可以作为对总体较好的近似代替。从总体中抽取不止一个由新观测值构成的数据集，用它们估计总体的相关分布特性。换句话说，Bootstrap 方法的原理是基于这样一个假设，即生成的数据与原始观测数据具有相同的属性。Bootstrap 一般有放回抽取的样本数量不一定非要与原样本量相等，重复抽样 B 次，每次可以求出一个统计量值。

举一个例子来详细讲解 Bootstrap 方法的深层含义：

给定一个随机变量 $H(X, F)$ ，该随机变量是关于 X 和 F 的，其中 $X = [x_1, x_2, \dots, x_n]$ 是独立同分布样本，且服从的分布为 $F(x)$ 。随机变量 $H(X, F)$ 的相关分布特征未知，需要利用观测样本来进行估计。具体的情景案例为：定义随机变量 $H(X, F)$ 为

$$H(X, F) = \hat{\theta}(F_n) - \theta(F) \triangleq R_n. \quad (3-1)$$

这里的 $\theta(F)$ 表示总体分布 F 的某个参数, F_n 为随机样本 $X = [x_1, x_2, \dots, x_n]$ 计算出的经验分布函数, 利用经验分布函数对 $\theta(F)$ 进行估计, 即估计值 $\hat{\theta} = \hat{\theta}(F_n)$, 定义随机变量 $H(X, F)$ 为所估计的相关误差。由于 $H(X, F)$ 的分布特征未知, 利用 Bootstrap 方法计算 $H(X, F)$ 的分布特征的相关步骤如下:

首先利用观测到的样本 $X = [x_1, x_2, \dots, x_n]$ 计算经验分布函数 F_n , 这一步骤是常规易行的, 接着从 F_n 中抽取 Bootstrap 样本 $X^* = [x_1^*, x_2^*, \dots, x_n^*]$, 利用抽取到的样本计算 Bootstrap 统计量 $H^*(X^*, F_n)$ 的值, 定义 F_n^* 是 Bootstrap 样本的经验分布函数, $H^*(X^*, F_n)$ 的表达式为

$$H^*(X^*, F_n) = \hat{\theta}(F_n^*) - \hat{\theta}(F_n) \triangleq H_n. \quad (3-2)$$

这里的 H_n 为 R_n 的 Bootstrap 统计量。重复上述过程 N 次, 可以计算得到 N 个 Bootstrap 统计量 $H^*(X^*, F_n)$ 的可能取值, 对应可求得参数 $\theta(F)$ 的 N 个取值。最终, $H(X, F)$ 分布可以用 $H^*(X^*, F_n)$ 来不断逼近, 即用 H_n 的分布去近似 R_n 的分布, 参数 θ 的特征值和分布亦可获得。

通过分析利用 Bootstrap 方法对总体参数估计的情景案例可知, Bootstrap 方法的有效可靠度由 H_n 和 R_n 的近似程度决定的, 二者越接近, 对总体的参数估计越准确。

3.2 基于置换检验利用 Bootstrap 方法对高维数据两样本均值假设检验

本文使用基于边缘标准化统计量的置换检验对高维数据两样本的均值进行假设检验, 因此也满足该检验统计量的一致性条件。构造两样本数 $m = n$ 时, 采用 Bootstrap 方法从样本观测值较多的一方抽取部分样本, 抽取的 Bootstrap 样本数量和样本观测值较少一方的样本数相同, 满足基于边缘标准化统计量置换检验的一致性条件, 从而可以使用该方法。并将新提出的方法与现存的几种方法进行比较, 比较其控制犯第一类错误的的能力以及检验的经验势能。

3.2.1 检验统计量

本文在 Kong et al(2020)提出的边缘标准化统计量的基础上, 改变了满足置换检验一致有效性的方式。而边缘标准化统计量是由改进后的 Hotelling's T^2 演变的, 之前提到的改进后的 Hotelling's T^2 方法如下:

定义 $\delta_N = (\delta_{N,1}, \dots, \delta_{N,p})^T = N^{1/2}(\bar{X}_m - \bar{Y}_n)$, $\tilde{\Omega}_N$ 是 $\tilde{\Omega} = c^{-1}\Sigma_X + (1-c)^{-1}\Sigma_Y$ 的一致估计, 其中当 $m, n \rightarrow \infty$ 时, $m/N \rightarrow c$ 。令 $h_N = \{\tilde{\Omega}_N\}^{-1/2} \delta_N$ 。定义改进后 Hotelling's T^2 的检验统计量为

$$H_N^2(Z^N) \equiv \|h_N\|_2. \quad (3-3)$$

该文提出的边缘标准化统计量是用协方差矩阵的对角矩阵 D_N 去代替 Hotelling's T^2 中的 $\tilde{\Omega}_N$ ，这里 $\tilde{D}_N = \text{diag}(v_{N,k}^2, k=1, \dots, p)$ ，

$$v_{N,k}^2 = \frac{N}{m} \hat{\sigma}_{m,k}^2(X_1, \dots, X_m) + \frac{N}{n} \hat{s}_{n,k}^2(Y_1, \dots, Y_n). \quad (3-4)$$

而 $\hat{\sigma}_{m,k}^2(X_1, \dots, X_m)$ ， $\hat{s}_{n,k}^2(Y_1, \dots, Y_n)$ 分别是两样本协方差矩阵 $\hat{\sigma}_m^2(X_1, \dots, X_m)$ ， $\hat{s}_n^2(Y_1, \dots, Y_n)$ 的第 k 个对角线元素。对任意的 $\gamma > 0$ ，给出边缘标准化统计量

$$S_N^\gamma(Z^N) \equiv S_N^\gamma(Z_1, \dots, Z_N) = \|\tilde{D}_N^{-1/2} \delta_N\|_\gamma = \left[p^{-1} \sum_{k=1}^p |\delta_{N,k} / v_{N,k}|^\gamma \right]^{1/\gamma}. \quad (3-5)$$

这样做的好处是，由于高维数据的协方差矩阵本身较为复杂，基于边缘标准化统计量的置换检验，只需计算协方差矩阵的对角线元素，相比起改进后的 Hotelling's T^2 检验，形式简单，计算高效。且应用于置换检验时，随着置换样本的增多，计算量也不会过分的增加，节省了计算时间和空间，更好的应用于高维数据的检验。

3.2.2 利用 Bootstrap 方法构造检验的一致有效性

本节根据检验的一致有效性条件，通过 Bootstrap 方法产生两个大小一样的伪样本，使得基于边缘标准化统计量的置换检验对高维两样本均值的假设检验是一致有效的。

定理 3.1 在满足一定的假设条件下，Kong et al(2020)证明了

$$\sup_{t \in R} \left| \frac{1}{N!} \sum_{\pi \in G_N} I\{S_N^\infty(Z_\pi^N) < t\} - \Pr(|\Xi|_\infty < t) \right| \xrightarrow{P} 0, \quad (3-6)$$

这里的 Ξ 是 p 维的高斯分布，其协方差矩阵为 $[\text{diag}(\Sigma(\bar{P}))]^{-1/2} \Sigma(\bar{P}) [\text{diag}(\Sigma(\bar{P}))]^{-1/2}$ ，相关系数矩阵和 $\Sigma(\bar{P})$ 有关；在另一方面，

$$\text{在 } H_0 \text{ 下, } \sup_{t \in R} \left| P(S_N^\infty(Z^N) \leq t) - \Pr(|\tilde{\Xi}|_\infty < t) \right| \xrightarrow{P} 0, \quad (3-7)$$

这里的 $\tilde{\Xi}$ 也是 p 维的高斯分布，其协方差矩阵为 $[\text{diag}(\tilde{\Sigma})]^{-1/2} \tilde{\Sigma} [\text{diag}(\tilde{\Sigma})]^{-1/2}$ ，相关系数矩阵为给定的 $\tilde{\Sigma}$ 。

对于统计量 $S_N^1(\cdot)$ 也有平行的结果，在满足一定的条件下，有

$$\sup_{t \in R} \left| \frac{1}{N!} \sum_{\pi \in G_N} I\{S_N^1(Z_\pi^N) < t\} - \Pr(|\Xi|_\infty < t) \right| \xrightarrow{P} 0, \quad (3-8)$$

这里的 Ξ 是 p 维的高斯分布，其协方差矩阵为 $[\text{diag}(\Sigma(\bar{P}))]^{-1/2} \Sigma(\bar{P}) [\text{diag}(\Sigma(\bar{P}))]^{-1/2}$ ，相关系数矩阵和 $\Sigma(\bar{P})$ 有关；在另一方面，

$$\text{在 } H_0 \text{ 下, } \sup_{t \in R} \left| P(S_N^1(Z^N) \leq t) - \Pr(|\tilde{\Xi}|_\infty < t) \right|^p \rightarrow 0, \quad (3-9)$$

这里的 $\tilde{\Xi}$ 也是 p 维的高斯分布，其协方差矩阵为 $[\text{diag}(\tilde{\Sigma})]^{-1/2} \tilde{\Sigma} [\text{diag}(\tilde{\Sigma})]^{-1/2}$ ，相关系数矩阵为给定的 $\tilde{\Sigma}$ 。

基于定理 3.1，可知在原假设成立的情况下， $\Sigma(\bar{P}) = \tilde{\Sigma}$ ，即满足 $\Sigma(\bar{P}) = c\Sigma_X + (1-c)\Sigma_Y$ 。要使得基于 $S_N^1(\cdot)$ 的置换检验是一致有效的，则需要排除总体方差的影响，即满足 $\Sigma_X = \Sigma_Y$ 或者 $c=1/2$ 。因此基于边缘标准化统计量的置换检验有效性条件为两总体方差相等或两样本数相同。

为了满足检验的一致有效性条件，可利用 Bootstrap 方法构造两样本数相同。基于 Bootstrap 方法的原理，从已有的样本数据中再继续抽样，抽样得到的子样本看作伪样本，具有操作步骤如下：

假定样本 $\{X_i, i=1, \dots, m\}$ 相互独立源自 X 总体，样本 $\{Y_j, j=1, \dots, n\}$ 相互独立源自 Y 总体，并且都是 p 维的随机变量。其中总体 X 的均值向量为 μ^X ，协方差矩阵为 Σ_X ；总体 Y 的均值向量为 μ^Y ，协方差矩阵为 Σ_Y 。设 $m > n$ ，定义 $X'_i = X_i - \mu^X$ ， $Y'_j = Y_j - \mu^X$ ， $i=1, \dots, m$ ， $j=1, \dots, n$ ，这里对样本进行“中心化处理”是有必要的，事实上，可以用 \bar{X}_m 去代替 μ^X 。其中 $\bar{X}_m \equiv (\bar{X}_{m,1}, \dots, \bar{X}_{m,p})^T = \frac{1}{m} \sum_{i=1}^m X_i$ ， $\bar{Y}_n \equiv (\bar{Y}_{n,1}, \dots, \bar{Y}_{n,p})^T = \frac{1}{n} \sum_{j=1}^n Y_j$ 。

将 m 个 $X'_i (i=1, \dots, m)$ 和 n 个 $Y'_j (j=1, \dots, n)$ 组合，令 $Z = \{Z_1, \dots, Z_{m+n}\}$ 。其中 $Z_i = X'_i$ ， $1 \leq i \leq m$ ， $Z_{m+j} = Y'_j$ ， $1 \leq j \leq n$ ，利用 Bootstrap 方法抽取伪样本，由于 X 总体的样本数 m 大于 Y 总体的样本数 n ，所以需要从 X 总体的样本中抽出 Bootstrap 样本。抽取方式为随机抽取 $X'_i (i=1, \dots, m)$ 的前 n 个样本，并将新得到的样本定义为 $X''_i (i=1, \dots, n)$ ，将 $X''_i (i=1, \dots, n)$ 和 $Y'_j (j=1, \dots, n)$ 定义为两个伪样本，两个伪样本的构造特点是样本数相同，数目都为 n ；

从现在起，所有置换检验步骤中的原始样本都用伪样本代替。令 $Z^n = \{Z_1, \dots, Z_{2n}\}$ ，其中 $Z_i = X''_i$ ， $1 \leq i \leq n$ ； $Z_{n+j} = Y'_j$ ， $1 \leq j \leq n$ 。即 Z^n 的前 n 个元素是 $X''_i (i=1, \dots, n)$ ，剩下的元素是 $Y'_j (j=1, \dots, n)$ 。将新构造的样本带入如下的检验统计量中：

$$S_1^{\gamma}(Z^n) = \left\{ p^{-1} \sum_{k=1}^p |\delta_{n,k}^* / \nu_{n,k}^*|^{\gamma} \right\}^{1/\gamma}. \quad (3-10)$$

令 $\bar{X}^* = n^{-1} \sum_{i=1}^n X_i''$, $\bar{Y}^* = n^{-1} \sum_{j=1}^n Y_j'$ 为两个伪样本的样本均值, 其中 $\delta_n^* = (\delta_{n,1}^*, \dots, \delta_{n,p}^*) = n^{\frac{1}{2}} (\bar{X}^* - \bar{Y}^*)$ 。 $v_{n,k}^* = \left\{ \hat{\sigma}_{n,k}^2 (X_1^*, \dots, X_n^*) + \hat{s}_{n,k}^2 (Y_1', \dots, Y_n') \right\}^{\frac{1}{2}}$ 是两样本协方差矩阵第 k 个对角线元素和的平方根, 也是 $\delta_{n,k}^*$ 均方差的估计值。

置换样本集 Z^n 中样本的顺序, G_{2n} 为所有置换可能的集合。对任意 $\pi \in G_{2n}$, 定义 Z_π^n 为 Z^n 的重新排列。将重新排列后的样本带入检验统计量, 即 $S_1^\gamma(Z_\pi^n) \equiv S(Z_{\pi(1)}^n, \dots, Z_{\pi(2n)}^n)$, 检验统计量的值构成经验分布函数, 利用经验分布函数代替原假设成立下的渐进分布, 经验分布函数以及检验统计量的初始值做出假设检验的判断。

3.3 本章小结

本章首先介绍了 Bootstrap 方法的定义及原理, 其原理是基于这样一个假设, 即生成的数据与原始观测数据具有相同的属性, 而且 Bootstrap 方法的一般抽样方式都是“有放回的抽取”。

其次介绍了基于边缘标准化统计量的置换检验, 该检验统计量只需计算协方差矩阵的对角线元素, 相比起改进后的 Hotelling's T^2 检验, 形式简单, 计算高效。且应用于置换检验时, 随着置换样本的增多, 计算量也不会过分的增加, 节省了计算时间和空间, 更好的应用于高维数据的检验。

但使用基于边缘标准化统计量的置换检验对高维数据两样本的均值进行假设检验, 需满足其一致有效性条件, 就需要排除对总体方差的影响, 经验证有效性条件为 $\Sigma_X = \Sigma_Y$ 或者 $c=1/2$ 。

构造两样本数 $m=n$, 利用 Bootstrap 方法抽取伪样本, 由于 X 总体的样本数 m 大于 Y 总体的样本数 n , 所以需要从 X 总体的样本中抽出 Bootstrap 样本。抽取方式为随机抽取 $X_i'(i=1, \dots, m)$ 的前 n 个样本, 并将新得到的样本定义为 $X_i''(i=1, \dots, n)$, 将 $\{X_i''(i=1, \dots, n)\}$ 和 $\{Y_j', j=1, \dots, n\}$ 定义为两个伪样本, 两个伪样本的构造特点是样本数相同, 数目都为 n , 从而满足基于边缘标准化统计量置换检验的一致性条件, 将伪样本带入到检验统计量中, 形成经验分布函数。并将新提出的方法与现存的几种方法进行比较, 比较其控制犯第一类错误的的能力以及检验的经验势能。

第四章 数值模拟分析

本章节主要研究两个模拟实验和一个实际案例。第一个模拟实验，计算本文提出的基于 $S_1^y(\cdot)$ 统计量的相关置换检验对第一类错误的控制，并与其他现存的几种检验方法进行比较；第二个模拟实验，计算 $S_1^y(\cdot)$ 统计量在备择假设成立下的经验势能，将其与其他方法的经验势能进行比较，得出结论。现存的方法包括 Chen and Qin(2010)(CQ), Cai et al.(CAI), Xu et al.(2016)(XLWP) 和 Xue and Yao(2018)(XY)。由于 Chung and Romano(2016)方法过于繁重的计算负担，该方法不在我们的比较实验中。最后，在实际案例中对我国 A 股市场不同行业的股票月收益率进行对比研究，检验各行业两个不同时间段的平均月收益率是否有差异。

4.1 前期准备

4.1.1 数据来源

模拟数据来源如下的模型， $X_i = (x_{i,1}, \dots, x_{i,p}) + \mu^X$, $Y_j = (y_{j,1}, \dots, y_{j,p}) + \mu^Y$; 这里的 μ^X 和 μ^Y 都是 p 维向量，并且对任意给定的 $i=1, \dots, m$ 和 $j=1, \dots, n$ ， $\{x_{i,k}, k=1, 2, \dots\}$ 和 $\{y_{j,k}, k=1, 2, \dots\}$ 服从以下时间序列模型

$$x_{i,k+1} = a_i x_{i,k} + \xi_k, \quad y_{j,k+1} = b_j y_{j,k} + \eta_k, \quad k=1, 2, \dots, \quad (4-1)$$

其中 ξ_k , η_k 为随机误差项， $\{a_i\}_{i=1}^m$, $\{b_j\}_{j=1}^n$ 为超参数，可以是固定的也可以是随机的。针对不同的 a_i , b_j 和 ξ_k , η_k ，我们考虑以下三种模拟模型。

模型 1: $a_i, b_k, i=1, \dots, m, k=1, \dots, n$ 相互独立同服从于 $[0, 0.95]$ 的均匀分布；对任意的 $k \in \{1, \dots, p-1\}$, $\xi_k \sim N(0,1)$, $\eta_k \sim N(0,4)$ 。

模型 2: 和模型 1 相同，但 X_i, Y_j 偶数下标对应的数值乘以 2。

模型 3: $a_i \equiv -0.2, b_j \equiv 0.7$ ，且 $\xi_k \sim t(3), \eta_k \sim 2t(3)$ 。满足这样分布生成的数据是重尾的。

在模拟对第一类错误控制的实验中，设定 $\mu^X = \mu^Y = 0$ ；在模拟经验势能的实验中，设定 $\mu^Y = 0$ 。并且针对 p 维的均值向量 $\mu^X = (\mu_1^X, \dots, \mu_p^X)$ ，有如下两种不同的方案。这两种设定分别基于稠密的情况以及稀疏的情况，要进行分类讨论的原因是，两总体均值的差异有多数不为 0 的情况，也有多数为 0 的情况，前者要按稠密分类，后者要按稀疏分类。针对两种不同的情况，模拟数据的均值向量设定也有差异，如下所示：

(I) 稠密的情况: μ_1^X, \dots, μ_p^X 相互独立同分布于 $[0, c_{n,p}]$ 上的均匀分布, 其中 $c_{n,p} = s / (p^{0.25} \times \min(m, n)^{0.5})$ 。针对于上述提及的模型 1-3, 分别选择 $s=6, 9, 11$ 来控制整体的信噪比。

(II) 稀疏的方案: 从 μ_1^X, \dots, μ_p^X 中随机选择 $0.2 \times p^{0.5}$ 个元素, 这些元素值仍然相互独立同分布且服从 $[0, c_{n,p}]$ 上的均匀分布, 并赋值 $c_{n,p} = s / \min(m, n)^{0.5}$ 。剩下未选中的元素将其赋值为 0。对于不同的模型 1-3, 选择不同的 $s=7, 8, 9$ 来控制整体的信噪比。

4.1.2 样本数量及维数选择

为了更加直观的对比检验结果的好坏, 对样本容量的选择和样本维数的确定, 我们考察几种不同的设定, 其中包括样本维数小于样本容量或样本维数大于样本容量这两种情况。样本观测值的多少从数量较少的 ($m=75, n=50$), 到数量中间的 ($m=300, n=200$), 再到数量较大的 ($m=600, n=400$) 这三种设定, 而维数的选择有 $p=10, 100, 1000$ 。

对应上文的三种模型, 分别给定不同的样本维数和样本数量, 生成对应的模拟数据, 进行后续的实验操作。

4.1.3 确定置换次数

在本文的模拟实验中, 经验分布函数是通过将伪样本置换 5000 次, 把每一次置换后的样本带入到检验统计量中形成的, 并且用经验分布函数代替原假设成立下的渐进分布函数。而在计算经验概率时, 需要计算 10000 个检验统计量的值并与初始值进行比较, 言外之意就是需要置换 10000 次样本。在计算经验势能时, 即在原假设不成立的前提下, 需要置换 3000 次样本。

4.2 模拟犯第一类错误的经验概率

经验概率是判断检验方法效果好坏的重要衡量指标, 其表达公式简单定义如下:

$$P = \frac{\{S_1^r(\cdot) \geq S_0\} \text{ 的个数}}{10000}.$$

其中 S_0 表示检验统计量的初始值, $S_1^r(\cdot)$ 表示将置换样本带入统计量中得到的值, 二者进行比较, 计算 $S_1^r(\cdot) \geq S_0$ 的个数, 分母的 10000 表示总的置换次数, 分子分母相比即得到犯第一类错误的经验概率 P 。

分别比较我们的检验方法与 CQ(2010)、XLWP(2016)的检验方法, 其他现存且

有效的方法暂时先不进行比较。在原假设成立的条件下,基于不同的三种假设模型,计算出新检验统计量 $S_1^y(\cdot)$ 以及 CQ(2010)、XLWP(2016)这三种方法犯第一类错误的经验概率。并将经验概率与显著性水平 $\alpha = 0.05$ 作比较,结果与 0.05 越接近,说明效果越好,经验概率偏大或偏小说明检验方法不能有效的控制犯第一类错误的概率。表中加粗的数据表示在 0.05 附近,未加粗的数据表示偏离 0.05 多于 20%以上。

表 4-1 基于模型 1 下的犯第一类错误的经验概率

Model	n	p	CQ	XLWP		$S_1^y(\cdot)$	
			L_2	L_2	L_∞	L_2	L_∞
1	50	10	4.51	7.13	3.16	4.76	5.53
		100	4.51	5.24	4.18	4.89	5.67
		1000	4.97	6.41	4.46	5.39	5.89
	200	10	4.76	6.32	2.97	4.12	4.18
		100	5.02	5.87	3.27	5.68	5.37
		1000	5.12	4.35	3.34	5.34	5.88
		10	4.17	6.98	3.28	4.78	4.79
	400	100	4.76	6	3.9	4.66	6.2
		1000	4.62	4.7	4.4	4.92	6.81

从表 4-1 基于模型 1 下的犯第一类错误的经验概率的结果表中,可总结出结论:

(1) 基于不一样的数据维数 p 和样本容量 n 的设置组合,计算出大部分犯第一类错误的经验概率都在 $\alpha = 0.05$ 的附近波动,在这三种检验方法比较中,我们提出的检验统计量可以更好的控制犯第一类错误的概率;

(2) 我们提出的方法和 CQ(2010)、XLWP(2016)这两个方法相比较,检验概率值随着样本维数 p 和样本容量 n 的增大离 0.05 更近。纵向来看,举一个例子,当样本量 $n=400$ 时,犯第一类错误的经验概率随着维数 p 的增大逐渐趋于 0.05,特别是在 L_2 的情形下,新的检验统计量没有产生不正常的结果值,检验过程更加稳定;XLWP 方法纵向来看,随着维数与样本量比值的增大,结果也越来越不好,逐渐偏离 0.05;CQ 方法在模型 1 中的表现也比较优异,而且随着维数的增大,也没有出现异常值,一直稳定在 0.05 附近。从横向来看,新检验统计量的模拟结果比其他两种方法表现更优。

表 4-2 的结果表明,基于模型 2 的假设下,第一类经验错误率的表现要优于基于模型 1 的假设,全部的值都落在 0.05 附近,而且犯第一类错误的经验概率随着维数 p 的增大逐渐趋于 0.05,不管在 L_2 的情形或者是在 L_∞ 的情形下,表现都是稳

定的；CQ 方法在模型 2 中的表现也是可圈可点，都能比较好的控制犯第一类错误的概率；反观 XLWP 方法，在均值稀疏差异的情形下表现要略好于在均值稠密差异情形，但总体表现还是偏于 0.05 超于 20% 的比率比较大，这也和该方法的经验势能表现有关，犯第一类错误的经验概率反而随着数据维数与样本量比值的增大慢慢趋于 0.05。

表 4-2 基于模型 2 下的犯第一类错误的经验概率

Model	n	p	CQ	XLWP		S_1^y	
			L_2	L_2	L_∞	L_2	L_∞
2	50	10	4.82	7.19	3.87	5.88	5.45
		100	5.28	6.50	5.10	5.10	5.04
		1000	5.15	8.15	7.85	4.15	5.77
	200	10	4.70	7.34	3.17	4.56	4.82
		100	4.41	6.39	4.36	5.73	4.70
		1000	5.26	4.94	4.98	6.68	4.71
	400	10	4.51	7.05	3.23	5.47	5.49
		100	4.02	6.09	4.16	5.88	5.76
		1000	5.05	4.71	4.33	4.96	4.36

表 4-3 基于模型 3 下的犯第一类错误的经验概率

Model	n	p	CQ	XLWP		S_1^y	
			L_2	L_2	L_∞	L_2	L_∞
3	50	10	4.85	8.56	3.69	4.06	4.79
		100	4.73	8.32	4.97	5.78	5.63
		1000	5.57	12.11	6.32	6.92	6.60
	200	10	4.56	6.67	3.33	4.24	5.32
		100	4.87	6.35	3.97	6.89	5.94
		1000	5.06	4.92	5.04	7.63	7.98
	400	10	4.57	7.19	3.28	5.43	4.56
		100	4.45	6.23	3.90	5.70	6.32
		1000	4.86	5.14	4.40	6.37	5.61

从表 4-3 基于模型 3 下的犯第一类错误的经验概率的结果表中，我们可以的到以下结论：

(1) 基于模型 3，存在较多犯第一类错误的经验概率值偏离标准的显著性水平，而且随着维数 p 的增大，即 p/n 的比值增大，检验结果出现不稳定的现象，但我们的检验方法还是要优于 XLWP 方法，比 XLWP 方法能更好的控制犯第一类错误的概率，略差于 CQ 检验方法；

(2)这三个模型的结果进行比较,在模型 1 和模型 2 假设下实验结果相类似:首先,随着维度 p 和样本量 n 的比值的增大,犯第一类错误的经验概率都较为稳定,并且都在显著性水平 $\alpha = 0.05$ 附近波动。其次,总的来说,我们提出的基于边缘标准化统计量的置换检验比 CQ(2010)、XLWP(2016)的检验方法要更加优越。基于模型 3 的模拟结果表现有些差强人意,这也和模型 3 的数据来源有关,模型 3 生成的数据是重尾的数据。

4.3 模拟经验检验势能

在备择假设 H_1 下,即 $\mu_x \neq \mu_y$,我们分别模拟新检验统计量的经验势能以及 CQ(2010)和 XLWP(2016)的经验势能。分别比较高维数据均值稠密差异下 (L_2) 新检验统计量和 CQ(2010)方法、XLWP(2016)方法的表现优劣,以及在均值稀疏差异下 (L_∞) 新检验统计量和 XLWP(2016)方法的表现优劣。

经验势能作为检验方法好坏的评判标准其一,也发挥着举足轻重的作用,而且和犯第一类错误的经验概率之间存在着相互制约的关系。经验检验功效的公式为:

$$T = \frac{\{S_1^r(\cdot) > \text{置换分布 } (1-\alpha) \text{ 分位数}\} \text{ 的个数}}{3000},$$

这里的 $S_1^r(\cdot)$ 表示通过模拟样本数据计算得到的检验统计量的值,分位数表示在原假设下进行 5000 次模拟得到的经验分布的 $100(1-\alpha)\%$ 分位数,两者相比并计算比较后的个数,分母的 3000 表示总的置换次数,分子分母相比即得到经验检验势能。

表 4-4 模拟了基于三个模型假设下 CQ、XLWP 以及 $S_1^r(\cdot)$ 这三种方法针对不同维度和样本量在显著性水平 $\alpha = 0.05$ 下的经验检验功效。从表 4-4 的结果可以得到以下结论:

(1)在大多数 (n, p) 的假设组合下,对于高维数据均值稠密差异的检验统计量 $S_1^2(\cdot)$ 都要比 CQ 或 XLWP 方法表现出更高的检验势能,并且随着维数 p 的增大, $S_1^2(\cdot)$ 方法的增长速度要快一些。

(2)在模型 1 的假设下, CQ 方法和我们的方法 $S_1^2(\cdot)$ 计算的表现很相似,而且都很接近 1,表现较为优异;但 XLWP 方法在模型 1 中具有非常低的势能,大多数都在 0.3 附近徘徊。

(3)在模型 2 的假设下, CQ 和 XLWP 的势能都低于 $S_1^2(\cdot)$ 方法。虽然在模型 2 中, CQ 方法可以更好的控制犯第一类错误的概率,但在检验势能方面,表现不佳;与模型 1 相比, XLWP 方法有所进步,在势能方面的表现优于 CQ 方面,这很

像是对经验概率表现不好的补偿。

(4)在模型 3 的假设下, $S_1^2(\cdot)$ 方法计算的经验检验势能在维数较低时没有 CQ 和 XLWP 方法表现好, 但随着维数 p 的增大, 在高维数据的情形下, 表现越来越好, 逐渐优于其他两种方法; XLWP 方法在模型 3 中的表现也优于在模型 1 和模型 2 中的表现; 但从模拟的这三个模型来看, 这三种方法在模型 3 中的经验势能都不是很高。

表 4-4 基于三种模型下稠密假设下的经验检验功效

Model	n	p	CQ	XLWP	$S_1^2(\cdot)$
1	50	10	0.76	0.31	0.82
		100	0.85	0.32	0.86
		1000	0.92	0.32	0.96
	200	10	0.78	0.32	0.84
		100	0.86	0.33	0.91
		1000	0.90	0.32	0.95
	400	10	0.72	0.35	0.87
		100	0.84	0.40	0.92
		1000	0.91	0.39	0.94
2	50	10	0.40	0.46	0.73
		100	0.41	0.47	0.78
		1000	0.43	0.51	0.86
	200	10	0.39	0.47	0.56
		100	0.42	0.48	0.73
		1000	0.40	0.52	0.88
	400	10	0.35	0.48	0.64
		100	0.41	0.51	0.78
		1000	0.43	0.53	0.90
3	50	10	0.58	0.59	0.46
		100	0.63	0.65	0.54
		1000	0.66	0.67	0.76
	200	10	0.58	0.60	0.43
		100	0.62	0.66	0.65
		1000	0.67	0.71	0.76
	400	10	0.54	0.61	0.37
		100	0.62	0.68	0.63
		1000	0.64	0.73	0.82

均值稠密差异讨论过后, 再看看均值稀疏差异下, 我们的检验方法与其他检验方法的比较结果。在高维数据两样本均值稀疏差异情形中, 比较 XLWP 方法和我们的方法 $S_1^\infty(\cdot)$ 计算得到的经验检验势能, 这里的 $S_1^\infty(\cdot)$ 的计算公式为:

$$S_1^\infty(Z^n) = \left\{ p^{-1} \sum_{k=1}^p |\delta_{n,k}^* / \nu_{n,k}^*|^\infty \right\}^{1/\infty}. \quad (4-2)$$

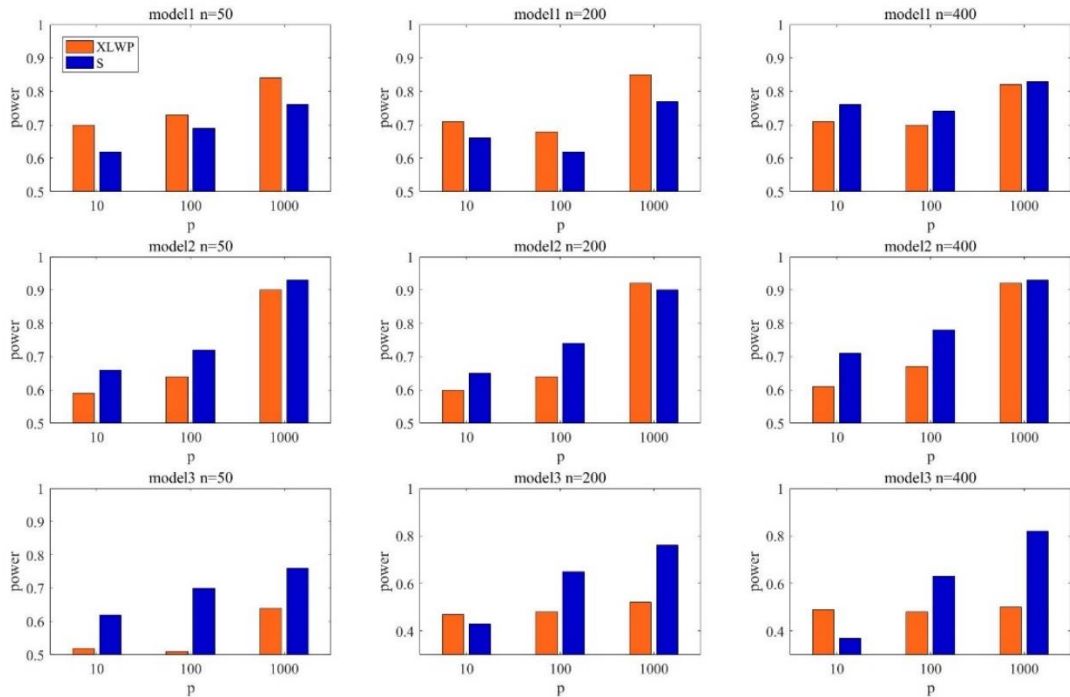


图 4-1 基于三种模型下稀疏假设下的经验检验功效

根据 XLWP 方法和本文提出的 $S_1^\infty(\cdot)$ 方法计算基于三种模型以及在不同的 (n, p) 组合下得到的经验检验势能，并绘制如上九宫格图表 4-1，可以更直观的比较这两种方法的差异，其中橙色柱子代表的是根据 XLWP 方法计算得到的经验检验势能，蓝色柱子代表的是根据 $S_1^\infty(\cdot)$ 方法计算得到的经验检验势能。观察图 4-1 可知：

(1) 从横向来看，即比较同一个模型下的检验势能差异，可以看出在模型 1 假设下，XLWP 方法表现略优于 $S_1^\infty(\cdot)$ 方法，随着样本量的不断增多， $S_1^\infty(\cdot)$ 方法的检验势能逐渐回升，而且随着维数的变大，两种方法的势能都在稳步上升；在模型 2 的假设下， $S_1^\infty(\cdot)$ 方法的经验检验势能要更高，而且随着维数的不断增大，这两种方法的经验势能渐渐地向着 1 靠近，尤其是在 $p=1000$ 时，经验检验势能与 1 非常接近；在模型 3 的假设下， $S_1^\infty(\cdot)$ 方法的优势就更加明显了，计算得到的经验检验势能远远领先于 XLWP 方法，而 XLWP 方法在模型 3 中的表现有些差强人意。但在样本数量比较少时，检验势能都比较低，随着样本数的升高，势能逐渐增加。

(2) 从纵向来看， $S_1^\infty(\cdot)$ 方法在模型 2 和模型 3 中的表现较好，在模型 1 中的计算得到的经验检验势能比较低；XLWP 方法在模型 1 和模型 2 中经验检验势能较大，但在模型 3 中表现不太好；

(3) 总体来说，在局部备择假设成立下，本文提出的方法 $S_1^\infty(\cdot)$ 基于三种模型

假设下，具有较好的经验势能。并且随着维数的增大，对应的经验检验功效都显著的增大。与 XLWP 方法相比，具有更稳定的检验势能，而且计算得到的经验势能趋于 1 的速度更快。

4.4 实例分析

本节进行实例分析，研究对象是中国 A 股市场不同行业的股票收益，想判断不同行业的股票月收益率是否具有明显的“五月卖出”效应^[43]，即在两个不同的时间段，比较每个行业平均月收益率是否有差异（两个时间段为当年 5 月至 10 月与上一年 11 月至 4 月）。我们从国泰安 CSMAR 数据库中找到股票收益数据，时间范围从 2001 年 5 月至 2017 年 10 月。不同的股票对应于不同的行业，对于高维数据的行业，采用 CQ(2010)方法以及我们提出的基于边缘标准化统计量的置换检验方法进行检验，比较这两个方法的优劣。

对数据进行预处理，剔除那些缺失月收盘价格数据的公司，收集到的数据共有 387526 条。记 p_l 为行业 l 中所包含的股票支数， $P_{i,t}^l$ 为行业 l 的第 i 支股票第 t 个月的月收盘价，将月收盘价格数据转换为月收益率数据：

$$r_t^l = \left(\log \frac{P_{1,t+1}^l}{P_{1,t}^l}, \log \frac{P_{2,t+1}^l}{P_{2,t}^l}, \dots, \log \frac{P_{p_l,t+1}^l}{P_{p_l,t}^l} \right), t = 1, \dots, 198. \quad (4-3)$$

这 387526 条数据中共有 3531 支 A 股，从中选出 227 支股票的月收盘价格数据，这 227 支股票共来自 12 个行业，这 12 个行业中数据维数 p 大于样本量 n 的有 6 个行业，分别是^[43]房地产业（FDC），电力、煤气及水的生产和供应业（SCGY），制造业（ZZY），仓储业（JTYS），信息技术（XXJS），批发和零售贸易（PFLS）。在进行检验之前，我们发现基于边缘标准化统计量的置换检验，要求样本之间相互独立且同分布，因此需要先计算股票收益率之间的 Spearman 相关系数，对我国 A 股两个不同时间段总体的样本股相关性进行分析。发现 5 月到 10 月总体中两支股票的相关系数绝对值大于等于 0.6 即属于强相关的有 1017 组。上一年 11 月到 4 月总体中相关系数绝对值大于等于 0.6 的有 1359 组，说明大部分行业样本股的收益率都具有相关性，因此对每一个行业的月收盘价格建立时间序列拟合模型，将拟合后的残差作为样本观测值，以此来去除样本间的相关性，再进行后续的置换步骤。要比较的 CQ 方法是不需要去除相关性的。

根据我们的目的提出高维数据两样本股票月收益率均值的假设检验问题，原假设为：

$$H_0^l : \mu_X^l = \mu_Y^l \quad (4-4)$$

其中 $\mu_X^l = E(X_i^l)$ ， $\mu_Y^l = E(Y_j^l)$ ， $l = 1, \dots, 6$ ，共 6 个行业。这里的 $X_i^l = \bar{r}_{X,i}^l$ ， $i = 1, 2, \dots, m^l$ ，

$Y_j^l = \bar{r}_{Y,j}^l, j=1,2,\dots,n^l$, 行业 l 在总体 X 中的样本量用 m^l 表示, 在总体 Y 中的样本量用 n^l 表示, 每年的第一个时间段 l 行业的平均对数月收益率向量用 $\bar{r}_{X,i}^l$ 表示, $\bar{r}_{Y,j}^l$ 表示第二个时间段 l 行业的平均对数月收益率向量。经过整理我们发现, 见表 4-5: 要分析的 227 支股票中, 房地产业 (FDC) 中的股票有 15 支, 在总体 X 的样本量有 17, 在总体 Y 的样本量有 16, 电力、煤气及水的生产和供应业 (SCGY) 中股票有 17 支, 制造业 (ZZY) 中的股票种类较多有 115 支, 仓储业 (JTYS) 股票有 17 支可以达到平均水平, 信息技术 (XXJS) 和批发和零售贸易 (PFLS) 分别有 18 和 23 支, 它们在总体 X 和总体 Y 的样本量分别为 17 和 16。

表 4-5 高维数据各行业两总体的样本维数和样本量

行业	p	m	n
FDC	15	17	16
SCGY	17	17	16
ZZY	115	17	16
JTYS	17	17	16
XXJS	18	17	16
PFLS	23	17	16

下面我们在显著性水平 $\alpha=0.05$ 下, 进行假设检验, 判断数据维数 p 大于样本量 n 的 6 个行业的月收益率在两个不同的时间段是否有明显的差异, 并且比较 CQ(2010)方法以及我们提出的基于边缘标准化统计量的置换检验方法的检验势能。

表 4-6 高维数据各行业两总体的检验概率

行业	p	m	n	P_s	P_{CQ}
FDC	15	17	16	0.1076	0.1236
SCGY	17	17	16	0.0019	0.0015
ZZY	115	17	16	0.0056	0.0025
JTYS	17	17	16	0.0037	0.0033
XXJS	18	17	16	0.0000	0.0002
PFLS	23	17	16	0.0042	0.0059

从表 4-6 可以看出, 这两种方法的判断结果是相同的, 在显著性水平 $\alpha=0.05$ 下, 电力、煤气及水的生产和供应业 (SCGY), 制造业 (ZZY), 仓储业 (JTYS), 信息技术 (XXJS), 批发和零售贸易 (PFLS) 这 5 个行业的检验 p 值都显著小于 0.05, 应当拒绝原假设, 这些行业 5 月到 10 月的平均对数收益率与上一年 11 月到 4 月的平均对数收益率不相等, 存在显著的“五月卖出”效应。而房地产业 (FDC) 两种方法对应得检验概率都大于 $\alpha=0.05$, 即不能拒绝原假设。

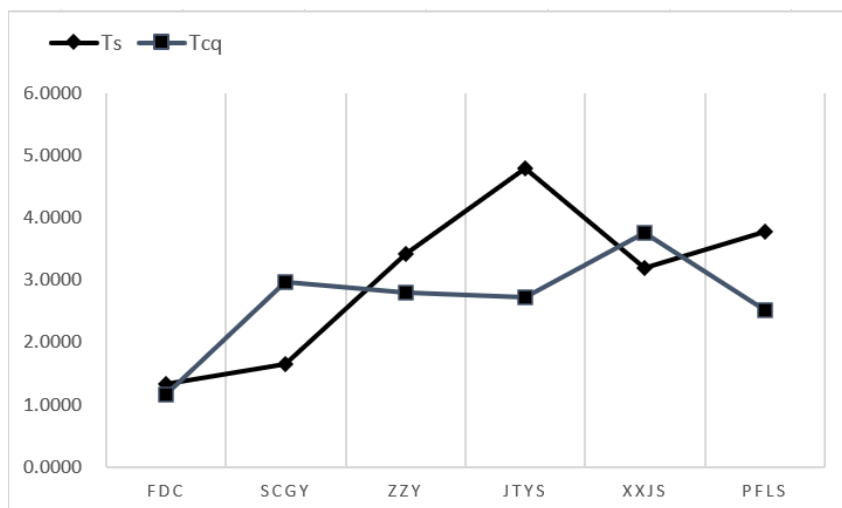


图 4-2 两种方法的经验势能比较

接下来比较这两种方法的检验势能，谁更高一些。观察图 4-2 可知，我们提出的新检验统计量的势能 T_s 总体来看是高于 CQ(2010)方法的检验势能，相较于 CQ(2010)方法假设检验效果更好。

4.5 本章小结

在模拟实验中，我们提出的检验统计量和其他两种方法相比可以更好的控制犯第一类错误的概率以及具有更高的检验势能。在实例分析中，对高维的 6 个行业两个不同时间段的平均月收益率进行比较，发现在显著性水平 $\alpha = 0.05$ 下，电力、煤气及水的生产和供应业 (SCGY)，制造业 (ZZY)，仓储业 (JTYS)，信息技术 (XXJS)，批发和零售贸易 (PFLS) 这 5 个行业的检验 p 值都显著小于 0.05，存在明显的“五月卖出”效应。而房地产业 (FDC) 两种方法对应得检验概率都大于 $\alpha = 0.05$ ，可以认为不存在“五月卖出”效应。

第五章 全文总结与展望

5.1 全文总结

本文以高维数据两样本均值的假设检验为研究背景，考虑基于边缘标准化统计量的置换检验方法，该方法不需假定总体分布，且计算协方差矩阵时只需计算对角线元素，相比起改进后的 Hotelling's T^2 检验，形式简单，计算高效。基于边缘标准化统计量的置换检验一致有效性条件为 $\Sigma_X = \Sigma_Y$ 或者 $c=1/2$ ，即为两总体方差相等或两样本数相同。本文研究的重点是利用 Bootstrap 方法构造两个样本数量相等的伪样本，满足 $m=n$ ，利用了 Bootstrap 方法抽样生成的数据与原始观测数据具有相同的属性的原理。最后进行模拟实验，比较新检验方法与 CQ(2010)、XLWP(2016)这两种方法的犯第一类错误的经验概率以及经验检验势能，验证了新检验方法可以更好的控制犯第一类错误的概率以及具有更高的检验势能。以下分点总结本文得到的主要结论：

(1) 本文提出的边缘标准化统计量是由改进后的 Hotelling's T^2 演变的，是用协方差矩阵的对角矩阵 \tilde{D}_N 去代替 Hotelling's T^2 中的 $\tilde{\Omega}_N$ ，新检验统计量为 $S_N^z(Z^N) \equiv S_N^z(Z_1, \dots, Z_N) = \|\tilde{D}_N^{-1/2} \delta_N\|_y = \left[p^{-1} \sum_{k=1}^p |\delta_{N,k} / v_{N,k}|^y \right]^{1/y}$ 。这样做的好处是，由于高维数据的协方差矩阵本身较为复杂，基于边缘标准化统计量的置换检验，只需计算协方差矩阵的对角线元素，相比起改进后的 Hotelling's T^2 检验，计算更为高效。且应用于置换检验时，计算量也不会过分的增加，更好的应用于高维数据的检验。

(2) 进一步的，要使用置换检验方法，必须保证检验方法的一致有效性。基于相关定理，可知在原假设成立的情况下， $\Sigma(\bar{P}) = \tilde{\Sigma}$ ，即满足 $\Sigma(\bar{P}) = c\Sigma_X + (1-c)\Sigma_Y$ 。要使得基于 $S_N^z(\cdot)$ 的置换检验是一致有效的，则需要排除总体方差的影响，即满足 $\Sigma_X = \Sigma_Y$ 或者 $c=1/2$ 。利用 Bootstrap 方法抽取伪样本构造两样本数 $m=n$ ，由于 X 总体的样本数 m 大于 Y 总体的样本数 n ，所以抽取方式为从 X 总体中随机抽取 $X'_i (i=1, \dots, m)$ 的前 n 个样本，并将新得到的样本定义为 $X''_i (i=1, \dots, n)$ ，将 $X''_i (i=1, \dots, n)$ 和 $\{Y'_j, j=1, \dots, n\}$ 定义为两个伪样本，两个伪样本的构造特点是样本数相同，数目都为 n ，将伪样本带入到检验统计量中，形成经验分布函数，进行后续的假设检验。

(3) 最后将本文提出的方法与其他现存的对高维数据均值检验的方法进行比较。发现在不同的数据维数 p 以及样本容量 n 的设置下，基于不同的假设模型大多数犯第一类错误的经验概率都在标准值附近上下波动，较为稳定。随着样本维数 p

和样本容量 n 的增大, 我们提出的新检验方法和 CQ(2010)、XLWP(2016)的检验方法相比, 得到的经验概率结果更好且更接近 0.05, 总的来说, 我们提出的检验统计量可以更好的控制犯第一类错误的概率; 而且在大多数 (n, p) 的假设组合下, 对于高维数据均值稠密差异的检验统计量 $S_1^2(\cdot)$ 都要比 CQ 或 XLWP 方法表现出更高的检验势能, 并且随着维数 p 的增大, 即 p/n 的比值增大, 三种方法的检验势能都呈现增长的趋势, $S_1^2(\cdot)$ 方法的增长速度要快一些。分析我国 A 股 6 个高维数据行业在两个不同时间段的平均月收益率是否有差异, 发现有 5 个行业具有明显的“五月卖出”效应。

5.2 展望与不足

本文虽有一定的研究意义, 但不足之处也不容忽略, 即之后可以继续深入研究的方向。

首先, 基于边缘标准化统计量的置换检验有效性条件为两样本数相等或两总体方差相等。本文利用 Bootstrap 方法, 从大样本中抽取和小样本数量相同的 Bootstrap 样本, 从而达到一致性条件。但其他构造样本相等的方式并没有去验证, 例如可以通过从小样本中以再抽样的方式来扩大小样本的数量, 使其与大样本数相同。这种从“小”变“大”的方法也可以去尝试, 比较基于哪种构造方式下的置换检验效果是最好的。

其次, 也可以从总体方差这一角度切入, 即通过构造两总体方差相同, 来满足置换检验的一致性条件。但由于高维数据的方差计算特别复杂, 而且这方面的研究比较少, 所以如何构造两总体的方差相同, 还需要继续深思。

最后, 观察模拟结果发现在高维数据两样本均值稀疏差异的情形下, $S_1^\infty(\cdot)$ 方法在模型 1 假设下, 经验检验势能略低于 XLWP 方法, 即模拟的经验检验势能不够理想, 还需要改进。现存的高维均值向量的检验方法已经逐渐趋于成熟, 我们仍要不断学习和改进, 提出新的更好的检验方法。

致 谢

时光转瞬即逝，回望这三年研究生的求学经历，发现自己不光独立思考和解决问题的能力得到了提升，专业知识和相关技能也颇有长进，这要感谢帮助我的各位老师、同学、家人和朋友们，正因为有她们，才成就了现在的我！

首先，感谢我的导师孔娴芳教授。在三年的硕士攻读期间，孔老师在学习和生活中都提供了很多帮助，让我学会如何面对研究生学习和生活中遇到的难题。在学习方面，孔老师对待科研十分认真严谨，注重细节，从刚入学时便认真叮嘱我要专心科研。在生活中也是一位热爱生活、爱运动且博学多才的学者，经常开导和鼓励我不断前进。在本次论文的书写完成过程中孔老师都悉心指导，耐心督促，认真修改。而且在未来工作的选择上也会给我相关建议，我的研究生学习阶段离不开孔老师对我的教导。

同时也感谢团队中的夏应存教授和秦旭副教授，感谢他们对我学习和科研方面的指导，以及对我生活方面的关心与关怀。感谢学校的其他老师们和各位领导，给我提供了一个好的学习和生活环境，让我可以更便利的享受到优秀的教师和学习资源。

感谢我的父母，是他们给予我一个友好和谐的家庭环境，在你们的关爱和鼓励中，度过了人生的求学时光，让我没有任何顾虑脚踏实地的追逐梦想，勇敢的继续向前。

感谢我的同学、室友、男友和朋友们，感谢你们的陪伴、关心以及困难时刻给予的帮助，很开心可以遇到你们，让我的研究生时光变得丰富多彩，不再孤单，希望我们的未来都如我们所愿，继续努力，在顶峰相见。

最后，再次感谢所有我爱的人和爱我的人，这一路上，有欢乐，有收获，有成长，有付出，是你们的一路陪伴，让我觉得前进的路途不再孤单，不再迷茫，让我可以成为更好的自己。接下来，我将踏上新的征程，我将怀感恩之心，不忘初心，继续向前！

参考文献

- [1] Zhu L P , Zhu L X . Nonconcave penalized inverse regression in single-index models with high dimensional predictors[J]. Journal of Multivariate Analysis, 2009, 100(5):862-875
- [2] Bai Z , Saranadasa H . Effect of high dimension: By an example of a two sample problem[J]. Statistica Sinica, 1996, 6(2):201-362
- [3] Chen S X , Qin Y L . A two-sample test for high-dimensional data with applications to gene-set testing[J]. Annals of Statistics, 2010, 38(2):808-835
- [4] Cai T T , Liu W , Yin X . Two-sample test of high dimensional means under dependence[J]. Journal Of The Royal Statistical Society, 2013, 76(2):20-86
- [5] Xu G , Lin L , Wei P , et al. An adaptive two-sample test for high-dimensional means[J]. Biometrika, 2016, 103(3):609-624
- [6] Chang J , Chao Z , Zhou W , et al. Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity[J]. Biometrics, 2017, 73(4)
- [7] Hall P , Neeman M A . Geometric Representation of High Dimension, Low Sample Size Data[J]. Journal of the Royal Statistical Society, 2005, 67(3):427-444
- [8] Donoho D , Tanner J . Neighborliness of randomly projected simplices in high dimensions[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(27):9452-9457
- [9] 茆诗松, 濮晓龙, 程依明. 概率论与数理统计简明教程[M]. 高等教育出版社, 2012
- [10] 高惠璇. 应用多元统计分析[M].北京:北京大学出版社, 2015
- [11] Engelhardt H T . Diagnosing Well and Treating Prudently: Randomized Clinical Trials and the Problem of Knowing Truly[M]. Springer Netherlands, 1988
- [12] Pitman E. Significance tests which may be applied to samples from any populations[J]. Supplement to the Journal of the Royal Statistical Society, 1937, 4(1):119-130
- [13] Pitman E . SIGNIFICANCE TESTS WHICH MAY BE APPLIED TO SAMPLES FROM ANY POPULATIONS III. THE ANALYSIS OF VARIANCE TEST[J]. Biometrika, 1938,3(4):13-38
- [14] F Wilcoxon. Individual Comparisons by Ranking Methods[J]. Biometrics Bulletin, 1945, 1(6):80-83
- [15] Efron B , Tibshirani R . An Introduction to the Bootstrap / B[M]. Efron, R.J. Tibshirani. EP, 2005
- [16] 冯变英. 关于均值向量的置换检验的研究[D]. 华东师范大学, 2009
- [17] Stanberry L . Permutation Test[J]. Springer New York, 2013

-
- [18] Chung E Y , Romano J P . Multivariate and multiple permutation tests[J]. *Journal of Econometrics*, 2016:76-91
- [19] Hotelling H . The Generalization of Student's Ratio[J]. *The Annals of Mathematical Statistics*, 1992, 2(3):36-89
- [20] Hotelling H . Multivariate Quality Control[J]. *Techniques of Statistical Analysis*, 1947, volume 31(3):17-20
- [21] Srivastava M S , Du M . A test for the mean vector with fewer observations than the dimension[J]. *Journal of Multivariate Analysis*, 2008, 99(3):386-402
- [22] Levina E , Bickel P J . Maximum Likelihood Estimation of Intrinsic Dimension[C]// *Advances in Neural Information Processing Systems 17* [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]. 2004
- [23] 赵俊光. 高维数据下总体均值的假设检验问题[D]. 北京理工大学, 2016
- [24] Hjort N L , McKeague I W , Keilegom I V . Extending the scope of empirical likelihood[J]. *The Annals of Statistics*, 2009, 37(3):1079-1111
- [25] Wang R , Peng L , Qi Y . JACKKNIFE EMPIRICAL LIKELIHOOD TEST FOR EQUALITY OF TWO HIGH DIMENSIONAL MEANS[J]. *Statistica Sinica*, 2013
- [26] Li G , Lin L , Zhu L . Empirical likelihood for a varying coefficient partially linear model with diverging number of parameters[J]. *Journal of Multivariate Analysis*, 2012, 105(1):85-111.
- [27] Liang P , Qi Y , Fang W . Test for a Mean Vector with Fixed or Divergent Dimension[J]. *Statistical Science*, 2014, 29(1):113-127
- [28] Chen L S , Paul D , Prentice R L , et al. A Regularized Hotelling's T₂ Test for Pathway Analysis in Proteomic Studies[J]. *Publications of the American Statistical Association*, 2011, 106(496):1345-1360
- [29] Shen Y , Lin Z , Zhu J . Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis[J]. *Computational Statistics & Data Analysis*, 2011, 55(7):2221-2233
- [30] Ahmad M R , Rosen D V , Singull M . A note on mean testing for high dimensional multivariate data under non-normality[J]. *Statistica Neerlandica*, 2013, 67(1):81-99
- [31] Ahmad M R . A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens–Fisher setting[J]. *Annals of the Institute of Statistical Mathematics*, 2014
- [32] Katayama S , Kano Y , Srivastava M S . 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
- [33] Park J , Ayyala D N . A test for the mean vector in large dimension and small samples[J]. Journal of Statistical Planning & Inference, 2013, 143(5):929-943
- [34] Maboudou-Tchao E M , Silva I . Tests for mean vectors in high dimension[J]. Statistical Analysis & Data Mining, 2013, 6(6):578-598
- [35] Efang K, Lengyang W, Yingcun X, Jin L. A Permutation Test for Two-Sample Means and Signal Identification of High-dimensional Data[J]. Statistica Sinica (in press)
- [36] Efron B . Bootstrap Methods: Another Look at the Jackknife[J]. The Annals of Statistics, 1979, 7(1):1-26
- [37] Brown R M , Siler C D . Maximum Likelihood Bootstrap Topology[J]. 2013.
- [38] 陈文华, 李奇志, 张为鄂,等. 产品可靠性的 Bootstrap 区间估计方法[J]. 机械工程学报, 2003, 39(006):106-109
- [39] 罗巍, 张春华, 谭源源,等. 基于 Bootstrap 的可修系统贮存可用度近似置信下限评估方法 [N]. 兵工学报, 2010
- [40] Li D Q , Tang X S , Phoon K K . Bootstrap method for characterizing the effect of uncertainty in shear strength parameters on slope reliability[J]. Reliability Engineering & System Safety, 2015, 140(aug.):99-106
- [41] 胡正东, 曹渊, 张士峰,等. 特小子样试验下导弹精度评定的 Bootstrap 方法[J]. 系统工程与电子技术, 2008(08):1493-1497
- [42] 孙权, 周星, 冯静,等. 寿命分布的参数 Bootstrap 拟合优度检验方法[J]. 国防科技大学学报, 2014(6):82-113
- [43] 黄彩珠. 基于投影的两个总体均值向量的高维检验[D]. 广州大学, 2019
- [44] Johnston J E , Berry K J , Mielke P W . Permutation tests: precision in estimating probability values[J]. Perceptual & Motor Skills, 2007, 105(3 Pt 1):915

攻读硕士学位期间取得的成果

- [1] Jing Yang, Bo Chen, Quan Nie. Permutation Tests for Two-Sample Means of High-dimensional Data. IEEE CPS[C]. ITCA 2020