分类号______ 密级______

UDC______ 编号______

華中師範大學

硕士学位论文

基于 bootstrap 方法的高维非参数 Behrens-Fisher 问题的假设检验

学位申请人姓名: 贾婉茹

申请学位学生类别: 全日制硕士

申请学位学科专业: 统计学

指导教师姓名: 李正帮 副教授

硕士学位论文
MASTER'S THESIS

硕士学位论文

基于 bootstrap 方法的高维非参数 Behrens-Fisher 问题的假设检验

论文作者：贾婉茹

指导教师：李正帮 副教授

学科专业：统计学

研究方向：数理统计学

华中师范大学数学与统计学学院

2021 年 3 月

硕士学位论文
MASTER'S THESIS

Hypothesis test of high-dimensional nonparametric Behrens-Fisher problem based on bootstrap method

A thesis

submitted in Partial Fulfillment of the Requirements

for the M.S.Degree in Mathematics

By

Wanru Jia

Postgradute Program

School of Mathematics and Statistics

Central China Normal University

Supervisor **Zhengbang Li**

Academic Title **Associate Professor** **Signature** 李正帮

Approved

March 2021

硕士学位论文
MASTER'S THESIS

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: 贾婉茹　　　　日期: 2021年 5月 30日

学位论文版权使用授权书

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

作者签名: 贾婉茹　　　　导师签名: 李正帮

日期: 2021 年 5 月 30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: 贾婉茹　　　　导师签名: 李正帮

日期: 2021年 5 月 30日　　　　日期: 2021 年 6月 1 日

摘 要

随着科学技术的迅猛发展, 人类社会迎来"大数据"时代, 很多领域出现了高维数据, 然而传统的统计学方法不能简单的应用在高维数据上, 因此我们需要寻找新的统计方法来研究处理这些数据. 在研究非参数 Behrens-Fisher 问题时, 已有的检验方法 (如 O'Brien 秩和检验、修正的 O'Brien 秩和检验和最大秩检验) 都是研究低维数据下的渐进分布理论, 对于高维数据的情况目前还没有很好的方法来进行假设检验.

基于此本文提出了一种新的检验统计量来对高维数据下的非参数Behrens-Fisher问题进行假设检验. 由于在高维情况下要推导出新的统计量较为精确的均值估计与协方差估计较为困难, 因此本文运用 Stationary Bootstrap 方法检验新统计量的统计性质, 提出了运用 Stationary Bootstrap 方法计算新检验统计量 p 值的算法, 从而避免了推导均值估计与协方差估计的复杂过程. 通过 R 语言进行大量数据模拟实验并与 O'Brien 秩和检验、修正的 O'Brien 秩和检验和最大秩检验的模拟结果进行比较, 我们发现本文提出的检验统计量对于高维非参数 Behrens-Fisher 问题不仅可以较好的控制第 I 类错误, 还有非常好的检验功效.

关键词: 高维数据; 非参数 Behrens-Fisher 问题; Stationary Bootstrap.

Abstract

With the rapid development of science and technology, human society has ushered in the era of "big data", and high-dimensional data has appeared in many fields. However, traditional statistical methods cannot be simply applied to high-dimensional data, so we need to find new statistical methods to study and process these data. When studying the non-parametric Behrens-Fisher problem, the existing test methods, such as O'Brien's rank sum test, modified O'Brien's rank sum test and maximum rank test, are all researches on the theory of asymptotic distribution under low-dimensional data, so there is currently no good way to perform hypothesis testing in the case of high-dimensional data.

Based on this, this thesis proposes a new test statistic to perform hypothesis testing on the non-parametric Behrens-Fisher problem under high-dimensional data. In the case of high-dimensional data, new statistics are required to derive more accurate mean estimation and covariance. Estimation is more difficult, so this article uses the Stationary Bootstrap method to test the statistical properties of the new statistic, and proposes an algorithm to calculate the p-value of the new test statistic using the Stationary Bootstrap method, thereby avoiding the complicated process of deriving the mean estimation and the covariance estimation. Through a large number of data simulation experiments by R language and comparison with the simulation results of the O'Brien rank sum test, the modified O'Brien rank sum test and the maximum rank test. we find that the test statistics proposed in this thesis are more suitable for high-dimensional non-parametric Behrens-Fisher problem because of the better first type of error and the good test power.

Keywords: High-dimensional data; Non-parametric Behrens-Fisher problem; Stationary Bootstrap.

目录

第一章 引言

1.1 研究背景及意义

随着科学技术的迅猛发展, 人类社会迎来"大数据"时代. 大数据已经渗透到当今每一个行业和业务职能领域, 成为重要的生产因素, 在物理学, 生物医学、气象学、金融、环境生态学等的运用尤为突出. 海量的数据为我们带来更多信息的同时也为如何进行数据分析来提取有效的信息带来了巨大的挑战. 随着数据爆发式增长, 高维数据集(高维数、小容量)会越来越多, 比如在生物医学中研究人的基因序列变化与人体疾病表征, 往往是对几十或几百名患者进行数千个基因测序, 这就造成了样本容量 n 小于变量的维数 p. 传统的统计学方法研究的是变量的维数远远低于样本容量, 这种情况下可以保证样本协方差是正定的, 但是当变量的维数高于样本的容量时, 样本协方差却不一定是正定的, 这时传统的统计学方法不能简单的应用在高维数据上, 迫切需求新的统计方法来研究处理高维数据, 建立高维数据下的统计推断理论.

假设检验是统计推断的一个重要内容, 假设检验的基本原理是先对总体的特征做出某种假设, 然后根据从该总体中抽取的随机样本进行统计推理, 对此假设应该被拒绝还是接受做出推断. 假设检验理论在现实生活中的用途非常广泛, 例如: 检测一种新药能否减轻关节炎病人的疼痛; 研究土壤成分对植物生长的影响; 判断两种方案对公司发展前景的利弊等. 在现实生活中, 假设检验中的 Behrens-Fisher 问题在统计研究中经常遇到. 假设有两个总体分别服从正态分布 $N(\mu_1,\sigma_1)$ 和 $N(\mu_2,\sigma_2)$, 我们从中分别抽取样本容量为 n_1 和 n_2 的两个简单随机样本 $\mathbf{X}=(X_1,\cdots,X_{n_1})$ 和 $\mathbf{Y}=(Y_1,\cdots,Y_{n_2})$, 这两个样本相互独立且 $\sigma_1\neq\sigma_2$, 在 $\mu_1,\mu_2,\sigma_1,\sigma_2$ 均未知时, 根据样本我们要检验 $H_0:\mu_1=\mu_2\quad vs\quad H_1:\mu_1\neq\mu_2$. 此问题首先由著名统计学家 Fisher 进行研究, 因此被称为 Behrens-Fisher 问题. Behrens-Fisher 问题应用范围非常广泛, 如可应用于空间飞行器性能参数的地面试验与飞行试验的天地差分析、航天产品在不同试验条件的试验参数对比分析等.

然而, 在现实中并不是所以的样本都满足正态性条件, 这时我们考虑更一般的情况: 假设两个相互独立的随机样本 $\mathbf{X}=(X_1,\cdots,X_{n_1})$ 和 $\mathbf{Y}=(Y_1,\cdots,Y_{n_2})$ 分别服从于 $\mathbf{F}_1$ 和 $\mathbf{F}_2$, 根据样本我们要检验 $H_0:M_x=M_y\quad vs\quad H_1:M_x\neq M_y$. 其中 M_x, M_y 分别为 $\mathbf{X}$, $\mathbf{Y}$ 各维变量的中位数组成的向量, 这个问题称为非参数 Behrens-Fisher 问题. 非参数 Behrens-Fisher 问题在生物医学、心理学、人文科学、电气工程等方面均有广泛的应用, 如分析两种治疗药物对于人体疾病的治疗效果是否具有显著性差异、自我满意度对人体行为的影响等.

对于非参数 Behrens-Fisher 问题, 传统的检验方法如 Hotelling T^2 检验[1] 或多元 Wilcoxon 检验[2] 可能会因为总体的方差不等而产生错误的推断, 因此许多学者提出了新的检验方法. 1984 年 O'Brien 提出了非参数的秩和检验[3](下文简称为 O'Brien 秩和检验), 当样本量较大时, 秩和之间的相关性会变小, 因而 O'Brien 秩和检验可以将一般的多变量问题简化为单变量问题. O'Brien (1984) 还表明, 当样本量小于结果维数且分布偏斜或存在异常值时, 秩和检验是稳健的. Sankoh et al.(1999)[4] 还通过模拟评估了 O'Brien 秩和检验在各种协方差结构下的性能. O'Brien 秩和检验已广泛用于生物医学研究, 如 Kaufman et al.(1998)[5] 通过O'Brien 秩和检验对皮肤病的随机化临床试验的样本数据进行分析; Shames et al.(1998)[6] 通过 O'Brien 秩和检验对妇女经期前后哮喘的随机化试验的样本数据进行分析; Li et al.(2001)[7] 通过 O'Brien 秩和检验对多发性硬化症的随机化临床试验的样本数据进行分析, Tilley et al.(2000)[8] 通过 O'Brien 秩和检验对一系列类风湿性关节炎临床试验的数据进行二次分析等.

Brunner et al.(2002)[9] 提出了 Wald 检验和 ANOVA 检验, 用于一般的非参数 Behrens-Fisher 假设问题, 并提出描述一般的非参数 Behrens-Fisher 问题的具体形式为:

假设随机向量 $\mathbf{X} = (X_1, \cdots, X_p)^T$ 和 $\mathbf{Y} = (Y_1, \cdots, Y_p)^T$ 分别来自两个不同的总体, $\mathbf{X}$ 和 $\mathbf{Y}$ 的维数都为 p, 假设 $\mathbf{X}$ 和 $\mathbf{Y}$ 分别服从 $\mathbf{F}$ 分布和 $\mathbf{G}$ 分布, X_a 和 Y_a 的边际分布为 F_a 和 G_a, 其中

$$F_a(x) = P(X_a < x) + \frac{1}{2}P(X_a = x),$$

$$G_a(y) = P(Y_a < y) + \frac{1}{2}P(Y_a = y).$$

令 $\theta = (\theta_1, \cdots, \theta_p)^T$, 其中 $\theta_a = P(X_a < Y_a) - P(X_a > Y_a)$, $a \in \{1, \cdots, p\}$, 通常我们称 $\theta_a/2 = Pr(X_a < Y_a) - 1/2$ 为样本 $\mathbf{Y}$ 的第 a 个变量对于样本 $\mathbf{X}$ 的第 a 个变量的相对效应, 于是一般非参数 Behrens-Fisher 问题的原假设和备择假设可以定义为:

$H_0 : \theta = (0, \cdots, 0)^T_{p\times1} \quad vs \quad H_1 : \theta \neq (0, \cdots, 0)^T_{p\times1}$

另一种经常使用的限制性更强的原假设和备择假设可以写为:

$H_0' : \mathbf{F} = \mathbf{G} \quad vs \quad H_1' : \mathbf{F} \neq \mathbf{G}$

对于一般非参数 Behrens-Fisher 假设检验问题, Huang et al.(2005)[10] 表明, 当两组总体具有不同的方差时, O'Brien 秩和检验具有较高的第 I 类错误率. 他们在 O'Brien 秩和检验的基础上, 通过调整方差提出了新的统计量(下文简称修正的

O'Brien 秩和检验)以控制检验的第 I 类错误率, 利用帕金森综合症病人的随机化临床试验的样本数据进行实例分析, 并得出相应结论.

Liu et al. (2010)[11] 指出在原假设不成立的情况下, 如果两个样本之间的结果差异方向不同时, O'Brien 秩和检验统计量和修正的 O'Brien 秩和检验统计量测试会遭受巨大的功效亏损. 为了克服这个问题, 他们提出了一个简单而强大的统计量: 列平均秩和差的最大值(以下简称最大秩检验统计量), 通过大量的仿真模拟表明无论差异的方向如何, 该统计量都可以保持令人满意的功率, 并且能控制第 I 类错误. 利用自闭症或自闭症谱系障碍 (ASD) 儿童与生长相关的激素数据进行实例分析, 得出较好的统计结论, 但是此检验却只适用于变量维度 p 小于样本容量 n 的情况.

杨贵军等(2015)[12] 指出 O'Brien 秩和检验统计量和修正的 O'Brien 秩和检验统计量在大样本的情况下近似于服从正态分布. 但是, 当样本量较小的时候, 无法近似于服从正态分布, 若样本量较小仍使用这两种统计量所得出的检验结果偏差较大, 可能导致结论不正确. 于是他们对 O'Brien 秩和检验统计量进行改进, 使用的方法是对观测值的秩进行正态转换, 使得转换后的检验统计量的真实分布与正态分布的偏离程度较小, 从而使得假设检验的统计量服从于正态分布, 提高检验统计量的功效, 但是此检验仍只适用于变量维度 p 小于样本容量 n 的情况.

已有的针对一般非参数 Behrens-Fisher 问题的检验方法都是研究低维数据下的渐进分布理论, 在高维数据下这些统计量不具备良好的统计性质, 基于此本文提出了一种新的统计量来对高维数据下的非参数 Behrens-Fisher 问题进行检验, 并通过数据模拟与 O'Brien 秩和检验统计量、修正的 O'Brien 秩和检验统计量和最大秩检验统计量进行比较, 从而检测本文提出的统计量的统计性质.

1.2 研究现状

假设第一个总体有 m 个观测值, 第二个总体有 n 个观测值, 则两个总体共有 $N = m + n$ 个观测值. 令 $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$, $i \in \{1, \cdots, m\}$ 为总体 $\mathbf{X}$ 的第 i 个样本观测值, $Y_j = (y_{j1}, y_{j2}, \cdots, y_{jn})^T$, $j \in \{1, \cdots, n\}$ 为总体 $\mathbf{Y}$ 的第 j 个样本观测值. 对于总体 $\mathbf{X}$ 和总体 $\mathbf{Y}$ 的第 a 个变量, 我们结合这两组样本将 $\{x_{1a}, \cdots, x_{ma}, y_{1a}, \cdots, y_{na}\}$ 这 N 个观测值进行排序, 所得观测值 $\{x_{1a}, \cdots, x_{ma}\}$ 对应的秩为 $\{R_{x1a}, \cdots, R_{xma}\}$, 观测值 $\{y_{1a}, \cdots, y_{na}\}$ 对应的秩为 $\{R_{y1a}, \cdots, R_{yna}\}$, 分别记为 R_{xia}, R_{yja}, $i \in \{1, \cdots, m\}, j \in \{1, \cdots, n\}$.

1.2.1 O'Brien 秩和检验

O'Brien(1984)[3] 将两个样本 t 检验的原理应用于所得观测值的秩中, 令 $R_{xi\cdot}=\sum_{a=1}^{p}R_{xia}$, $R_{yj\cdot}=\sum_{a=1}^{p}R_{yja}$, 将多因素样本转化为对秩和 $R_{xi\cdot}$, $R_{yj\cdot}$ 的样本比较, 而后采用学生化 t 检验方法进行假设检验, 提出的检验统计量为:

$$T_1=\frac{\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot}}{\hat{\sigma}\sqrt{\frac{1}{m}+\frac{1}{n}}} \tag{1.1}$$

$$T_2=\frac{\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot}}{\sqrt{\frac{\hat{\sigma}^2_{x\cdot\cdot}}{m}+\frac{\hat{\sigma}^2_{y\cdot\cdot}}{n}}} \tag{1.2}$$

其中 $\bar{R}_{x\cdot\cdot}=\frac{1}{m}\sum_{i=1}^{m}R_{xi\cdot}$, $\bar{R}_{y\cdot\cdot}=\frac{1}{n}\sum_{j=1}^{n}R_{yj\cdot}$, $\hat{\sigma}^2_{x\cdot\cdot}=\frac{1}{m-1}\sum_{i=1}^{m}(R_{xi\cdot}-\bar{R}_{x\cdot\cdot})^2$, $\hat{\sigma}^2_{y\cdot\cdot}=\frac{1}{n-1}\sum_{j=1}^{n}(R_{yj\cdot}-\bar{R}_{y\cdot\cdot})^2$, $\hat{\sigma}^2=\frac{(m-1)\hat{\sigma}^2_{x\cdot\cdot}+(n-1)\hat{\sigma}^2_{y\cdot\cdot}}{N-2}$.

在原假设 $H_0':\mathbf{F}=\mathbf{G}$ 成立时, 统计量 T_1 服从自由度为 $m+n-2$ 的 t 分布, 统计量 T_2 服从自由度为 $[\zeta^2/(m-1)+(1-\zeta)^2/(n-1)]^{-1}$ 的 t 分布, 其中 $\zeta=(\hat{\sigma}^2_{x\cdot\cdot}/m)/(\hat{\sigma}^2_{x\cdot\cdot}/m+\hat{\sigma}^2_{y\cdot\cdot}/n)$.

Huang et al.(2005)[10] 指出, 在限制性更强的零假设 $H_0':\mathbf{F}=\mathbf{G}$ 成立的条件下, T_1 和 T_2 都渐进地服从标准正态分布. 当 $\mathbf{F}\neq\mathbf{G}$ 时, 这两个统计量渐进服从正态分布. 而当用于检验广义 Behrens-Fisher 零假设 $H_0:\theta=(0,\cdots,0)^T_{p\times 1}$, 即当两组数据具有不同的方差时, O'Brien 秩和检验具有较高的第 I 类错误率.

1.2.2 修正的 O'Brien 秩和检验

为了使 O'Brien 秩和检验统计量即 (1.1) 和 (1.2) 适合于检验零假设 $H_0:\theta=(0,\cdots,0)^T_{p\times 1}$ 的情况, Huang et al.(2005)[10] 推导了两个统计量的渐近方差, 并证明了在零假设 H_0 成立条件下, 当 $m/n\longrightarrow\lambda$, $0<\lambda<+\infty$, $N=(m+n)\longrightarrow+\infty$ 时, 统计量 T_1 渐进服从均值为0, 方差为 h_1 的正态分布, 统计量 T_2 渐进服从均值为0, 方差为 h_2 的正态分布. 其中

$$h_1=\frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}[e_{ab}\lambda^3+(d_{ab}+2f_{ab})\lambda^2+(c_{ab}+2\eta_{ab})\lambda+\xi_{ab}]}$$

$$h_2 = \frac{\sum_{a=1}^{p}\sum_{b=1}^{p}(1+\lambda)^2(c_{ab}+d_{ab}\lambda)}{\sum_{a=1}^{p}\sum_{b=1}^{p}[d_{ab}\lambda^3+(e_{ab}+2\eta_{ab})\lambda^2+(\xi_{ab}+2f_{ab})\lambda+c_{ab}]}$$

$$c_{ab}=cov(G_a(X_a),G_b(X_b)), d_{ab}=cov(F_a(Y_a),F_b(Y_b)), e_{ab}=cov(F_a(X_a),F_b(X_b))$$

$$f_{ab}=cov(F_a(X_a),G_b(X_b)), \xi_{ab}=cov(G_a(Y_a),G_b(Y_b)), \eta_{ab}=cov(G_a(Y_a),F_b(Y_b))$$

因此提出以下两个修改后的检验统计量:

$$T_3 = \frac{\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot}}{\hat{\sigma}\sqrt{\hat{h}_1(\frac{1}{m}+\frac{1}{n})}} \tag{1.3}$$

$$T_4 = \frac{\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot}}{\sqrt{\hat{h}_2(\frac{\hat{\sigma}^2_{x\cdot\cdot}}{m}+\frac{\hat{\sigma}^2_{y\cdot\cdot}}{n})}} \tag{1.4}$$

在原假设 $H_0:\theta=(0,\cdots,0)^T_{p\times1}$ 成立时, 若样本量足够大, 统计量 T_3 和 T_4 均渐进服从标准正态分布, 并且都能很好地控制第 I 类错误率.

Liu ct al.(2010)[11] 指出在原假设不成立的情况下, 当 θ_a ($a\in\{1,\cdots,p\}$) 方向相同, 即 θ_a 同正或者同负时, O'Brien 秩和检验统计量和修正的 O'Brien 秩和检验统计量有良好的功效, 但是当 θ_a ($a\in\{1,\cdots,p\}$) 方向不同, 即 θ_a 正负相间时, O'Brien 秩和检验统计量和修正的 O'Brien 秩和检验统计量测试会遭受巨大的功效亏损. 假设对于修正的 O'Brien 秩和检验统计量的双边检验的检验水平为 α, 令 $\bar{R}_{x\cdot a}=\frac{1}{m}\sum_{i=1}^{m}R_{xia}$, $\bar{R}_{y\cdot a}=\frac{1}{n}\sum_{i=1}^{m}R_{yja}$, 则 $\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot}=\sum_{a=1}^{p}(\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a})$, 那么 $E(\bar{R}_{y\cdot\cdot}-\bar{R}_{x\cdot\cdot})=\frac{m+n}{2}\sum_{a=1}^{p}\theta_a$. 显然, 对于子空间 $\{\theta_a:\sum_{a=1}^{p}\theta_a=0,\theta_a$ 不全为0$\}$, 统计量 T_1, T_2, T_3, T_4 的功效是渐进为 α 的. 例如, 当 $p=2$, $\theta_1=-\theta_2\gg0$ 时, 统计量测试的功效仍然大致为 α.

1.2.3 最大秩检验

Liu et al.(2010)[11] 为了避免功效的损失提出了一个新的统计量:

$$T_5=\max_{a\in\{1,\cdots,p\}}(|\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a}|) \tag{1.5}$$

在原假设 H_0 成立的条件下, 当 $min\{m,n\}\to\infty$ 且 $0<\frac{m}{n}\to\lambda_0<+\infty$ 时 $(\bar{R}_{y\cdot1}-\bar{R}_{x\cdot1},\cdots,\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p})'$ 服从均值为 $(0,\cdots,0)'$, 相关系数矩阵为 $\Lambda=$

$(\rho_{ab})_{p\times p}$ 的多元正态分布. 其中 $\rho_{ab} = c_{ab} + \lambda_0 d_{ab}/\sqrt{[c_{aa}+\lambda_0 d_{aa}][c_{bb}+\lambda_0 d_{bb}]}$, $c_{ab} = cov(G_a(X_a), G_b(X_b))$, $d_{ab} = cov(F_a(Y_a), F_b(Y_b))$.

如果 $T_{max} > c_{max}$ 则拒绝原假设, 根据 $Pr_{H_0}(T_{max} > t) = 1 - Pr_{H_0}(|\bar{R}_{y\cdot 1} - \bar{R}_{x\cdot 1}| < t, \cdots, |\bar{R}_{y\cdot p} - \bar{R}_{x\cdot p}| < t)$ 和多元积分, 我们可以评价 T_{max} 的显著性水平.

统计量 T_3, T_4 可以看成为 $\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}$ 的线性组合, 而统计量 T_{max} 可以看成为 $\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}$ 的非线性组合. 如果观察到的 Y 样本相对于 X 样本的相对效应较大, 则会拒绝 H_0, 因此无论相对效应的方向如何, 都有望使测试保持令人满意的功效. 因此测试的结果不受 θ_a 方向的影响, 可以保持很好的功效. 但是此统计量有一个使用前提 $m, n > p$.

1.3 本文的主要内容及组织结构

现阶段, 对于非参数 Behrens-Fisher 问题研究的统计量仅适用于样本容量 n 大于数据维数 p $(n > p)$ 时, 而当 $n < p$ 时已提出的统计量则失去了良好的统计性质, 本文所提出的统计量可以很好的解决当 $n < p$ 时的非参数 Behrens-Fisher 问题. 在传统方法下检验统计量的优劣时我们往往需要推导统计量的近似分布从而计算其第一类错误率和功效, 而在推导高维统计量分布的均值和协方差时往往较为复杂, 需要良好的数学功底与统计知识, 利用 Stationary Bootstrap 方法我们可以直接计算统计量的第一类错误率和功效, 避免了复杂的估计均值和协方差的过程.

本文共分为五章, 全文组织结构如下:

第一章首先介绍了本文研究背景及意义, 然后详细地叙述了非参数 Behrens-Fisher 问题的国内外研究现状及其现有的主要参考文献, 回顾了非参数 Behrens-Fisher 问题的三个检验方法, 最后对本文的研究内容以及组织结构做了详细说明.

第二章详细介绍了 Stationary Bootstrap 抽样方法, 并介绍了相关数据(均值)的 Stationary Bootstrap 算法.

第三章在秩和检验的基础上提出了一个新的统计量对非参数 Behrens-Fisher 问题进行假设检验, 运用 Stationary Bootstrap 方法计算统计量的第 I 类错误率和功效, 并对算法做了详细说明, 最后对所需的协方差进行了推导, 并提出了其渐近协方差矩阵的估计程序.

第四章利用 R 软件对前文介绍的 5 种统计量及新的统计量进行模拟, 在正态分布和 Laplace 分布两种模型下分别计算出这六种统计量的第 I 类错误率和检验功效, 并对模拟的结果进行比较分析.

第五章对本文的研究作出总结, 并对今后进一步研究提出意见和建议.

本文的创新点:

1. 对于一般非参数 Behrens-Fisher 问题, 目前国内外所提出的方法在 $n > p$ 时具有良好的性质, 而当 $n < p$ 时还没有提出较好的统计量去进行假设检验. 本文提出的方法可以较好的解决这个问题.

2. 利用 Stationary Boostrap 方法去计算统计量的第 I 类错误率和功效, 避免了复杂的推导过程.

第二章 相关数据的 Stationary Bootstrap 算法

Bootstrap 方法是 1979 年美国统计学家 Efron[13] 提出的原理较为简单、直接的一种方法, 其本质是样本的重抽样. Bootstrap 方法并不需要对总体分布作假设或事先推导估计量的解析式, 只需要重构样本并不断计算估计值, 显然, 它本质上是一种非参数方法[14]. 该方法在四十多年间已经得到了极大的发展并被广泛地应用于统计学的各个领域, 其研究方向包括独立同分布数据的 Bootstrap[13], 基于回归分析的 Bootstrap[13], 块状 Bootstrap[15], Sieve Bootstrap[16] 和基于变换的 Bootstrap[17] 等.

2.1 Stationary Bootstrap 算法

Stationary Bootstrap 算法是Politis & Romano(1994)[18] 提出的块状 Bootstrap 方法的运用, 其抽样方法如下:

假设有 p 个观测值 $T=\{t_1,\cdots,t_p\}$, 这 p 个观测值是严格稳健弱相关的时间序列, 令 q 为 $[0,1]$ 上的一个常数, 在我们对这 p 个观测值进行 Stationary Bootstrap 抽样之前, 对来自 T 的子样本组成的块做如下定义: 不妨记每个块为 $B_{I_i,k_i}=\{t_{I_i},\cdots,t_{I_i+k_i-1}\}$, 其中 $i\in\{1,2,\cdots\}$, I_i 是来自离散均匀分布 $\{1,2,\cdots,p\}$ 的独立同分布样本, 表示 B_{I_i,k_i} 这个块的第一个元素是 t_{I_i}; k_i 是独立服从几何分布的随机变量序列且 $P(k_i=m)=q(1-q)^{m-1}$ $(m\in\{1,2,\cdots\})$, 表示这个块的元素是从原观测值中抽取的由 t_{I_i} 开始往后的 k_i 个观测值.

抽取样本的过程为: 先生成服从离散均匀分布总体 $\{1,2,\cdots,p\}$ 的数值 I_1, 再生成服从 $q(1-q)^{m-1}$ $(m\in\{1,2,\cdots\})$ 的几何分布数值 k_1, 则第一个块里就有 k_1 个元素, 令 $t_1^*=t_{I_1}$ 作为 Stationary Bootstrap 方法抽取的第一个样本且作为第一个块里的第一个元素, 并将其在 T 中原始位置的后 k_1-1 个观测值全放入第一个块里为 $t_2^*,\cdots,t_{k_1}^*$. 接着生成服从离散均匀分布总体 $\{1,2,\cdots,p\}$ 的数值 I_2和服从 $q(1-q)^{m-1}$ $(m\in\{1,2,\cdots\})$ 的几何分布数值 k_2, 令 $t_{k_1+1}^*=t_{I_2}$ 作为第二个块里的第一个元素, 并将其在 T 中原始位置的后 k_2-1 个观测值全放入第二个块里为 $t_{k_1+2}^*,\cdots,t_{k_1+k_2}^*$. 第三个块以及之后的 Stationary Bootstrap 方法抽取的样本按照此原则进行下去直到 $k_1+k_2+\cdots\geq p$ 为止. 若 $k_1+k_2+\cdots>p$, 我们只需要取前 p 个样本作为 Stationary Bootstrap 方法抽取的样本. 如果 $I_i+a_i>p$, 则 $t_{I_i+a_i}=t_{I_i+a_i-p}, 0<a_i\leq k_i-1$.

一旦 $t_1^*,\cdots,t_p^*$ 生成, 我们可以用生成的伪时间序列来计算我们感兴趣的量 ($R_p(t_1^*,\cdots,t_p^*;\ T)$). 然后以完全相同的方式重复抽样模拟 B 次伪时间序列, 通过 B 次 $R_p(t_1^*,\cdots,t_p^*;T)$ 的经验分布来近似 $R_p(t_1,\cdots,t_p;\mu)$ 的真实分布.

令 $\theta=\mu$, $\hat{\theta}=\bar{T}$, 其中 μ 是$\{t_1,\cdots,t_p\}$ 联合分布的均值, $\hat{\theta}$ 是参数 θ 的估计量, 且 $\bar{T}=\frac{1}{p}\sum\limits_{i=1}^{p}t_i$. 假设 Stationary Bootstrap 方法抽取出来的样本为 $t_1^*,\cdots,t_p^*$, 记 $\hat{\theta}^*=\bar{T}^*,\bar{T}^*=\frac{1}{p}\sum\limits_{i=1}^{p}t_i^*$, 考虑统计量 $W_p=p^{1/2}[\hat{\theta}-\theta]$, 根据 Politis & Romano(1994)[18] 有:

$$E(W_p^*)=E(p^{1/2}[\hat{\theta}^*-\hat{\theta}])=0$$

更多理论证明详见 Politis & Romano(1994)[18].

2.2 Stationary Bootstrap 算法的应用

2.2.1 Stationary Bootstrap 方法应用原理

在许多研究领域中我们需要检验来自两个总体的高维随机样本的均值是否相同. 假设 $X^{(1)}=(x_{11},\cdots,x_{1p})^T,\cdots,X^{(m)}=(x_{m1},\cdots,x_{mp})^T$ 是来自总体均值为 $\mu_1=(\mu_{11},\cdots,\mu_{1p})^T$, 协方差阵为 $\Sigma_1=(\sigma_{1:q,s})_{q,s\in\{1,\cdots,p\}}$ 的独立样本, $Y^{(1)}=(y_{11},\cdots,y_{1p})^T,\cdots,Y^{(n)}=(y_{n1},\cdots,y_{np})^T$ 是来自总体均值为 $\mu_2=(\mu_{21},\cdots,\mu_{2p})^T$, 协方差阵为 $\Sigma_2=(\sigma_{2:q,s})_{q,s\in\{1,\cdots,p\}}$ 的独立样本. 其中 μ_1,μ_2 未知, Σ_1,Σ_2 是未知的非奇异矩阵. 于是双样本均值检验问题为:

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

记 $\bar{\mathbf{X}}=(\bar{X}_1,\cdots,\bar{X}_p)^T$, $\bar{\mathbf{Y}}=(\bar{Y}_1,\cdots,\bar{Y}_p)^T$, 其中 $\bar{X}_k=\frac{1}{m}\sum\limits_{i=1}^{m}x_{ik}$, $\bar{Y}_k=\frac{1}{n}\sum\limits_{j=1}^{n}y_{jk}$, $k=1,\cdots,p$, $S_1=\frac{1}{m-1}\sum\limits_{i=1}^{m}(X^{(i)}-\bar{\mathbf{X}})(X^{(i)}-\bar{\mathbf{X}})^T=(\hat{\sigma}_{1:q,s})_{q,s\in\{1,\cdots,p\}}$, $S_2=\frac{1}{n-1}\sum\limits_{j=1}^{n}(Y^{(j)}-\bar{\mathbf{Y}})(Y^{(j)}-\bar{\mathbf{Y}})^T=(\hat{\sigma}_{2:q,s})_{q,s\in\{1,\cdots,p\}}$.

当 $m,n>p$ 即样本容量大于样本维数时, 我们可以用传统的 Hotelling T^2 检验统计量来处理双样本均值检验问题; 当 $m,n<p$ 时,传统的 Hotelling T^2 检验统计量没有定义, 后续有许多学者进一步研究了此问题, 提出了很多统计量, 其中有一类统计量是关于两均值向量之差 $\bar{\mathbf{X}}-\bar{\mathbf{Y}}$ 的形式, 这种类型的检验统计量对于检验 $\mu_1-\mu_2$ 非零成分相对密集的差异是非常有效的. 当列式变量的方差相同时, Bai and Saranadasa(1996)[19], Chen and Qin(2010)[20] 提出了非学生化 $\{\bar{X}_1-\bar{Y}_1,\cdots,\bar{X}_p-\bar{Y}_p\}$

列统计量平方和的统计量; 当列式变量的方差不同时, Srivastava et al.(2013)[21], Gregory et al.(2015)[22] 提出了学生化 $\{\frac{\bar{X}_1-\bar{Y}_1}{\sqrt{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n}},\cdots,\frac{\bar{X}_p-\bar{Y}_p}{\sqrt{\hat{\sigma}_{1:p,p}/m+\hat{\sigma}_{2:p,p}/n}}\}$ 各项平方之和的统计量.

在实践中, 我们不知道所有列式变量的方差是否相同. 因此, 基于研究学生化列统计量的平方和可能是稳健的. 为了做统计推断, 学生化列统计量需要估计 trR_1, trR_2, trR_1^2, trR_2^2 和 trR_1R_2, 其中 R_1, R_2 分别是 Σ_1, Σ_2 的相关矩阵, 这个计算过程非常复杂, 基于此 Li et al.(2020)[23] 提出了利用Stationary Bootstrap 方法来计算学生化列统计量的平方和统计量的第 I 类错误率和功效, 其具体原理如下:

当 $p<m$、n 时, $(m+n)^{\frac{1}{2}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})$ 渐进正态分布[24], 当 $p>m$, n 时, Li et al.(2020)[23] 推导了在特定条件下 $(m+n)^{\frac{1}{2}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})$ 的渐进分布:

设 p 维实数向量 $\mathbf{l}=(l_1,\cdots,l_p)'_{p\times 1}$ 满足 $\parallel l\parallel_2^2=\sum\limits_{s=1}^{p} l_s^2<M_p$ (M_p 是与 p 相关的有限正实数), 定义 $\mathbf{V}=\mathbf{l}'(\bar{\mathbf{X}}-\bar{\mathbf{Y}})$, 且满足条件 $\lim\limits_{m,n\to\infty}\frac{m}{m+n}=\theta(0<\theta<1)$. 如果对于任意正整数 p, 存在一个正实数 α, 满足条件 $\frac{l'\Sigma_1 l}{p^\alpha}<C$ 和 $\frac{l'\Sigma_2 l}{p^\alpha}<C$, 其中 C 是与 p 无关的有限正实数. 记 $h_1(l_1,\cdots,l_p)$ 为 $(\frac{(m+n)^{\frac{1}{2}}(\bar{X}_1-\bar{Y}_1)}{p^{\alpha/2}},\cdots,\frac{(m+n)^{\frac{1}{2}}(\bar{X}_p-\bar{Y}_p)}{p^{\alpha/2}})$ 的特征函数. 对于任意的固定正数 p, 在零假设下有

$$\lim_{m,n\to\infty} log[h_1(l_1,\cdots,l_p)]=-\frac{l'(\frac{\Sigma_1}{\theta}+\frac{\Sigma_2}{1-\theta})l}{2p^\alpha}.$$

因此当 $p>m$, n 或 $m+n$ 且 $m+n\to\infty$ 时, 统计量 $(m+n)^{\frac{1}{2}}(\bar{\mathbf{X}}-\bar{\mathbf{Y}})$ 是渐进多元正态分布的. 当 m, n 很大,并且维数 p 与样本容量 m, n 独立时 $\{(m+n)^{\frac{1}{2}}(\bar{X}_1-\bar{Y}_1),\cdots,(m+n)^{\frac{1}{2}}(\bar{X}_p-\bar{Y}_p)\}$ 渐进一个高斯序列, 因此可以将 Stationary Bootstrap 算法运用于学生化列统计量平方和的统计量上.

2.2.2 Stationary Bootstrap 算法检验步骤

Li et al.(2020)[23] 在文中指出: 因为没有计算估计量 $\{\hat{\sigma}_{1:1,1},\cdots,\hat{\sigma}_{1:p,p},\hat{\sigma}_{2:1,1},\cdots,\hat{\sigma}_{2:p,p}\}$ 具体的值, 所以没有给出 $t_{s,k}=\frac{(\bar{X}_k-\bar{Y}_k)^2}{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n}$, $k\in\{1,\cdots,p\}$ 的渐进分布. 但是 $\{\hat{\sigma}_{1:1,1},\cdots,\hat{\sigma}_{1:p,p},\hat{\sigma}_{2:1,1},\cdots,\hat{\sigma}_{2:p,p}\}$ 是 $\sigma_{1:1,1},\cdots,\sigma_{1:p,p}$, $\sigma_{2:1,1},\cdots$, $\sigma_{2:p,p}$ 的一致估计, 因此猜想 $\{t_{s,j}\}_{j=1}^p$ 的渐进分布可能为多元正态分布. 而 Stationary Bootstrap 算法只关注计算统计量的第 I 类错误率和功效, 因此没有给出 $\{t_{s,k}\}_{k=1}^p$ 的渐进分布为多元正态分布的严格证明.

假设 $\{t_{s,1}=\frac{(\bar{X}_1-\bar{Y}_1)^2}{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n},\cdots,t_{s,p}=\frac{(\bar{X}_p-\bar{Y}_p)^2}{\hat{\sigma}_{1:p,p}/m+\hat{\sigma}_{2:p,p}/n}\}$ 是渐进严平稳弱相关时间序列, 其均值在零假设下都渐进为 1, 则基于 Stationary Bootstrap 算法的检验过

程如下:

(1). 计算 p 个学生化列统计量的值 $t_{s,1}=\frac{(\bar{X}_1-\bar{Y}_1)^2}{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n},\cdots,t_{s,p}=\frac{(\bar{X}_p-\bar{Y}_p)^2}{\hat{\sigma}_{1:p,p}/m+\hat{\sigma}_{2:p,p}/n}$, 并计算统计量

$$S_0=\frac{p(\bar{t}_s-1)^2}{\bar{v}_s}$$

其中 $\bar{t}_s=\frac{1}{p}\sum_{k=1}^{p}t_{s,k}$, $\bar{v}_s=\frac{1}{p-1}\sum_{k=1}^{p}(t_{s,k}-\bar{t}_s)^2$.

(2). 按照Stationary Bootstrap方法抽样的原则从 $t_{s,1}=\frac{(\bar{X}_1-\bar{Y}_1)^2}{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n},\cdots,t_{s,p}=\frac{(\bar{X}_p-\bar{Y}_p)^2}{\hat{\sigma}_{1:p,p}/m+\hat{\sigma}_{2:p,p}/n}$ 中抽取 p 个样本, 记为 $t^*_{s,1},\cdots,t^*_{s,p}$, 并计算其均值 $\bar{t}_{s,(1)}=\frac{1}{p}\sum_{k=1}^{p}t^*_{s,k}$ 和方差 $\bar{v}_{s,(1)}=\frac{1}{p-1}\sum_{k=1}^{p}(t^*_{s,k}-\bar{t}_{s,(1)})^2$, 接着计算 Bootstrap 统计量的值, 记为

$$S^*_{(1)}=\frac{p(\bar{t}_{s,(1)}-\bar{t}_s)^2}{\bar{v}_{s,(1)}}.$$

(3). 重复步骤 (2) B 次, 得到 B 个 Bootstrap 统计量值, 记为 $S^*_{(1)},S^*_{(2)},\cdots,S^*_{(B)}$.

(4). 计算 P 值:

$$P=\frac{\sum_{b=1}^{B}I_{\{S^*_{(b)}>S_0\}}}{B}.$$

第三章 基于 Stationary Bootstrap 方法的非参数 Behrens-Fisher 检验

在本章我们在秩和检验的基础上提出了一个新的统计量对非参数 Behrens-Fisher 问题进行检验, 运用 Stationary Bootstrap 算法计算统计量的第 I 类错误和功效, 并对算法做了详细说明, 最后对 U 型统计量 (U_a) 的协方差进行了推导, 并提出了其渐近协方差矩阵的估计程序.

假设随机向量 $\mathbf{X}=(X_1,\cdots,X_p)^T$ 和 $\mathbf{Y}=(Y_1,\cdots,Y_p)^T$ 分别来自两个不同的总体, $\mathbf{X}$ 和 $\mathbf{Y}$ 的维数都为 p, 假设 $\mathbf{X}$ 和 $\mathbf{Y}$ 分别服从 $\mathbf{F}$ 分布和 $\mathbf{G}$ 分布, X_a 和 Y_a 的边际分布为 F_a 和 G_a, 其中

$$F_a(x)=P(X_a<x)+\frac{1}{2}P(X_a=x),$$

$$G_a(y)=P(Y_a<y)+\frac{1}{2}P(Y_a=y), a\in\{1,\cdots,p\}.$$

令 $\theta=(\theta_1,\cdots,\theta_p)^T$, 其中 $\theta_a=P(X_a<Y_a)-P(X_a>Y_a)$, 于是, 一般非参数 Behrens-Fisher 问题的原假设和备择假设可以定义为:

$H_0:\theta=(0,\cdots,0)^T_{p\times1}\quad vs\quad H_1:\theta\neq(0,\cdots,0)^T_{p\times1}$

假设第一个总体有 m 个观测值, 第二个总体有 n 个观测值, 则两个总体共有 $N=m+n$ 个观测值. 令 $X_i=(x_{i1},x_{i2},\cdots,x_{ip})^T, i\in\{1,\cdots,m\}$ 为总体 $\mathbf{X}$ 的第 i 个样本观测值, $Y_j=(y_{j1},y_{j2},\cdots,y_{jp})^T, j\in\{1,\cdots,n\}$ 为总体 $\mathbf{Y}$ 的第 j 个样本观测值. 对于总体 $\mathbf{X}$ 和总体 $\mathbf{Y}$ 的第 a 个变量, 我们结合这两组样本将 $\{x_{1a},\cdots,x_{ma},y_{1a},\cdots,y_{na}\}$ 这 N 个观测值进行排序, 所得观测值 $\{x_{1a},\cdots,x_{ma}\}$ 对应的秩为 $\{R_{x1a},\cdots,R_{xma}\}$, 观测值 $\{y_{1a},\cdots,y_{na}\}$ 对应的秩为 $\{R_{y1a},\cdots,R_{yna}\}$, 分别记为 $\{R_{xia},R_{yja},\ i\in\{1,\cdots,m\},\ j\in\{1,\cdots,n\}\}$. 然后计算 $\bar{R}_{x\cdot a}=\frac{1}{m}\sum\limits_{i=1}^{m}R_{xia}$, $\bar{R}_{y\cdot a}=\frac{1}{n}\sum\limits_{j=1}^{n}R_{yja}$, 其中 $a\in\{1,\cdots,p\}$.

令 $U_a=\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a}$, 其为 U 型统计量, 经过简单的计算(具体推算过程见本章第3节)我们可得

$$U_a=\frac{m+n}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I_{\{x_{ai}<y_{aj}\}}-I_{\{x_{ai}>y_{aj}\}}],\ a\in\{1,\cdots,p\}.$$

令 $\mathbf{U}=(U_1,U_2,\cdots,U_p)^T_{p\times1}$, $\Sigma=(\sigma_{ab})_{a,b\in\{1,\cdots,p\}}$ 为 $\mathbf{U}$ 的协方差矩阵. 当样本的规模 p 小于样本容量 m, n 时, Huang et al.(2005)[10] 提出基于 $\sum\limits_{a=1}^{p}U_a$ 的统计量并研

究了它的低维渐进分布理论, 推导出随机向量 $\mathbf{U}$ 的渐进分布并且提出了其渐进协方差矩阵的估计方法. Liu et al.(2010)[11] 提出基于 $\max\limits_{s=1,\cdots,p}\{|U_a|\}$ 的统计量并研究了它的低维渐进分布理论. 但当样本的规模 p 变大甚至大于样本量 m, n 时, O'Brien (1984)[3] 和 Huang et al.(2005)[10] 中的四个检验因为 $\theta_a(a\in\{1,\cdots,p\})$ 方向的不同可能具有较低的功效, 而 Liu et al.(2010)[11] 提出的检验由于 m, $n<p$ 可能会有较高的第 I 类错误.

3.1 非参数 Behrens-Fisher 问题的检验统计量

当 $p>m$, n 时, 受高维数据下的双样本均值检验问题的学生化列统计量 $\{\frac{(\bar{X}_1-\bar{Y}_1)^2}{\sqrt{\hat{\sigma}_{1:1,1}/m+\hat{\sigma}_{2:1,1}/n}},\cdots,\frac{(\bar{X}_p-\bar{Y}_p)^2}{\sqrt{\hat{\sigma}_{1:p,p}/m+\hat{\sigma}_{2:p,p}/n}}\}$ 平方和统计量[21, 22] 的启发, 我们提出了一种新的统计量:

$$T_{BS}=Z_1+\cdots+Z_q=\frac{U_1^2}{\hat{\sigma}_{11}}+\frac{U_2^2}{\hat{\sigma}_{22}}+\cdots+\frac{U_p^2}{\hat{\sigma}_{pp}}=\sum_{a=1}^{p}\frac{U_a^2}{\hat{\sigma}_{aa}} \tag{3.1}$$

其中 $U_a=\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a}, a\in\{1,\cdots,p\}$, $\hat{\sigma}_{aa}$ 为 σ_{aa} 的估计 (见本章第 3 节).

Huang et al.(2005)[10] 推导了在原假设 H_0 成立的条件下, $(\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1},\cdots,\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p})'$ 的渐进协方差矩阵的估计为 $\hat{\sigma}=\hat{\sigma}_{ab}$, $a,b\in\{1,\cdots,p\}$. 其中

$$\hat{\sigma}_{ab}=\begin{cases}\dfrac{N^2[(n-1)\widehat{cov}(G_a(X_a),G_b(X_b))+(m-1)\widehat{cov}(F_a(Y_a),F_b(Y_b))+\eta]}{mn}, & a\neq b\\[2ex] \dfrac{N^2[(n-1)\widehat{cov}(G_a(X_a),G_a(X_a))+(m-1)\widehat{cov}(F_a(Y_a),F_a(Y_a))+0.25]}{mn}, & a=b\end{cases}$$

类似于高维数据下的双样本均值检验问题中的学生化列统计量的平方和统计量, 我们猜想在一定条件下, 当零假设成立时有:

$$Z_{SB}=\frac{T_{BS}-p}{\sqrt{\widehat{Var(T_{BS})}}}\xrightarrow{L}N(0,1) \tag{3.2}$$

由于计算统计量 T_{BS} 的渐进方差过于复杂, 此处我们不做具体证明, 只是猜想. 参考 Li et al.(2020)[23] 中将 Stationary Bootstrap 方法运用到高维数据双样本均值检验的原理, 在本文我们运用 Stationary Bootstrap 方法求统计量的第 I 类错误率和功效, 由于 Stationary Bootstrap 方法需要进行大量的重抽样, 我们借助计算机来计算本文提出的统计量的第 I 类错误率和功效, 从而对统计量的优劣进行讨论.

3.2 基于 Stationary Bootstrap 算法的检验步骤

假设$\{\frac{(\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1})^2}{\hat{\sigma}_{11}},\cdots,\frac{(\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p})^2}{\hat{\sigma}_{pp}}\}$是渐进严平稳弱相关时间序列, 其均值在零假设下都渐进为1, 则基于Stationary Bootstrap算法的检验过程如下:

(1). 计算p个U型统计量$\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1},\cdots,\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p}$的渐进方差的值$\hat{\sigma}_{11},\cdots,\hat{\sigma}_{pp}$.

(2). 计算p个学生化列统计量的值$Z_1=\frac{(\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1})^2}{\hat{\sigma}_{11}},\cdots,Z_p=\frac{(\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p})^2}{\hat{\sigma}_{pp}}$, 并计算统计量

$$W_0=\frac{p(\bar{Z}-1)^2}{\bar{D}}$$

其中 $\bar{Z}=\frac{1}{p}\sum\limits_{k=1}^{p}Z_k$, $D=\frac{1}{p-1}\sum\limits_{k=1}^{p}(Z_k-\bar{Z})^2$.

(3). 按照 Stationary Bootstrap 抽样方法的原则从 $Z_1=\frac{(\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1})^2}{\hat{\sigma}_{11}},\cdots,Z_p=\frac{(\bar{R}_{y\cdot p}-\bar{R}_{x\cdot p})^2}{\hat{\sigma}_{pp}}$ 中抽取 p 个样本, 记为 $Z^*_{(1)1},\cdots,Z^*_{(1)p}$, 并计算抽取出来的 p 个样本的均值 $\bar{Z}_{(1)}=\frac{1}{p}\sum\limits_{k=1}^{p}Z^*_{(1)k}$ 和方差 $D_{(1)}=\frac{1}{p-1}\sum\limits_{k=1}^{p}(Z^*_{(1)k}-\bar{Z}_{(1)})^2$, 接着计算 Bootstrap 统计量的值, 记为

$$W^*_{(1)}=\frac{p(\bar{Z}_{(1)}-\bar{Z})^2}{D_{(1)}}.$$

(4). 重复步骤(3) B 次, 得到 B 个 Bootstrap 统计量值, 记为 $W^*_{(1)},W^*_{(2)},\cdots,W^*_{(B)}$.

(5). 计算 P 值:

$$P=\frac{\sum\limits_{b=1}^{B}I_{\{W^*_{(B)}>W_0\}}}{B}.$$

3.3 $\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a}$ 协方差矩阵的推导

在运用 Stationary Bootstrap 算法来计算统计量的经验水平和功效时, 我们需要计算学生化列统计量的方差的渐进估计, 其具体计算过程如下:

假设 $\{x_{1a},\cdots,x_{ma},y_{1a},\cdots,y_{na}\}$ 对应的秩为 $\{R_{x1a},\cdots,R_{xma},R_{y1a},\cdots,R_{yma}\}$. 令 $\bar{R}_{x\cdot a}=\frac{1}{m}\sum\limits_{i=1}^{m}R_{xia}$,$\bar{R}_{y\cdot a}=\frac{1}{n}\sum\limits_{i=1}^{m}R_{yja}$, 则

$$\begin{aligned}\bar{R}_{y\cdot a}-\bar{R}_{x\cdot a}&=\frac{1}{n}\sum_{j=1}^{n}R_{yja}-\frac{1}{m}\sum_{i=1}^{m}R_{xia}\\&=\frac{1}{n}\sum_{j=1}^{n}\Big(\sum_{i=1}^{m}I_{\{y_{ja}>x_{ia}\}}+\sum_{l\neq j}I_{\{y_{ja}>y_{la}\}}+1\Big)-\frac{1}{m}\sum_{i=1}^{m}\Big(\sum_{j=1}^{n}I_{\{x_{ia}>y_{ja}\}}+\sum_{k\neq i}I_{\{x_{ia}>x_{ka}\}}+1\Big)\end{aligned}$$

$$
\begin{aligned}
&=\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{\{y_{ja}>x_{ia}\}}+\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{\{y_{ja}>x_{ia}\}}-n+\frac{m-1}{2}-\frac{n-1}{2}\\
&=\frac{m+n}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{1}{2}[I_{\{y_{ja}>x_{ia}\}}-I_{\{y_{ja}<x_{ia}\}}]+\frac{1}{2}\right)-\frac{m+n}{2}\\
&=\frac{m+n}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(I_{\{x_{ia}<y_{ja}\}}-I_{\{x_{ia}>y_{ja}\}}\right)
\end{aligned}
$$

其中 $I_{\{y_{ja}>x_{ia}\}}=1-I_{\{y_{ja}<x_{ia}\}}$, 则

$$
\begin{pmatrix}\bar{R}_{y\cdot 1}-\bar{R}_{x\cdot 1}\\ \vdots\\ \bar{R}_{y\cdot p}-\bar{R}_{x\cdot p}\end{pmatrix}=\frac{N}{2mn}\begin{pmatrix}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}(I_{\{x_{i1}<y_{j1}\}}-I_{\{x_{i1}>y_{j1}\}})\\ \vdots\\ \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}(I_{\{x_{ip}<y_{jp}\}}-I_{\{x_{ip}>y_{jp}\}})\end{pmatrix},
$$

其为U型统计量, 当 $min\{m,n\}\longrightarrow\infty$, 且 $0<m/n\longrightarrow\lambda_0<\infty$ 时渐进正态分布. 令 $\xi_{ab}^{(i_1i_2j_1j_2)}=cov(I_{\{x_{i_1a}<y_{j_1a}\}}-I_{\{x_{i_1a}>y_{j_1a}\}},I_{\{x_{i_2b}<y_{j_2b}\}}-I_{\{x_{i_2b}>y_{j_2b}\}})$, $i_1,i_2\in\{1,\cdots,m\}$, $j_1,j_2\in\{1,\cdots,n\}$. 我们有:

(1). 当 $i_1\neq i_2,j_1\neq j_2$ 时, $\xi_{ab}^{(i_1i_2j_1j_2)}=0$;

(2). 当 $i_1=i_2=i,j_1\neq j_2$ 时, 在零假设成立的条件下 $E[G_a(X_a)]=E[G_b(X_b)]=0.5$, 因此有:

$$
\begin{aligned}
\xi_{ab}^{(i_1i_2j_1j_2)}&=E[(I_{\{x_{ia}<y_{j_1a}\}}-I_{\{x_{ia}>y_{j_1a}\}})(I_{\{x_{ib}<y_{j_2b}\}}-I_{\{x_{ib}>y_{j_2b}\}})]\\
&=E\{E[(I_{\{x_{ia}<y_{j_1a}\}}-I_{\{x_{ia}>y_{j_1a}\}})(I_{\{x_{ib}<y_{j_2b}\}}-I_{\{x_{ib}>y_{j_2b}\}})|(x_{ia},x_{ib})]\}\\
&=E\{[1-2G_a(X_a)][1-2G_b(X_b)]\}\\
&=4cov(G_a(X_a),G_b(X_b))
\end{aligned}
$$

(3). 当 $i_1\neq i_2,j_1=j_2$ 时, 同理有 $\xi_{ab}^{(i_1i_2j_1j_2)}=4cov(F_a(Y_a),F_b(Y_b))$;

(4). 当 $i_1=i_2,j_1=j_2$ 时, $\xi_{ab}^{(i_1i_2j_1j_2)}\hat{=}\eta*I_{\{a\neq b\}}+I_{\{a=b\}}$, $|\eta|\leq 1$.

综合(1)(2)(3)(4)我们可得:

$$
cov(I_{\{x_{i_1a}<y_{j_1a}\}}-I_{\{x_{i_1a}>y_{j_1a}\}},I_{\{x_{i_2b}<y_{j_2b}\}}-I_{\{x_{i_2b}>y_{j_2b}\}})
$$

$$
= \begin{cases} 0, & i_1 \neq i_2, j_1 \neq j_2 \\ 4cov(G_a(X_a), G_b(X_b)), & i_1 = i_2, j_1 \neq j_2 \\ 4cov(F_a(Y_a), F_b(Y_b)), & i_1 \neq i_2, j_1 = j_2 \\ \eta * I_{\{a \neq b\}} + I_{\{a=b\}} & i_1 = i_2, j_1 = j_2 \end{cases}
$$

因此我们可以得到:

(1). 当 $a \neq b$ 时

$$
\begin{aligned}
& cov(\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}, \bar{R}_{y\cdot b} - \bar{R}_{x\cdot b}) \\
&= \frac{N^2}{4m^2n^2} \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} cov(I_{\{x_{i_1 a} < y_{j_1 a}\}} - I_{\{x_{i_1 a} > y_{j_1 a}\}}, I_{\{x_{i_2 b} < y_{j_2 b}\}} - I_{\{x_{i_2 b} > y_{j_2 b}\}}) \\
&= N^2 \frac{4mn(n-1)cov(G_a(X_a), G_b(X_b)) + 4nm(m-1)cov(F_a(Y_a), F_b(Y_b)) + mn\eta}{4m^2n^2} \\
&= N^2 \frac{(n-1)cov(G_a(X_a), G_b(X_b)) + (m-1)cov(F_a(Y_a), F_b(Y_b)) + \eta}{mn}
\end{aligned}
$$

(2). 当 $a = b$ 时

$$
\begin{aligned}
& cov(\bar{R}_{y\cdot a} - \bar{R}_{x\cdot a}, \bar{R}_{y\cdot b} - \bar{R}_{x\cdot b}) \\
&= N^2[(n-1)cov(G_a(X_a), G_a(X_a)) + (m-1)cov(F_a(Y_a), F_a(Y_a)) + 0.25]/(mn)
\end{aligned}
$$

对任意的 $a \in \{1, \cdots, p\}$, 我们定义 $R_y(x_{ia})$ 为 x_{ia} 在 $\{x_{ia}, y_{1a}, \cdots, y_{na}\}$ 中的秩, $R_x(x_{ia})$ 为 x_{ia} 在 $\{x_{1a}, \cdots, x_{ma}\}$ 中的秩, $R_x(y_{ja})$ 为 y_{ja} 在 $\{x_{1a}, \cdots, x_{ma}, y_{ja}\}$ 中的秩, $R_y(y_{ja})$ 为 y_{ja} 在 $\{y_{1a}, \cdots, y_{na}\}$ 中的秩. 令

$$
\begin{aligned}
\hat{\theta}_a &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \{I_{\{x_{ia} < y_{ja}\}} - I_{\{x_{ia} > y_{ja}\}}\}, \\
P_1 &= (p_{1,ia})_{m \times p}, \\
P_2 &= (p_{2,ia})_{m \times p}, \\
Q_1 &= (q_{1,ja})_{n \times p}, \\
Q_2 &= (q_{2,ja})_{n \times p},
\end{aligned}
$$

其中 $p_{1,ia} = 2R_y(x_{ia}) - 2 - n + n\hat{\theta}_a$, $p_{2,ia} = 2R_x(x_{ia}) - 1 - m$, $q_{1,ja} = 2R_x(y_{ja}) - 2 - m - m\hat{\theta}_a$, $q_{2,ja} = 2R_y(y_{ja}) - 1 - n, i \in \{1, \cdots, m\}, j \in \{1, \cdots, n\}, a \in \{1, \cdots, p\}$. 因此 $cov(G_a(X_a), G_b(X_b))$, $cov(F_a(Y_a), F_b(Y_b))$, $\hat{h_1}$, $\hat{h_2}$ 的一致估计分别为:

$$
c\hat{o}v(G_a(X_a), G_b(X_b)) = \frac{P_1^T P_1}{4mn^2}
$$

$$\hat{cov}(F_a(Y_a), F_b(Y_b)) = \frac{Q_1^T Q_1}{4m^2 n}$$

$$\hat{h_1} = \{\frac{N^2}{mn}\} \times \frac{J^T(P_1^T P_1 + Q_1^T Q_1)J}{J^T\{(P_1+P_2)^T(P_1+P_2) + (Q_1+Q_2)^T(Q_1+Q_2)\}J}$$

$$\hat{h_2} = \frac{N^2 J^T(P_1^T P_1 + Q_1^T Q_1)J}{J^T\{n^2(P_1+P_2)^T(P_1+P_2) + m^2(Q_1+Q_2)^T(Q_1+Q_2)\}J}$$

其中 $J = (1, \cdots, 1)_{1\times p}$.

第四章 统计模拟

在本章将通过 MonteCarlo 模拟, 从数值上研究 O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4)、最大秩检验统计量 (T_5) 以及本文提出的统计量 (T_{BS}) 的统计性质, 计算在不同模型和不同样本容量下六种统计量犯第 I 类错误的概率和功效. 通过比较犯第 I 类错误的概率和功效来分析这六种检验统计量的统计性质的优劣. 最后结合一个实例来对这几种统计量进行分析.

本章使用了多元正态分布模型和 Laplace 分布模型来检验统计量的统计性质. 正态分布有极其广泛的实际背景, 生产与科学实验中很多随机变量, 例如: 同一种生物体的身长、体重等指标, 测量同一物体的误差等, 都可以近似地用正态分布来描述. Laplace 分布在证券金融经济领域有着重要的作用, 在工程中对于测绘数据的处理以及在语音和图像数据等领域上也有着广泛的应用. 因此由多元正态分布模型和 Laplace 分布模型产生的数据来检验统计量的统计性质具有一定的代表性.

4.1 模拟及分析

4.1.1 模拟方法

对于多元正态分布模型, 我们从服从均值为 $\mu_1 = (\mu_{1,1}, \cdots, \mu_{1,p})^T$ 和协方差为 Σ_1 的多元正态分布中产生 m 个 p 维独立样本 $X_i = (X_{i1}, \cdots, X_{ip})^T (i \in \{1, \cdots, m\})$, 从服从均值为 $\mu_2 = (\mu_{2,1}, \cdots, \mu_{2,p})^T$ 和协方差为 Σ_2 的多元正态分布中产生 n 个 p 维独立样本 $Y_j = (Y_{j1}, \cdots, Y_{jp})^T (j \in \{1, \cdots, n\})$.

对于 Laplace 分布模型, 我们从服从均值为 $\mu_1 = (\mu_{1,1}, \cdots, \mu_{1,p})^T$ 和协方差为 Σ_1 的 Laplace 分布中产生 m 个 p 维独立样本 $X_i = (X_{i1}, \cdots, X_{ip})^T (i \in \{1, \cdots, m\})$, 从服从均值为 $\mu_2 = (\mu_{2,1}, \cdots, \mu_{2,p})^T$ 和协方差为 Σ_2 的 Laplace 分布中产生 n 个 p 维独立样本 $Y_j = (Y_{j1}, \cdots, Y_{jp})^T (j \in \{1, \cdots, n\})$.

考虑一般性, 我们采用两种协方差结构. 令对角阵 $D = (d_{q,s})_{q,s \in \{1, \cdots, p\}}$, 其对角元素 $d_{q,q} (q \in \{1, \cdots, s\})$ 产生于均匀分布 $(1, 3)$. 令 $R_1 = (\rho_1^{|q-s|})_{q,s \in \{1, \cdots, p\}}$, $R_2 = (\rho_2^{|q-s|})_{q,s \in \{1, \cdots, p\}}$, 其中 $\rho_1, \rho_2 \in \{0.2, 0.3, 0.4\}$, 两种协方差结构的设置如下:

模型 I (致密结构): $\Sigma_1 = D^{1/2} R_1 D^{1/2}$, $\Sigma_2 = D^{1/2} R_2 D^{1/2}$.

模型 II (稀疏结构): $\Sigma_1 = D^{1/2} R_1^{-1} D^{1/2}$, $\Sigma_2 = D^{1/2} R_2^{-1} D^{1/2}$.

针对第 I 类错误率, 我们将参数设置为 $\mu_1 = (\mu_{1,1}, \cdots, \mu_{1,p})^T = (0, \cdots, 0)^T$, $\mu_2 = (\mu_{2,1}, \cdots, \mu_{2,p})^T = (0, \cdots, 0)^T$. 针对检验功效, 我们令 $\mu_1 = (\mu_{1,1}, \cdots, \mu_{1,p})^T =$

$(0,\cdots,0)^T$, 而参数 $\mu_2=(\mu_{2,1},\cdots,\mu_{2,p})^T$ 设置为有 $[p^{0.9}]$ 个非零项, 这些非零项从 $\{1,\cdots,p\}$ 中随机均匀抽取. 为了更好的比较不同统计量的功效, 我们经过多次的测试, 对于不同协方差下的不同分布的非零项设置如下:

当维数 $p=100$ 不变时, 对于第一种协方差下的多元正态分布, 我们将这些非零项的值一半设置为 0.28, 另一半设置为 -0.28. 对于第二种协方差下的多元正态分布, 我们将这些非零项的值一半设置为 0.31, 另一半设置为 -0.31. 对于第一种协方差下的 Laplace 分布, 我们将这些非零项的值一半设置为 0.24, 另一半设置为 -0.24. 对于第二种协方差下的 Laplace 分布, 我们将这些非零项的值一半设置为 0.26, 另一半设置为 -0.26;

当维数 $p=100,200,300,400$ 变动时, 对于多元正态分布我们将这些非零项一半设置为 0.18, 另一半设置为 -0.18. 对于拉普拉斯分布我们将这些非零项一半设置为 0.14, 另一半设置为 -0.14.

4.1.2 模拟结果及分析

在进行 MonteCarlo 模拟时, 我们令显著性水平(也称为名义水平) $\alpha=0.05$, 循环次数 $N=2000$, 样本量 (m,n) 分别取 $(50,40),(50,60),(50,80)$ 的情形(具体代码见附录).

表 1 和表 2 分别是基于模型 I 、 II 在正态分布下六种统计量犯第 I 类错误的概率和功效, 我们将样本容量固定为 $(50,80)$, 协方差中的 $\rho_1=0.2$, $\rho_2=0.4$, 总体变量维数 p 分别取 100、200、300、400. 表 3 和表 4 分别是基于模型 I 、 II 在 Laplace 分布下六种统计量犯第 I 类错误的概率和功效, 我们将样本容量 (m,n) 固定为 $(50,60)$, 协方差中的 $\rho_1=0.2,\rho_2=0.4$, 总体变量维数 p 分别取 100、200、300、400.

从表 1 、表 2 可以看出在正态分布下 O'Brien 秩和检验统计量 (T_2)、修正的 O'Brien 秩和检验统计量 (T_3,T_4) 和本文提出的统计量 (T_{BS}) 都将第 I 类错误率控制在名义水平 (0.05) 左右, 而 O'Brien 秩和检验统计量 (T_1) 在模型 I 下第 I 类错误率相对保守, 在模型 II 第 I 类错误率又相对较为膨胀, 最大秩检验统计量 (T_5) 无论在哪种模型下都有很高的第 I 类错误率, 并且随着维数p的增加其第 I 类错误率越来越高. 例如在模型 II (表2) 下, 当 $p=400$ 时, O'Brien 秩和检验统计量 (T_2) 的第 I 类错误率为 0.058, 修正的 O'Brien 秩和检验统计量 $(T_3,\ T_4)$ 的第 I 类错误率分别为 0.0585 和 0.0565, 本文所提出的统计量 (T_{BS}) 的第 I 类错误率为 0.05, O'Brien 秩和检验统计量 (T_1) 的第 I 类错误率为 0.0685, 最大秩检验统计量 (T_5) 的第 I 类

错误率为 0.1195.

从表 3、表 4 可以看出在 Laplace 分布下 O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4) 和本文提出的统计量 (T_{BS}) 都将第 I 类错误率控制在名义水平 (0.05) 左右, 而最大秩检验统计量 (T_5) 无论在哪种模型下都有很高的第 I 类错误率. 例如在模型 I (表3) 下, 当 $p = 200$ 时, O'Brien 秩和检验统计量 (T_1, T_2) 的第 I 类错误率分别为 0.0495 和 0.051, 修正的 O'Brien 秩和检验统计量 (T_3, T_4) 的第 I 类错误率分别为 0.0515 和 0.0515, 本文所提出的统计量 (T_{BS}) 的第 I 类错误率为 0.058, 最大秩检验统计量 (T_5) 的第 I 类错误率为 0.1345.

对于检验功效, 在相同的非零参数设置下, 我们可以看出 O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4) 无论在两种分布的哪种模型下都非常低, 并且随着维数 p 的增大其功效波动幅度非常小. 最大秩检验统计量 (T_5) 和本文提出的检验统计量 (T_{BS}) 对于两种模型检验功效都较好, 并且随着维数 p 的增大其检验功效也逐渐增大, 但是本文提出的检验统计量 (T_{BS}) 检验功效的增长幅度更快, 例如在正态分布模型 I (表1) 下随着维数 p 从 100 到 400, 最大秩检验统计量 (T_5) 的检验功效从 0.316 升到 0.4275, 而本文所提出的统计量 (T_{BS}) 的检验功效从 0.3345 升到 0.9485, 因而在一定程度下我们可以认为本文提出的检验统计量 (T_{BS}) 的检验功效大大优于最大秩检验统计量 (T_5).

表 1: 基于模型 I 在正态分布下统计量的第 I 类错误率和功效

(m,n)	ρ_1	ρ_2	维数p	第 I 类错误率					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.04	0.0505	0.051	0.0505	0.1095	0.054
			200	0.034	0.045	0.0445	0.0445	0.1205	0.0575
			300	0.0395	0.051	0.0505	0.05	0.1275	0.0555
			400	0.039	0.049	0.049	0.0485	0.155	0.0565
(m,n)	ρ_1	ρ_2	维数p	功效					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.0445	0.061	0.061	0.061	0.316	0.3345
			200	0.0455	0.0515	0.051	0.051	0.3725	0.7225
			300	0.04	0.049	0.0475	0.0475	0.3915	0.8655
			400	0.0395	0.047	0.0455	0.045	0.4275	0.9485

表 2: 基于模型 II 在正态分布下统计量的第 I 类错误率和功效

(m,n)	ρ_1	ρ_2	维数p	第 I 类错误率					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.058	0.0465	0.046	0.0435	0.0935	0.049
			200	0.0615	0.0515	0.0525	0.05	0.12	0.054
			300	0.0655	0.0505	0.0525	0.0505	0.14	0.042
			400	0.0685	0.058	0.0585	0.0565	0.1195	0.05
(m,n)	ρ_1	ρ_2	维数p	功效					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.058	0.0475	0.046	0.046	0.281	0.2535
			200	0.0655	0.051	0.051	0.051	0.332	0.609
			300	0.0615	0.0495	0.05	0.0495	0.3595	0.773
			400	0.066	0.0525	0.054	0.053	0.38	0.8875

表 3: 基于模型 I 在 Laplace 分布下统计量的第 I 类错误率和功效

(m,n)	ρ_1	ρ_2	维数p	第 I 类错误率					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.0445	0.047	0.047	0.047	0.114	0.0595
			200	0.0495	0.051	0.0515	0.0515	0.1345	0.058
			300	0.051	0.056	0.0545	0.0545	0.1485	0.061
			400	0.0465	0.0505	0.0495	0.0495	0.1425	0.063
(m,n)	ρ_1	ρ_2	维数p	功效					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.0465	0.0485	0.0485	0.0485	0.2775	0.226
			200	0.0535	0.056	0.0555	0.0555	0.311	0.507
			300	0.044	0.0485	0.046	0.046	0.319	0.6535
			400	0.0465	0.053	0.0525	0.0525	0.393	0.832

表 4: 基于模型 II 在 Laplace 分布下统计量的第 I 类错误率和功效

(m,n)	ρ_1	ρ_2	维数p	第 I 类错误率					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.0505	0.047	0.044	0.0435	0.1115	0.0585
			200	0.044	0.041	0.041	0.041	0.1235	0.061
			300	0.0585	0.051	0.05	0.0485	0.1235	0.0535
			400	0.0495	0.043	0.0435	0.0435	0.1505	0.0595
(m,n)	ρ_1	ρ_2	维数p	功效					
				T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,80)	0.2	0.4	100	0.057	0.051	0.0505	0.05	0.2445	0.1605
			200	0.0515	0.043	0.0415	0.0405	0.271	0.361
			300	0.063	0.0555	0.056	0.055	0.304	0.5805
			400	0.05	0.043	0.0435	0.0425	0.3555	0.662

表 5、表 6、表 7、表 8、表 9、表 10、表 11 和表 12 分别是基于模型 I 、II 在正态分布和 Laplace 下六种统计量的第 I 类错误率和功效, 我们将变量维数 p 固定为 100, 样本容量 $(m,n) = (50,40),(50,60),(50,80)$, 协方差中的 $\rho_1 = 0.2,0.3,0.4$, $\rho_2 = 0.2,0.3,0.4$.

从表中的数据可以看出随着样本量的增大, O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4) 和本文提出的统计量 (T_{BS}) 的第 I 类错误率也在上下浮动, 但从总体上看无论是在正态分布还是在 Laplace 分布下统计量 T_1, T_2, T_3, T_4, T_{BS} 的第 I 类错误率都大致在名义水平 $\alpha = 0.05$ 左右. 最大秩检验统计量 (T_5) 无论在哪种模型下都有很高的第 I 类错误率, 随着样本间的相关性的增加, 其第 I 类错误率越来越难以控制, 虽然随着样本量的增加其第 I 类错误率有所收敛, 但总体而言依旧较为膨胀. 因此 O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4) 和本文提出的统计量 (T_{BS})都可以较好的控制第 I 类错误, 而最大秩检验统计量 (T_5) 有较高的第 I 类错误率.

对于检验功效, 在相同的非零参数设置下, 我们可以看出 O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 无论在哪种模型下都非常低, 随着样本量的增大其功效增加的幅度也非常小, 如在模型 I 的正态分布中 (表7), 当 $(m,n) = (50,40)$, $\rho_1 = 0.2$, $\rho_2 = 0.2$ 时, O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 的功效分别为 0.055、0.054、0.055、0.054, 当 $(m,n) = (50,60)$, $\rho_1 = 0.2$, $\rho_2 = 0.2$ 时, O'Brien 秩和检验统计量

表 5: 基于模型 I 在正态分布下统计量的第 I 类错误率

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.0535	0.055	0.053	0.053	0.13	0.0405
		0.3	0.055	0.0505	0.052	0.052	0.13	0.053
		0.4	0.0555	0.051	0.052	0.052	0.117	0.0525
	0.3	0.2	0.041	0.0425	0.042	0.042	0.1235	0.0515
		0.3	0.0485	0.0485	0.0485	0.0475	0.1185	0.053
		0.4	0.0495	0.0455	0.044	0.044	0.1175	0.057
	0.4	0.2	0.0505	0.0575	0.0555	0.0555	0.1185	0.0555
		0.3	0.05	0.054	0.0535	0.0535	0.112	0.057
		0.4	0.0525	0.053	0.053	0.0525	0.12	0.065
(50,60)	0.2	0.2	0.052	0.052	0.0515	0.051	0.11	0.0515
		0.3	0.0415	0.0435	0.044	0.0435	0.0955	0.049
		0.4	0.0565	0.0625	0.061	0.061	0.1055	0.049
	0.3	0.2	0.05	0.049	0.049	0.0485	0.1015	0.0575
		0.3	0.0545	0.0555	0.0545	0.0545	0.095	0.056
		0.4	0.048	0.0495	0.0485	0.0485	0.1055	0.073
	0.4	0.2	0.056	0.0535	0.0535	0.0535	0.111	0.055
		0.3	0.056	0.0535	0.0535	0.0535	0.111	0.055
		0.4	0.0535	0.051	0.0505	0.0505	0.098	0.0645
(50,80)	0.2	0.2	0.046	0.0465	0.0465	0.046	0.1095	0.056
		0.3	0.0465	0.0565	0.0565	0.0555	0.1025	0.0575
		0.4	0.04	0.0505	0.051	0.0505	0.1095	0.054
	0.3	0.2	0.0505	0.046	0.0465	0.045	0.0955	0.057
		0.3	0.0445	0.044	0.045	0.0445	0.099	0.0605
		0.4	0.052	0.059	0.059	0.0585	0.1015	0.067
	0.4	0.2	0.065	0.055	0.054	0.0525	0.097	0.065
		0.3	0.0535	0.0465	0.047	0.047	0.1055	0.059
		0.4	0.058	0.0575	0.058	0.0575	0.0895	0.068

表 6: 基于模型II在正态分布下统计量的第 I 类错误率

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.0535	0.0545	0.0515	0.0515	0.124	0.0535
		0.3	0.0465	0.0475	0.046	0.046	0.11	0.048
		0.4	0.0365	0.0415	0.0385	0.0385	0.116	0.051
	0.3	0.2	0.054	0.0505	0.049	0.0485	0.126	0.049
		0.3	0.0435	0.0435	0.041	0.0405	0.1225	0.0585
		0.4	0.0425	0.047	0.044	0.044	0.1295	0.0575
	0.4	0.2	0.054	0.048	0.048	0.0475	0.1245	0.049
		0.3	0.0475	0.0435	0.042	0.042	0.1295	0.056
		0.4	0.0515	0.0525	0.0475	0.046	0.1295	0.051
(50,60)	0.2	0.2	0.0535	0.054	0.0525	0.0525	0.0965	0.0635
		0.3	0.0445	0.0405	0.0395	0.0395	0.1055	0.049
		0.4	0.0485	0.0435	0.043	0.0425	0.102	0.061
	0.3	0.2	0.0405	0.043	0.042	0.042	0.1125	0.0425
		0.3	0.05	0.0495	0.048	0.048	0.1065	0.052
		0.4	0.057	0.054	0.0535	0.0535	0.104	0.056
	0.4	0.2	0.038	0.0445	0.0435	0.0435	0.1175	0.0445
		0.3	0.05	0.0535	0.0515	0.051	0.0935	0.048
		0.4	0.046	0.046	0.0415	0.041	0.105	0.059
(50,80)	0.2	0.2	0.0435	0.044	0.0445	0.044	0.1015	0.052
		0.3	0.062	0.0585	0.0575	0.057	0.109	0.044
		0.4	0.058	0.0465	0.046	0.0435	0.0935	0.049
	0.3	0.2	0.0375	0.041	0.04	0.04	0.094	0.053
		0.3	0.0555	0.0565	0.0555	0.055	0.1055	0.054
		0.4	0.059	0.053	0.0525	0.052	0.0935	0.062
	0.4	0.2	0.0375	0.0465	0.048	0.048	0.105	0.0555
		0.3	0.048	0.052	0.0505	0.05	0.1085	0.0555
		0.4	0.057	0.056	0.0535	0.0535	0.106	0.057

表 7: 基于模型 I 在正态分布下统计量的功效

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.055	0.054	0.055	0.054	0.5555	0.7795
		0.3	0.0565	0.0545	0.0545	0.054	0.5505	0.759
		0.4	0.055	0.051	0.052	0.0515	0.5275	0.787
	0.3	0.2	0.058	0.0605	0.06	0.06	0.573	0.81
		0.3	0.0525	0.051	0.051	0.051	0.543	0.8095
		0.4	0.0585	0.0555	0.0555	0.055	0.54	0.7715
	0.4	0.2	0.0395	0.044	0.0455	0.0455	0.5305	0.7375
		0.3	0.0425	0.044	0.044	0.0435	0.5335	0.7825
		0.4	0.05	0.0505	0.051	0.051	0.5555	0.7645
(50,60)	0.2	0.2	0.0585	0.0585	0.061	0.061	0.5845	0.904
		0.3	0.049	0.0505	0.0515	0.0515	0.604	0.9085
		0.4	0.043	0.047	0.048	0.048	0.61	0.908
	0.3	0.2	0.0635	0.0605	0.0615	0.0615	0.608	0.9165
		0.3	0.049	0.049	0.049	0.049	0.6085	0.916
		0.4	0.0525	0.056	0.056	0.056	0.616	0.935
	0.4	0.2	0.0525	0.0465	0.047	0.046	0.608	0.901
		0.3	0.059	0.057	0.057	0.057	0.618	0.919
		0.4	0.048	0.046	0.0465	0.0465	0.585	0.911
(50,80)	0.2	0.2	0.0645	0.064	0.065	0.064	0.65	0.944
		0.3	0.047	0.055	0.0555	0.0555	0.654	0.957
		0.4	0.04	0.0495	0.049	0.049	0.6645	0.9415
	0.3	0.2	0.068	0.0625	0.063	0.0625	0.664	0.9585
		0.3	0.0425	0.043	0.043	0.0425	0.6625	0.962
		0.4	0.044	0.0485	0.0495	0.049	0.665	0.962
	0.4	0.2	0.061	0.049	0.0515	0.05	0.6345	0.938
		0.3	0.0525	0.049	0.05	0.049	0.6585	0.944
		0.4	0.0545	0.0535	0.0555	0.0545	0.6585	0.942

表 8: 基于模型 II 在正态分布下统计量的功效

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.063	0.0645	0.063	0.0625	0.606	0.872
		0.3	0.045	0.048	0.0465	0.0465	0.5875	0.8525
		0.4	0.049	0.053	0.0535	0.0535	0.5215	0.752
	0.3	0.2	0.043	0.0415	0.04	0.04	0.5815	0.82
		0.3	0.0505	0.052	0.0505	0.05	0.5515	0.7675
		0.4	0.047	0.049	0.046	0.046	0.5275	0.7315
	0.4	0.2	0.062	0.0575	0.057	0.057	0.5595	0.758
		0.3	0.0545	0.052	0.049	0.049	0.488	0.7685
		0.4	0.048	0.0465	0.044	0.044	0.516	0.664
(50,60)	0.2	0.2	0.0495	0.0495	0.047	0.047	0.6495	0.9515
		0.3	0.058	0.0555	0.0555	0.055	0.6755	0.9545
		0.4	0.057	0.0535	0.053	0.053	0.619	0.9215
	0.3	0.2	0.08	0.0835	0.0805	0.0805	0.6675	0.9365
		0.3	0.081	0.08	0.0775	0.0775	0.6055	0.906
		0.4	0.075	0.0705	0.0695	0.0695	0.587	0.8835
	0.4	0.2	0.045	0.047	0.047	0.047	0.5855	0.8945
		0.3	0.0505	0.055	0.053	0.053	0.6025	0.8955
		0.4	0.0575	0.0585	0.0525	0.052	0.545	0.8285
(50,80)	0.2	0.2	0.063	0.0615	0.061	0.0605	0.7245	0.981
		0.3	0.0625	0.056	0.0545	0.054	0.713	0.964
		0.4	0.0705	0.058	0.058	0.056	0.689	0.9565
	0.3	0.2	0.044	0.0495	0.049	0.049	0.666	0.958
		0.3	0.0515	0.054	0.055	0.053	0.6695	0.967
		0.4	0.058	0.0485	0.05	0.049	0.631	0.934
	0.4	0.2	0.0335	0.0445	0.0445	0.0445	0.6515	0.9385
		0.3	0.0475	0.0535	0.0525	0.052	0.637	0.9325
		0.4	0.0475	0.0475	0.0465	0.0455	0.564	0.8855

表 9: 基于模型 I 在 Laplace 分布下统计量的第 I 类错误率

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.051	0.052	0.052	0.0515	0.119	0.0465
		0.3	0.052	0.0485	0.048	0.0475	0.123	0.0425
		0.4	0.059	0.05	0.0505	0.05	0.12	0.055
	0.3	0.2	0.0505	0.052	0.054	0.054	0.123	0.055
		0.3	0.0485	0.049	0.049	0.048	0.127	0.0595
		0.4	0.059	0.056	0.0565	0.0565	0.1325	0.067
	0.4	0.2	0.047	0.0515	0.0515	0.051	0.12	0.059
		0.3	0.042	0.044	0.0445	0.0445	0.12	0.0575
		0.4	0.0555	0.0565	0.057	0.0565	0.116	0.066
(50,60)	0.2	0.2	0.0525	0.053	0.053	0.0525	0.104	0.054
		0.3	0.0405	0.0435	0.0435	0.0435	0.1055	0.051
		0.4	0.0445	0.047	0.047	0.047	0.114	0.0595
	0.3	0.2	0.0455	0.0435	0.043	0.043	0.1025	0.0615
		0.3	0.053	0.051	0.0505	0.0505	0.118	0.0475
		0.4	0.053	0.0555	0.055	0.055	0.115	0.0585
	0.4	0.2	0.052	0.048	0.0475	0.047	0.1055	0.0625
		0.3	0.0555	0.054	0.053	0.053	0.106	0.068
		0.4	0.049	0.048	0.0475	0.047	0.1	0.068
(50,80)	0.2	0.2	0.0495	0.0495	0.0505	0.0495	0.0995	0.059
		0.3	0.043	0.0495	0.0505	0.0505	0.1055	0.056
		0.4	0.04	0.052	0.0515	0.0515	0.1025	0.059
	0.3	0.2	0.0565	0.0505	0.051	0.05	0.1005	0.054
		0.3	0.047	0.05	0.049	0.049	0.107	0.062
		0.4	0.0475	0.0535	0.053	0.0525	0.109	0.0585
	0.4	0.2	0.0645	0.05	0.052	0.049	0.093	0.057
		0.3	0.0525	0.0445	0.0445	0.044	0.101	0.0685
		0.4	0.0535	0.054	0.054	0.054	0.1015	0.064

表 10: 基于模型 II 在 Laplace 分布下统计量的第 I 类错误率

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.046	0.0455	0.0435	0.043	0.1225	0.044
		0.3	0.047	0.0505	0.048	0.048	0.1265	0.0545
		0.4	0.0455	0.0485	0.0465	0.0465	0.1295	0.053
	0.3	0.2	0.067	0.06	0.0565	0.0565	0.1155	0.0505
		0.3	0.049	0.0485	0.049	0.0485	0.1335	0.06
		0.4	0.048	0.0505	0.0485	0.0485	0.121	0.053
	0.4	0.2	0.0595	0.052	0.0495	0.0495	0.119	0.061
		0.3	0.0495	0.047	0.0475	0.047	0.1305	0.0555
		0.4	0.0525	0.052	0.047	0.047	0.127	0.056
(50,60)	0.2	0.2	0.039	0.039	0.0385	0.0385	0.098	0.052
		0.3	0.0495	0.0455	0.045	0.0445	0.1035	0.054
		0.4	0.0505	0.047	0.044	0.0435	0.1115	0.0585
	0.3	0.2	0.05	0.053	0.052	0.052	0.1105	0.0545
		0.3	0.045	0.046	0.044	0.044	0.116	0.051
		0.4	0.05	0.046	0.046	0.0455	0.1075	0.05
	0.4	0.2	0.039	0.0435	0.043	0.043	0.1	0.053
		0.3	0.05	0.053	0.05	0.05	0.121	0.0545
		0.4	0.0495	0.0495	0.045	0.045	0.1005	0.061
(50,80)	0.2	0.2	0.0485	0.05	0.05	0.049	0.104	0.046
		0.3	0.061	0.0525	0.0525	0.051	0.1075	0.0395
		0.4	0.062	0.0425	0.0425	0.042	0.106	0.0575
	0.3	0.2	0.0415	0.0495	0.048	0.0475	0.102	0.065
		0.3	0.0515	0.0515	0.048	0.0475	0.102	0.059
		0.4	0.0635	0.0495	0.0495	0.049	0.0995	0.0625
	0.4	0.2	0.037	0.051	0.051	0.051	0.124	0.051
		0.3	0.0425	0.0475	0.045	0.0445	0.106	0.0585
		0.4	0.051	0.0495	0.048	0.0465	0.0955	0.0545

表 11: 基于模型 I 在 Laplace 分布下统计量的功效

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.054	0.0525	0.0525	0.0515	0.5675	0.82
		0.3	0.0525	0.0495	0.0495	0.049	0.517	0.758
		0.4	0.0545	0.0495	0.0495	0.0495	0.563	0.7865
	0.3	0.2	0.0525	0.057	0.058	0.058	0.5545	0.7925
		0.3	0.052	0.051	0.0505	0.0505	0.52	0.7545
		0.4	0.056	0.054	0.054	0.0525	0.4985	0.7295
	0.4	0.2	0.037	0.0425	0.042	0.042	0.524	0.7755
		0.3	0.048	0.0505	0.0505	0.0505	0.523	0.763
		0.4	0.0545	0.0545	0.0545	0.0545	0.522	0.707
(50,60)	0.2	0.2	0.0555	0.055	0.0555	0.055	0.709	0.9655
		0.3	0.0475	0.0515	0.0505	0.0505	0.631	0.929
		0.4	0.0495	0.0535	0.0535	0.0535	0.665	0.9405
	0.3	0.2	0.059	0.056	0.056	0.056	0.644	0.9435
		0.3	0.0545	0.055	0.054	0.054	0.6405	0.9345
		0.4	0.0415	0.043	0.0435	0.0435	0.65	0.952
	0.4	0.2	0.0565	0.0525	0.0525	0.052	0.6195	0.9245
		0.3	0.0545	0.0515	0.0535	0.0525	0.6395	0.9325
		0.4	0.041	0.0425	0.0425	0.0425	0.6425	0.917
(50,80)	0.2	0.2	0.0515	0.0485	0.0495	0.049	0.6775	0.957
		0.3	0.053	0.06	0.0605	0.0595	0.6455	0.953
		0.4	0.0375	0.0485	0.0485	0.0485	0.6335	0.9525
	0.3	0.2	0.0485	0.046	0.047	0.0465	0.68	0.971
		0.3	0.0515	0.053	0.0535	0.053	0.609	0.9415
		0.4	0.0365	0.043	0.0445	0.0435	0.6735	0.9605
	0.4	0.2	0.069	0.051	0.052	0.051	0.6275	0.9485
		0.3	0.055	0.0475	0.0475	0.0475	0.64	0.9335
		0.4	0.0415	0.041	0.041	0.041	0.598	0.933

表 12: 基于模型II在 Laplace 分布下统计量的功效

(m,n)	ρ_1	ρ_2	T_1	T_2	T_3	T_4	T_5	T_{BS}
(50,40)	0.2	0.2	0.0595	0.059	0.0565	0.0555	0.64	0.8855
		0.3	0.0405	0.0445	0.0405	0.0405	0.556	0.8325
		0.4	0.046	0.052	0.0505	0.0505	0.5645	0.7995
	0.3	0.2	0.0705	0.0645	0.062	0.0615	0.5945	0.847
		0.3	0.047	0.045	0.0435	0.0435	0.586	0.839
		0.4	0.0445	0.052	0.0465	0.046	0.522	0.762
	0.4	0.2	0.0555	0.046	0.046	0.0455	0.563	0.8115
		0.3	0.062	0.058	0.0535	0.0525	0.5475	0.759
		0.4	0.052	0.053	0.0495	0.0485	0.499	0.6745
(50,60)	0.2	0.2	0.056	0.0545	0.0545	0.0545	0.6595	0.946
		0.3	0.059	0.0565	0.055	0.055	0.6275	0.9295
		0.4	0.0765	0.07	0.068	0.0675	0.5755	0.8695
	0.3	0.2	0.048	0.0495	0.0495	0.049	0.618	0.9375
		0.3	0.0575	0.057	0.0535	0.0535	0.603	0.894
		0.4	0.056	0.0525	0.0485	0.047	0.528	0.8335
	0.4	0.2	0.0445	0.049	0.048	0.048	0.568	0.864
		0.3	0.052	0.0545	0.0505	0.0505	0.5375	0.8385
		0.4	0.049	0.0485	0.0485	0.0485	0.5015	0.7515
(50,80)	0.2	0.2	0.0655	0.064	0.064	0.0635	0.721	0.984
		0.3	0.0695	0.059	0.0585	0.0575	0.6955	0.973
		0.4	0.074	0.058	0.0565	0.055	0.665	0.964
	0.3	0.2	0.058	0.063	0.064	0.0625	0.7215	0.9705
		0.3	0.063	0.065	0.0615	0.0615	0.654	0.9535
		0.4	0.0515	0.047	0.0445	0.0445	0.626	0.951
	0.4	0.2	0.05	0.071	0.0665	0.0665	0.632	0.9255
		0.3	0.051	0.0575	0.056	0.0555	0.5995	0.8995
		0.4	0.047	0.0475	0.045	0.0445	0.5815	0.858

(T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 的功效分别为 0.0585、0.0585、0.61、0.061. 随着样本之间的相关性变大, 这四种统计量的功效反而有下降的趋势. 如在模型 II 的正态分布中 (表8), 当 $(m, n) = (50, 40)$, $\rho_1 = 0.2$, $\rho_2 = 0.2$ 时, O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 的功效分别为 0.063、0.0645、0.063、0.0625, 当 $(m, n) = (50, 40)$, $\rho_1 = 0.2$, $\rho_2 = 0.4$ 时, 功效分别为 0.049、0.053、0.0535、0.0535. 最大秩检验统计量 (T_5) 的检验功效与 O'Brien 秩和检验统计量 (T_1, T_2) 和修正的O'Brien 秩和检验统计量 (T_3, T_4) 相比表现得较好, 大多都在区间 (0.5, 0.7). 而本文提出的检验统计量 (T_{BS}) 对于两种模型检验功效都非常好, 某些情况甚至高达 0.962, 更加优越于最大秩检验统计量 (T_5). 总体来讲本文提出的检验统计量 (T_{BS}) 的检验功效相对于其他的统计量的检验功效是遥遥领先的.

综合而言, 本文提出的检验统计量 (T_{BS}) 具有良好的统计性质. 对于第 I 类错误率, 与最大秩检验统计量 (T_5) 相比它的第一类错误率控制得更好, 与 O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 相差也不大, 在某些情况下甚至表现得更好. 对于检验功效, 无论在哪种模型下本文提出的检验统计量 (T_{BS}) 都是遥遥领先, 而 O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 的检验功效都非常低.

4.2 实例分析

随着生物学技术的发展, 生物学家已经进行了全基因组研究以研究芽孢酵母中单倍体不足的生长表型. Ohnuki et al.(2018)[7] 使用具有 501 种形态特征的高维表型研究了杂合1,112个必需基因的菌株的单倍不足, 还提出了 114 种具有 501 个形态特征的野生型复制品. 该数据集为研究酿酒酵母中的基本基因功能提供了有用的资源. 我们将 O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4)、最大秩检验统计量 (T_5) 以及本文提出的统计量 (T_{BS}) 应用于 Ohnuki et al.(2018)[7] 所研究的数据集进行分析, 研究来自酵母必需基因的杂合子和野生型的 114 个复制品的所有 501 个形态特征之间是否存在差异. 该数据集可在 http://www.yeast.ib k.u-tokyo.ac.jp/SCMD/summary. php? pj = ess1112het 上获得.

首先, 我们测试了 1,112 种必需基因的酵母杂合子和 114 种野生型复制品的 501 种形态特征之间是否存在差异. O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4) 的 p 值分别为 0.06598726、0.05921012、0.05672050、0.05846328. 最大秩检验统计量 (T_5) 以及本文提出的统计量 (T_{BS})的

p 值都远小于 0.05. 由 p 值可以看出最大秩检验统计量 (T_5) 以及本文提出的统计量 (T_{BS})可以明显识别出差异. 其次, 我们将酵母必需 1,112 个基因的杂合子分为两组, 前 500 个基因被分为第 1 组, 后 612 个基因被分为第 2 组, 我们测试两组中所有 501 个形态特征之间是否存在差异. O'Brien 秩和检验统计量 (T_1, T_2)、修正的 O'Brien 秩和检验统计量 (T_3, T_4)、最大秩检验统计量 (T_5) 的 p 值分别为 0.05954172、0.05961865、0.06027383、0.06027928、0.02476340. 本文提出的统计量 (T_{BS})的 p 值 远小于 0.02. 由 p 值可以看出最大秩检验统计量 (T_5) 和本文提出的统计量 (T_{BS}) 可以显著识别出两组间的差异 (检验水平以 0.05 为标准), 而其他的统计量不能很好的识别出两组间的差异, 因此本文提出的统计量 (T_{BS}) 的检验较为敏锐.

第五章　总结与展望

在本文中我们提出了一种新的检验统计量来对高维数据下的非参数 Behrens-Fisher 问题进行假设检验. 由于在高维情况下要推导出新的统计量较为精确的均值估计与协方差估计较为困难, 因此参考Li et al.(2020)[23] 中将 Stationary Bootstrap 方法运用到高维数据双样本均值检验的原理, 本文运用 Stationary Bootstrap 算法求新的统计量的经验水平(也称 p 值)与O'Brien 秩和检验统计量、修正的O'Brien 秩和检验统计量和最大秩检验统计量进行比较, 从而讨论其统计性质.

通过大量的数据模拟, 我们从统计量的经验水平和经验功效两方面的综合表现来看, 本文提出的统计量要优于其他统计量:

①. 从经验水平来看: 最大秩检验统计量 (T_5) 无论在什么情况下其第 I 类错误率都难以控制, 并且随着样本间的相关性的增加其第 I 类错误率越来越膨胀, 随着维数p 的增加其第 I 类错误率也越来越高, 虽然随着样本量的增加其第 I 类错误有所收敛但是相对来讲依然很高. O'Brien 秩和检验统计量 (T_1) 可能因为协方差结构的不同第 I 类错误率的表现也不同, 在正态分布的模型 I 下, 随着维数 p 增加其第一类错误率相对较为保守, 而在正态分布的模型 II 下, 随着维数 p 增加其第 I 类错误率又相对较为膨胀. 从总体数据来看, 本文提出的检验统计量 (T_{BS}) 与 O'Brien 秩和检验统计量 (T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 都能较好的控制第 I 类错误.

②. 从检验功效来看: 对于相同的参数设置, O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 的检验功效都表现得非常差. 最大秩检验统计量 (T_5) 的功效与 O'Brien 秩和检验统计量 (T_1, T_2)、修正的O'Brien 秩和检验统计量 (T_3, T_4) 相比表现得较好, 但本文提出的统计量的检验功效明显地高于最大秩检验统计量 (T_5), 甚至在有些情况高达 0.9 以上.

综上而言, 本文提出的统计量对检验高维数据的非参数 Behrens-Fisher 问题具有一定的可行性. 运用 Stationary Bootstrap 方法求新的统计量的 p 值可以避免估计复杂的均值和方差. 虽然 Stationary Bootstrap 方法要通过大量的重抽样, 但是我们可以借助计算机来完成. 在大量的数据模拟下, 本文提出的统计量具有合理的经验水平和理想的功效, 因此可以运用本文提出的统计量来对高维数据的非参数 Behrens-Fisher 问题进行检验.

除了本文研究的工作以外, 对高维数据的非参数 Behrens-Fisher 问题的假设检验的研究还有进一步探讨的空间, 具体有如下几个方面:

①. 在本文中我们是基于数据是严平稳弱相关的, 对于其他情况的数据还有进

一步深入研究的价值.

②. 本文只对于服从多元正态分布和 Laplace 分布的高维数据进行模拟实验, 没有考虑其他类型的分布. 本文提出的统计量对于其他类型的分布是否适用仍需要进一步讨论.

③. 本文研究的是两个总体的分布相同的情况, 对于两个总体的分布不同或者某个总体的某些元素遵循不同的分布的情况有待进一步讨论.

参考文献

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

[2] Mann H B, Whitney D R. On a test of whether one of two random variables is stochastically larger than the other[J]. The annals of mathematical statistics, 1947: 50-60.

[3] O'Brien PC. Procedures for comparing samples with multiple endpoints[J]. Biometrics, 1984, 40: 1079-1087.

[4] Sankoh A J, Huque M F, Russell H K, et al. Global two-group multiple endpoint adjustment methods applied to clinical trials[J]. Drug information journal: DIJ/Drug Information Association, 1999, 33(1): 119-140.

[5] Kaufman K D, Olsen E A, Whiting D, et al. Finasteride in the treatment of men with androgenetic alopecia[J]. Journal of the American Academy of Dermatology, 1998, 39(4): 578-589.

[6] Shames R S, Heilbron D C, Janson S L, et al. Clinical differences among women with and without self-reported perimenstrual asthma[J]. Annals of Allergy, Asthma & Immunology, 1998, 81(1): 65-72.

[7] Li D K B, Zhao G J, Paty D W. Randomized controlled trial of interferon-beta-1a in secondary progressive MS: MRI results[J]. Neurology, 2001, 56(11): 1505-1513.

[8] Tilley B C, Pillemer S R, Heyse S P, et al. Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trials[J]. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 1999, 42(9): 1879-1888.

[9] Brunner E, Munzel U, Puri M L. The multivariate nonparametric Behrens-Fisher problem[J]. Journal of Statistical Planning and Inference, 2002, 108(1-2): 37-53.

[10] Huang P, Tilley BC, Woolson RF, Lipsitz S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem[J]. Biometrics, 2005, 61: 532-539.

[11] Liu A, Li Q, Liu C, et al. A rank-based test for comparison of multidimensional outcomes[J]. Journal of the American Statistical Association, 2010, 105(490): 578-587.

[12] 杨贵军,林珍英,张聪聪.关于一般非参数Behrens-Fisher 问题的秩和检验方法[J].统计与信息论坛,2015, 30(01): 3-8.

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

[14] Mooney C Z, Mooney C F, Mooney C L, et al. Bootstrapping: A nonparametric approach to statistical inference[M]. sage, 1993.

[15] Hall P. An Edgeworth View of the Bootstrap[M]. The Bootstrap and Edgeworth Expansion. Springer, New York, NY, 1992: 83-153.

[16] Shao J, Tu D. The jackknife and bootstrap[M]. Springer Science & Business Media, 2012.

[17] Davison A C, Hinkley D V. Bootstrap methods and their application[M]. Cambridge university press, 1997.

[18] Politis D N, Romano J P. The stationary bootstrap[J]. Journal of the American Statistical association, 1994, 89(428): 1303-1313.

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

[20] Chen S X, Qin Y L. A two-sample test for high-dimensional data with applications to gene-set testing[J]. The Annals of Statistics, 2010, 38(2): 808-835.

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

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

[23] Li Z, Liu F, Zeng L, et al. A stationary bootstrap test about two mean vectors comparison with somewhat dense differences and fewer sample size than dimension[J]. Computational Statistics, 2020: 1-20.

[24] Bilodeau M, Brenner D. Theory of multivariate statistics[M]. Springer, 1999.

[25] Ohnuki S, Ohya Y. High-dimensional single-cell phenotyping reveals extensive haploinsufficiency[J]. PLoS biology, 2018, 16(5): e2005130.

附录

R模拟程序代码

```
install.packages("np")
library(np)
install.packages("mvtnorm")
library(mvtnorm)
rm(list=ls())
## calculate rank
CAL.RANK <- function(X,Y)
{
     k <- ncol(X)
     m <- nrow(X)
     n <- nrow(Y)
     dx <- 1:m
     dy <- (m+1):(m+n)
     Rx <- matrix(0,nrow=m,ncol=k)
     Ry <- matrix(0,nrow=n,ncol=k)
     RR <- matrix(0,nrow=k,ncol=1)
     for (a in 1:k)
     {
         temp <- rank(c(X[,a],Y[,a]))
         Rx[, a] <- temp[dx]
         Ry[, a] <- temp[dy]
         RR[a,1] <- mean(temp[dy]) - mean(temp[dx])
         rm(temp)
     }
     #print(cbind(Rx,Ry))
     #print(RR)
     list(Rx,Ry,RR)
}
```

```
## calcualte hs
CAL.HS <- function(X,Y,Rx,Ry)
{
    k <- ncol(X)
    m <- nrow(X)
    n <- nrow(Y)

    # theta
    theta <- rep(0,k)
    for (a in 1:k)
    {
        for (i in 1:m)
        {
            temp <- sum(X[i,a]<Y[,a])
            theta[a] <- theta[a] + temp - (n-temp)
            rm(temp)
        }
    }
    theta <- theta/(m*n)

    A1 <- matrix(0,nrow=m,ncol=k)
    A2 <- matrix(0,nrow=m,ncol=k)
    B1 <- matrix(0,nrow=n,ncol=k)
    B2 <- matrix(0,nrow=n,ncol=k)
    for (a in 1:k)
    {
        req <- rank(X[,a])
        for (i in 1:m)
        {
            temp <- rank(c(X[i,a],Y[,a]))[1]
            A1[i,a] <- 2*temp - 2 - n + n*theta[a]
            A2[i,a] <- 2*req[i]- 1 - m
            rm(temp)
        }
```

```
    rm(req)

    req <- rank(Y[,a])
    for (j in 1:n)
    {
        temp <- rank(c(Y[j,a],X[,a]))[1]
        B1[j,a] <- 2*temp - 2 - m - m*theta[a]
        B2[j,a] <- 2*req[j] - 1 - n
        rm(temp)
    }
    rm(req)
}

# caluclate parameter
M1 <- t(A1) %*% A1 + t(B1) %*% B1
M2 <- t(A1+A2) %*% (A1+A2) + t(B1+B2) %*% (B1+B2)
M3 <- n^2 * t(A1+A2) %*% (A1+A2) + m^2 * t(B1+B2) %*% (B1+B2)

# Huang et al.
J <- matrix(rep(1,k),ncol=1)
temp1 <- t(J) %*% M1 %*% J
temp2 <- t(J) %*% M2 %*% J
temp3 <- t(J) %*% M3 %*% J
h1 <- (m+n)^2/(m*n) * temp1 / temp2
h2 <- (m+n)^2 * temp1 / temp3

# Max
G.mat <- t(A1) %*% A1 / (4*m*n^2)
F.mat <- t(B1) %*% B1 / (4*m^2*n)
V <- matrix(0,nrow=k,ncol=k)
for (a in 1:k)
{
    for (b in 1:k)
    {
```

```
            V[a,b] <- (m+n)^2*((n-1)*G.mat[a,b]+(m-1)*F.mat[a,b])/(m*n)
        }
        V[a,a] <- (m+n)^2*((n-1)*G.mat[a,a]+(m-1)*F.mat[a,a]+0.25)/(m*n)
    }
    #print(V)

    corrcoef <- matrix(0,nrow=k,ncol=k)
    for (a in 1:k)
    {
        for (b in 1:k)
        {
            corrcoef[a,b] <- V[a,b]/sqrt(V[a,a]*V[b,b])
        }
    }
    list(h1, h2, V, corrcoef, F.mat, G.mat)
}

###sampling SBboot sample
SBoot = function(x,b){
X = rep(x,times=2)
n = length(x)
J = rgeom(n,1/b)+1  #rgeom返回每组试验经历失败的次数
tt=(which(cumsum(J)>=n))[1] #取累加大于等于n时的第一个数
A = sample(1:n,tt,replace=TRUE) #从1:n中任意有重复抽tt个数，代表的是每
个块的第一个元素
Block = c()
for(j in 1:tt)
{
Block = c(Block,A[j]:(A[j]+J[j]-1))
 }
N = Block[1:n]
return(X[N])
}
######################sampling SBboot sample
```

```
CAL.P.VALUES <- function(X, Y)
{
    k <- ncol(X)
    m <- nrow(X)
    n <- nrow(Y)
    temp <- CAL.RANK(X,Y)
    Rx <- temp[[1]]
    Ry <- temp[[2]]
    RR <- temp[[3]]

    # OBrien's method
    RxS <- apply(Rx,1,sum)
    RyS <- apply(Ry,1,sum)
    x.bar <- mean(RxS)
    x.sig <- var(RxS)
    y.bar <- mean(RyS)
    y.sig <- var(RyS)
    sig <- ((m-1)*x.sig+(n-1)*y.sig)/(m+n-2)
    T1 <- (y.bar-x.bar)/sqrt(sig*(1/m+1/n))
    T1 <- abs(T1)
    p1 <- pt(-T1,(m+n-2))*2

    eta <- (x.sig/m)/(x.sig/m+y.sig/n)
    dff <- (eta^2/(m-1)+(1-eta)^2/(n-1))^{-1}
    T2 <- (y.bar-x.bar)/sqrt(x.sig/m+y.sig/n)
    T2 <- abs(T2)
    p2 <- pt(-T2,dff)*2

    # variance matrix
    qet <- CAL.HS(X,Y,Rx,Ry)
    FM = qet[[5]]
    GM = qet[[6]]
```

```
# Huang et al's method
h1 <- qet[[1]]
T3 <- T1/sqrt(h1)
p3 <- pt(-T3,(m+n-2))*2
h2 <- qet[[2]]
T4 <- T2/sqrt(h2)
p4 <- pt(-T4,dff)*2

# MAX
V <- qet[[3]]
obs <- max(abs(RR)/sqrt(diag(V)))
cor.mat <- qet[[4]]
p5 <- 1-pmvnorm(lower=rep(-obs,k), upper=rep(obs,k), corr=cor.mat)

#Boostrap
X0 = RR^2 /diag(V)
T0 = as.numeric(k*(mean(X0) - 1)^2/var(X0))
b = b.star(X0,round = TRUE)[1]
############
stf = function(y)
{
( k*(mean(y) - mean(X0))^2/var(y) )
}
############
az = c()
for(i in 1:3000)
{
boot1 = SBoot(X0,b)
T6 = stf(boot1)
az=c(az,T6)
}
pvMSs = mean( az > T0 )
p6 = pvMSs
list(RR,FM,GM,c(p1, p2, p3, p4, p5, p6))
```

```
}

###########################
covMed=function(rho,p)
{
    M=matrix(0,p,p)
    for(i in 1:p)
     {
    for(j in 1:p)
      {M[i,j]=rho^{abs(i-j)}}
     }
     return(M)
}

num.simu = 2000
p =  100
alpha = 0.05
N=c(40,60,80)
for(j in 1:3)
{
n1 = 50
n2 = N[j]
DD = diag(sqrt(runif(p,1,3)))
for(rho1 in seq(0.2,0.4,by=0.1))
{
names(rho1) =  "rho1"
print(rho1)
VV1 = DD %*% covMed(rho1,p) %*% DD
for (rho2 in seq(0.2,0.4,by=0.1))
{
names(rho2) =  "rho2"
print(rho2)
VV2 =  DD %*% covMed(rho2,p) %*% DD
mu1 = rep(0,p)
```

```
mu2 = rep(0,p)
pvs = c()
print(date())
for(i in 1:num.simu)
{
if(i %% 1000 ==0) print(i)
X = rmvnorm(n1, mu1, VV1)
Y = rmvnorm(n2, mu2, VV2)

cttj = CAL.P.VALUES(X=X, Y=Y)
pvalue=round(cttj[[4]],4)
pvs = rbind(pvs,pvalue)
}
sizes = colMeans(pvs <= alpha)
Para = c(p,n1,n2,alpha,rho1,rho2,num.simu)
A1 = c(Para,sizes)
names(A1) = c('p','n1','n2','alpha','rho1', 'rho2', 'num.simu', 'p1',
'p2', 'p3', 'p4', 'p5', 'p6')
print(A1)
write.table(t(as.matrix(A1)),file="p values of mvrnorm model 1.txt",
sep=",",row.names=F, col.names=T,append=T)
}
}
print(date())
}
```

致　谢

光阴荏苒, 研究生三年的学习之旅即将结束, 三年的学习生活使我受益匪浅. 回首三年在华师的学习生活, 我得到了许多的关怀和帮助, 现在向他们表达我最诚挚的谢意.

首先, 我要深深感谢我的导师李正帮老师. 李老师为人谦和, 平易近人, 对待我们提出的问题总是很有耐心的解释. 在论文的选题、方法使用和写作阶段, 李老师都倾注了极大的关怀和鼓励. 在编程模拟过程中, 每当我有所疑问, 李老师总会放下繁忙的工作, 不厌其烦地指点我; 在我论文写作遇到瓶颈时, 李老师又在百忙之中抽出空来对我的论文写作结构、方向和侧重点进行指导, 提出许多中肯的修改意见, 使我在研究和写作过程中不致迷失方向. 他严谨的治学之风和对事业的孜孜追求将影响和激励我的一生, 他对我的关心和教诲我更将永远铭记. 借此机会, 我谨向李老师致以深深地谢意.

其次我要感谢的是数统学院行政部门和任课的老师们, 谢谢李波老师、胡建伟老师、宁建辉老师在讲台上辛勤的讲授, 谢谢丁老师和舒老师在我担任学生助理期间对我工作的指导, 使我受益匪浅; 谢谢辅导员李东老师和张岚老师对我生活上的关心与帮助, 谢谢教学秘书熊老师对研究生工作的付出, 是你们让我感觉能在数统学院学习是一件非常幸福的事.

还要感谢的是我的室友陈静、杨兰萍和周子西, 三年来, 我们朝夕相处, 共同进步. 正是有你们的陪伴才会让我在困惑与迷茫的日子里继续勇往直前, 才会让我的生活到处充满欢声笑语, 才会让我在这个陌生的城市体会到家的温暖, 感谢你们给予我的所有关心和帮助, 谢谢你们对我的支持、包容与理解, 同窗之谊, 我将终生难忘!

最后, 感谢我的家人在我研究生期间给予我的包容、关爱和鼓励, 以及所有陪我一路走来的同学和朋友, 正是由于他们的支持和照顾, 我才能安心学习, 并顺利完成我的学业.

毕业在即, 在今后的工作和生活中, 我会铭记师长们的教诲, 继续不懈努力和追求, 来报答所有支持和帮助过我的人!

贾婉茹

2021 年 3 月