分类编号：_____ 单位代码：10065

密　　级：_____ 学　　号：1810050011

天津师范大学

研究生学位论文

论文题目：高维数据下正态总体的假设检验问题

学 生 姓 名：李 熠 璇　　申请学位级别：硕 士

申请专业名称：概率论与数理统计

研 究 方 向：多元统计

指导教师姓名：范 永 辉　　专业技术职称：副教授

提交论文日期：2021 年 3 月

天津师范大学学位论文原创声明

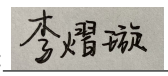
本人郑重声明: 此处所提交的学位论文《高维数据下正态总体的假设检验问题》，是本人在导师指导下，在天津师范大学攻读硕士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

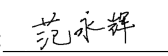
作者签名:  日期: 2021 年 5 月 30 日

天津师范大学学位论文使用授权书

《高维数据下正态总体的假设检验问题》系本人在天津师范大学攻读学位期间在导师指导下完成的学位论文。本论文的研究成果归天津师范大学所有，本论文的研究内容不得以其他单位的名义发表。本人完全了解天津师范大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅，同意学校将论文加入《中国优秀博硕士学位论文全文数据库》和编入《中国知识资源总库》。本人授权天津师范大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

本学位论文属于（请在以下相应方框内打" √");
保密□, 在_____年解密后适用本授权书
不保密□

作者签名:  日期: 2021 年 5 月 30 日

导师签名:  日期: 2021 年 5 月 30 日

摘　要

随着计算机科学技术迅猛发展, 人类进入了大数据时代, 数据收集越来越容易, 与此同时数据库的规模越来越大, 复杂性也越来越高, 呈现高维数据. 高维数据一个重要特点就是维数超过了样本容量, 此类数据在生物统计、气象统计、农业、金融等领域广泛存在. 在多元统计分析中, 许多经典的检验方法在高维数据背景下失效, 为解决这一问题, 本文对高维数据下有关正态总体的一系列假设检验问题进行了相关研究.

第一章首先介绍了高维数据下检验问题的相关研究发展历程, 并对下文要用到的相关定理及预备知识进行了简单介绍.

第二章运用Bonferroni校正思想对高维数据下正态分布均值的假设检验问题进行研究. 提出了一种按维数划分样本阵的新方法, 并结合经典的Hotelling T^2检验方法, 分别对单样本情形和双样本情形进行讨论, 并得到了新检验统计量, 通过数值模拟验证新检验具有优良性.

第三章在似然比的基础上结合并交原则对高维数据下的方差分析问题进行研究. 给出了该检验的广义似然比统计量及其相关性质, 并用蒙特卡洛方法模拟出检验水平, 结果显示该检验能很好的控制检验水平.

第四章在似然比的基础上结合并交原则对高维数据下正态均值线性约束的假设检验问题进行研究. 给出了该检验问题的广义似然比统计量及其相关性质, 并通过数值模拟方法验证该检验与Bai和Saranadasa的检验方法相比能更好的控制检验水平.

第五章在似然比的基础上结合并交原则对高维数据下随机变量的独立性检验问题进行研究. 重点讨论了高维数据下两组变量的独立性问题, 并给出了新检验统计量的定义、定理及检验方法.

关键词: 高维数据; Bonferroni校正; 并交原则(UIT); 广义似然比检验; 蒙特卡洛方法

Abstract

With the rapid development of computer science and technology, human beings have entered the era of big data. Data collection is becoming easier and easier. At the same time, the scale and complexity of database are becoming larger and larger, presenting high-dimensional data, This kind of data widely exists in the fields of Biostatistics, meteorological statistics, agriculture, finance and so on. In multivariate statistical analysis, many classical test methods fail under the background of high-dimensional data. In order to solve this problem, this paper studies a series of hypothesis testing problems about normal population under high-dimensional data

In the first chapter, we introduce the related research and development process of the test problem under high-dimensional data, and briefly introduce the related theorems and preparatory knowledge to be used in the following.

In the second chapter, we use Bonferroni correction to study the hypothesis test of the mean value of normal distribution under high-dimensional data. We propose a new method to divide the sample matrix according to the dimension. Combined with the classical Hotelling T^2test method, we discuss the single sample case and the double sample case respectively, and obtain the new test statistics, Numerical simulation shows that the test is good.

In the third chapter, based on the likelihood ratio, combining with the union-intersection principle, we study the ANOVA under high-dimensional data. We give the generalized likelihood ratio statistics of the test and its related properties, and simulate the test level with Monte Carlo method. The results show that the test can control the test level well.

In the fourth chapter, based on the likelihood ratio and the principle of intersection, we study the hypothesis testing problem of linear constraint of normal mean under high-dimensional data. We give a new definition of the generalized likelihood ratio statistics and its related properties. The data simulation method is used to verify that the test can better control the test level than Bai and saranadasa's test method.

In the fifth chapter, based on the likelihood ratio and the principle of intersection, we study the independence test of random variables in high-dimensional data. We focus on the independence test of two groups of variables in high-dimensional data, and propose the definition, theorem and test method.

Key words: High-dimensional data; Bonferroni correction ; union-intersection principle(UIT); generalized likelihood ratio test; Monte Carlo method

目　录

符 号 说 明

A'	矩阵A的转置
A^{-1}	矩阵A的逆
A^{+}	矩阵的Moore-Penrose广义逆
$A>0$	A为对称正定矩阵
$A\geq 0$	A为对称半正定矩阵
$\|A\|$	矩阵A的行列式
$\mathrm{tr}(A)$	矩阵A的迹
$\mathrm{rk}(A)$	矩阵A的秩
$\lambda_1(A)$	矩阵A的最大特征值
$\Re(A)$	矩阵A的列向量生成的线性子空间
$\Re^{\perp}(A)$	$\Re(A)$的正交补空间
P_A	$\Re(A)$上的正交投影阵
$\mathrm{Vec}(A)$	矩阵A的列向量依次相接排成的列矩阵
1_n	分量全为1的维列向量, 即$1_n=(1,1,\cdots,1)'$
H_n	$H_n=I_n-\frac{1}{n}1_n1_n'$
$\|a\|$	向量a的长度
χ_p^2	自由度为p的中心χ^2分布
$t(n)$	自由度为n的中心t分布
z_α	标准正态分布的上侧α分位点
$x\sim N_p(\mu,\Sigma)$	x服从均值为μ、协方差为Σ的p维正态分布
$W_p(n,\Sigma)$	参数为p,n,Σ的中心Wishart分布
$F(m,n),F_\alpha(m,n)$	自由度为m,n的中心F分布及其上侧α分位点
$T^2(p,n)$	参数p,n的中心T^2分布
$Beta(a,b)$	参数为a,b的$Beta$分布

第 1 章　绪论

1.1 选题背景及意义

随着科学技术的迅猛发展, 人类进入了大数据时代, 高维数据随之产生. Tian[1]对高维数据有如下定义: 高维数据是指具有大量变量且变量远大于观测值的数据, 即高维数据的一个特点就是维数远大于样本容量, 也就是少量的数据样本含有大量的属性. 这类数据在生物统计、气象统计、农业、经济金融、医疗等各个领域广泛存在, 例如包含多种信息的个人档案、多种指标的环境监测数据、DNA和基因的数据表达, 以及Hall[2]中提到的测量基因微阵列、医学成像分析和化学计量学中各种类型的光谱测量等, 因此对此类数据进行分析处理显得至关重要.

对高维数据进行分析是近几十年来最具困难和挑战的研究热点之一, 这是由于相对于传统数据来讲, 高维数据包含了更多更全面的信息, 维数的不断增加导致数据的复杂性越来越高. 传统数据即样本容量大于维数的数据, 很多经典的检验理论均在传统数据的条件下提出, 但却不能适用于高维数据, 例如经典的Hotelling T^2检验[3,4], 在正态分布时该检验对传统数据具有良好的检验效果, 但对于高维数据该检验方法失效, 这是由于此时样本协方差矩阵的逆是不存在的, 从而导致Hotelling T^2检验没有意义.

为弥补这一缺陷, 早期学者对高维数据提出了各种降维方法, 例如主成分分析, 但该方法不能适用于维数过高的情况, 这是因为随着维数的不断增大, 信息丢失的问题越来越严重使得该方法受限. 近年来, 对于如何在高维数据下运用主成分分析的方法, 许多学者也进行了相关研究, 见文献 [5–7]; 此外研究学者对于高维数据下的检验问题提出了另一种重要的思路－对经典的似然比检验方法进行修正, 似然比检验是数理统计中一种重要且有效的假设检验方法, 对似然比检验进行修正, 使其解决高维数据背景下的检验问题一直以来都是学者研究和分析的热点.

1.2 多元统计分析中关于正态分布的假设检验问题

统计假设检验是多元统计推断的基本问题之一, 正态总体均值和总体协方差检验是假设检验中的重要问题, 下面介绍一些关于多元正态总体的假设检验问题:

单样本问题

假设$X_1, X_2, \cdots, X_n$为来自p维正态分布$N_p(\mu, \Sigma)$的简单随机样本, 其中μ和Σ未知且$\Sigma > 0$, 考虑如下假设检验问题:

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0. \tag{1.1}$$

不失一般性地, 经过相应的变换, 可设$\mu_0 = \mathbf{0}$, 即

$$H_0 : \mu = \mathbf{0} \leftrightarrow H_1 : \mu \neq \mathbf{0}. \tag{1.2}$$

两样本问题

假设$X_1, X_2, \cdots, X_{n_1}$为来自$p$维正态分布$N_p(\mu_1, \Sigma_1)$的简单随机样本, $Y_1, Y_2, \cdots, Y_{n_2}$为来自$p$维正态分布$N_p(\mu_2, \Sigma_2)$的简单随机样本, 其中$\mu_1$、$\mu_2$和$\Sigma_1$、$\Sigma_2$均未知, $\Sigma_1 > 0, \Sigma_2 > 0$, 并且两样本相互独立, 考虑如下假设检验问题:

$$H_0 : \mu_1 = \mu_2 \leftrightarrow H_1 : \mu_1 \neq \mu_2. \tag{1.3}$$

方差分析问题

假设有$k(k \geq 2)$个相互独立的p维总体$\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_k$且$\mathrm{X}_i \sim N_p(\mu_i, \Sigma)$, 其中$\mu_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{ip})' \in R^p, \Sigma > 0$, $X_{i1}, X_{i2}, \cdots, X_{in_i}$是来自于总体$\mathrm{X}_i$的简单随机样本, $i = 1, 2, \cdots, k, n = \sum_{i=1}^{k} n_i$. 考虑如下假设检验问题:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \leftrightarrow H_1 : \mu_1, \mu_2, \cdots, \mu_k\text{不全相等}. \tag{1.4}$$

均值受约束问题

假设$X_1, X_2, \cdots, X_n$为来自p维正态分布$N_p(\mu, \Sigma)$的简单随机样本, 其中μ和Σ未知且$\Sigma > 0$. 考虑如下假设检验问题：

$$H_0 : R\mu = r \leftrightarrow H_1 : R\mu \neq r, \tag{1.5}$$

其中R为$q \times p$的行满秩矩阵, r为$q \times 1$向量.

变量的独立性检验问题

假设$X_1, X_2, \cdots, X_n$是来自p维正态总体$N_p(\mu, \Sigma)$的简单随机样本, 为方便起见, 我们记样本阵为X, 将X按维数分为q组, 其中$q \leq p$, 则样本阵X可表示为

$$X = (X_1, X_2, \cdots, X_n) = (X^{(1)}, X^{(2)}, \cdots, X^{(q)})',$$

相应的协方差矩阵可表示为：

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1q} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{q1} & \Sigma_{q2} & \cdots & \Sigma_{qq} \end{bmatrix},$$

Σ_{ij}是$p_i \times p_j$矩阵, 且$1 \leq i, j \leq q, \sum_{i=1}^{q} p_i = p$. 讨论$q$个分量中任意两个分量之间的独立性, 即考虑如下假设检验问题:

$$H_0 : \Sigma_{ij} = 0 \leftrightarrow H_1 : \Sigma_{ij} \neq 0, (i \neq j). \tag{1.6}$$

1.3 研究现状

1.3.1 Hotelling T^2检验

Hotelling T^2检验是多元统计分析中一种经典的检验方法, 是Hotelling[4]在1931年研究正态总体分布下均值向量的检验问题时所提出的, 该检验方法对于样本容量大于维数的问题具有良好的性质及检验效果, 但随着维数的不断增大, 当维数超过样本容量时该检验方法失效, 这是因为在高维数据下, 样本协方差不可逆导致该检验无定义. 为弥补这个缺陷, 许多学者进行了相关研究. Dempster[8,9]以两样本问题为基础, 首次指出Hotelling T^2检验不适用于高维数据的问题, 并在正态总体下就两样本显著性检验问题提出了一种非精确检验方法.

1.3.2 Bai和Saranadasa的检验

Bai和Saranadasa[10]以Dempster[8,9]的检验为基础, 得到了两样本问题的Hotelling T^2检验和Dempster非精确检验的渐进功效, 同时提出了一种渐近正态分布的新检验统计量T_{BS}, 新检验方法避免了求协方差矩阵逆, 即修正了Hotelling T^2检验中样本协方差矩阵不可逆的缺陷, 且该检验不依赖于正态假设.

首先作如下假设：

（1）$x_{ij}=\Gamma z_{ij}\mu_j, i=1,2,\cdots,N_j, j=1,2$, 其中$\Gamma$是一个$p\times m(m\leq\infty)$矩阵且$\Gamma\Gamma'=\Sigma$, z_{ij}是一个有m 个独立同分布分量的随机向量, 且满足$Ez_{ij}=0, Var(z_{ij})=I_m$, 设$v_1+v_2+v_3+v_4=4$, 若至少有一个$v_k=1$(相应的, 有两个$v_k=2$)时, 则有$Ez_{ijk}^4=3+\Delta<\infty$且$E\prod\limits_{k=1}^{m}z_{ijk}^{v_k}=0$(或1)；

（2）$p/n\longrightarrow y>0$且$N_1/(N_1+N_2)\longrightarrow\kappa\in(0,1)$；

（3）$\mu'\Sigma\mu=o(\tau\mathrm{tr}\Sigma^2), \lambda_{\max}=o(\sqrt{\mathrm{tr}\Sigma^2})$成立, 其中$\tau=(N_1+N_2)/N_1N_2, \mu=\mu_1-\mu_2$.

将以上假设为前提, Bai和Saranadasa提出了检验统计量：

$$\begin{aligned}T_{BS}&=\frac{(\overline{x}_1-\overline{x}_2)'(\overline{x}_1-\overline{x}_2)-\tau\mathrm{tr}(S_n)}{\tau\sqrt{\frac{2(n+1)n}{(n+2)(n-1)}[\mathrm{tr}S_n^2-\frac{1}{n}(\mathrm{tr}S_n)^2]}}\\&=\frac{\frac{N_1N_2}{N_1+N_2}(\overline{x}_1-\overline{x}_2)'(\overline{x}_1-\overline{x}_2)-\tau\mathrm{tr}S_n}{\sqrt{\frac{2(n+1)}{n}}B_n}\to N(0,1).\end{aligned}$$

其中

$$\overline{x}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}x_{ij},\ S_n=\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(x_{ij}-\overline{x}_i)(x_{ij}-\overline{x}_i)',\ i=1,2, n=N_1+N_2-2.$$

$$B_n^2=\frac{n^2}{(n+2)(n-1)}[\mathrm{tr}(S_n^2)-\frac{1}{n}(\mathrm{tr}S_n)^2].$$

当$T_{BS}>z_\alpha$时拒绝原假设.

1.3.3 Srivastava和Du的检验

Srivastava和Du[13]考虑了独立同分布的多元正态随机向量的均值检验问题, 为弥补高维数据下Hotelling T^2检验无定义这一缺陷, 用样本协方差矩阵S的对角线的元素构成一个新的对角阵来代替S, 即用一个非奇异矩阵代替奇异矩阵解决样本协方差矩阵不可逆的问题, 并通过模拟实验可知SD检验比Dempster的非精确检验以及Bai和Saranadasa的检验具有优越性.

对于假设检验问题(1.2), Srivastava和Du提出了检验统计量:

$$T_{SD}=\frac{n\overline{x}'D^{-1}\overline{x}-\frac{(n-1)p}{n-3}}{\sqrt{2(\mathrm{tr}R^2-\frac{p^2}{n-1})c_{p,n}}}.$$

其中$\overline{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i, S=\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})', D=\mathrm{diag}(s_{11},s_{22},\cdots,s_{pp}), R=D^{-\frac{1}{2}}SD^{-\frac{1}{2}}, c_{p,n}=1+\frac{\mathrm{tr}R^2}{p^{3/2}}$.

之后, Srivastava在文献 [14]中再次修正了该检验统计量, 使修正后的统计量不再局限于正态总体分布的假设前提.

1.3.4 Zhao和Xu的检验

Zhao和Xu在文献 [11, 12]中考虑高维数据总体均值的假设检验问题时, 首次在似然比检验的基础上借助并交原则, 提出了一种新的广义似然比检验. 在一定条件下, 得到了广义似然比检验统计量在原假设下的渐近分布, 并分析了广义似然比的渐近功效, 在渐近分布不成立的情形下, 用随机化方法得到了检验统计量的p值, 并通过模拟结果可知与其他检验相比, 该检验在某些情况下具有极大优势.

单样本情形: 假设$X_1,X_2,\cdots,X_n$来自p维正态分布$N_p(\mu,\Sigma), p\geq n$, 则假设检验问题(1.1)的广义似然比检验统计量为:

$$R(X)=\frac{1}{(2\pi)^{\frac{n}{2}}[1_n'(X'X)^{-1}1_n]^{-\frac{n}{2}}}\cdot \mathrm{e}^{-\frac{n}{2}},$$

其中$X=(X_1,X_2,\cdots,X_n)$, 当$R(X)$足够小时拒绝原假设.

两样本情形: 假设$X_1,X_2,\cdots,X_m$来自p维正态分布$N_p(\mu_1,\Sigma)$, $Y_1,Y_2,\cdots,Y_n$来自p维正态分布$N_p(\mu_2,\Sigma)$, $p\geq m+n$, 其中μ_1,μ_2和Σ未知, $\Sigma>0$. 假设问题(1.3)的广义似然比检验统计量为:

$$R(X,Y)=\frac{1}{(2\pi)^{\frac{m+n}{2}}[\frac{mn}{(m+n)^2}T(Z)]^{\frac{m+n}{2}}}\cdot \mathrm{e}^{-\frac{m+n}{2}}.$$

其中$J=\mathrm{diag}(1_m,1_n)$, $T(Z)=\begin{bmatrix}1 & -1\end{bmatrix}[J'(Z'Z)^{-1}J]^{-1}\begin{bmatrix}1\\-1\end{bmatrix}$.

由于$R(X,Y)$是$T(Z)$的单调减函数, 则可使用$T(Z)$作为假设检验问题(1.3)的广义似然比统计量, 当$T(Z)$足够大时拒绝原假设.

1.3.5 其他研究成果

Chen和Qin[15]以Bai和Saranadasa的研究为基础, 对两样本的均值向量的检验进行修正, 将T_{BS}中一项$\|\overline{X}_1 - \overline{X}_2\|^2$里对均值向量检验没有影响的一项去掉, 得到了一个具有更高检验功效的检验统计量, 并且新检验方法最大的优势是不再限制通常意义的“大p小n”, 此时所设的条件“大p小n”是指当n趋于无穷时, $\frac{p}{n}$ 趋于无穷, 这对解决更多高维数据的问题提供很大的灵活性.

Bai和Jiang[16]等人做了一项非常好的工作, 考虑了关于协方差矩阵的两个似然比检验, 先通过理论解释了这两个检验在高维数据下不成立的原因, 并运用高维随机F-矩阵线性谱统计量的中心极限定理对似然比检验进行修正, 通过模拟结果表明修正的检验在中等p值(约为20)和高维两种情况都有较好的性质.

Jiang和Yang[17]在$p, n \to \infty, \frac{p}{n} \to y \in (0, 1]$条件下, 利用中心极限定理研究了高维数据下正态分布均值和协方差矩阵的几种似然比检验问题, 并证明出该似然比统计量分布收敛于正态分布.

Schott[18]针对高维正态分布的数据, 以皮尔逊相关系数为基础(即通过任意两个变量间的线性相关性的大小来反应变量间独立性), 讨论了连续变量的完全独立性问题, 并表明当维数p和样本容量n满足一定条件时, 该统计量的渐进分布服从正态分布; 随后Schott[19]对高维数据下两组变量的独立性检验问题提出了新的检验方法, 并验证该检验在某种情况下具有优良性. Jiang, Bai和Zheng[20]针对两组变量的独立性问题, 当维数$p = p_1 + p_2$和样本量n同时成比例趋于无穷大时, 修正了似然比检验, 并通过数据模拟可知, 当维数相对于样本量较大时检验效果优良. 更多的关于高维独立性检验问题可参见文献 [21–23].

1.4 预备知识

1.4.1 似然比检验

似然比检验是数理统计中一种重要且有效的假设检验方法, 是Neyman和Pearson在1928年提出的检验方法, 这可以看成是Fisher极大似然估计法在检验问题中的延伸, 是一种构造检验的常用方法之一.

考虑p维总体$f(x, \theta)$, θ为未知参数, 取值空间为Θ, 又设Θ_0为Θ的一个子集, 当得到样本后要对下列假设

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta \backslash \Theta_0$$

做出判断. 设$x_1, x_2, \cdots, x_n$为来自具有密度$f(x, \theta)$的总体中取出的样本, 其似然函数记为

$$L(\theta; x) = L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta),$$

则似然比统计量定义为

$$\lambda(X) = \frac{L_0^*}{L_1^*},$$

其中

$$L_0^* = \max_{\theta\in\Theta_0} L(\theta; X), \quad L_1^* = \max_{\theta\in\Theta} L(\theta; X).$$

注意到$0 \le \lambda(X) \le 1$, 而且λ值越大越有利于原假设, 越小越不利于原假设. 用λ进行检验的法则是: 对给定的水平α, $0 < \alpha < 1$, 选取适当的临界值c, 使得对一切$\theta \in \Theta_0$, 有$P_\theta(\lambda \le c) \le \alpha$. 则当$\lambda \le c$时拒绝原假设, 否则接受原假设. 关于似然比检验更详细的内容见参考文献 [24].

1.4.2 并交原则(UIT)

并交原则(UIT)是由Roy[25]提出的一种假设检验构造方法, 该方法适合于原假设表示为集合的联合情况, 其原理可简述如下:

当原假设集容易表示成k个其他集的交集时, 即$\Theta_0 = \bigcap_{i=1}^{k} \Theta_i$, 此时假设检验问题可表示为:

$$H_0 : \theta \in \bigcap_{i=1}^{k} \Theta_i \leftrightarrow H_1 : \theta \in \bigcup_{i=1}^{k} \Theta_i^c. \tag{1.7}$$

$$H_{0i} : \theta \in \Theta_i \leftrightarrow H_{1i} : \theta \in \Theta_i^c. \tag{1.8}$$

即H_0成立的充要条件是所有的H_{0i}成立, 并称H_{0i}为H_0的一个成分. 每个H_{0i}的拒绝域记为R_i, 则H_0的拒绝域可表示为$R = \bigcup_{i=1}^{k} R_i$.

并交原则的基本思想是将原假设表示成更初等无穷个假设成分之交, 相比直接检验H_0的优点就是, 当原假设被拒绝时, 可以知道是哪一个成分假设被拒绝而造成的, 由此更容易了解偏离H_0 假设的某种本质原因. 在许多情况下, 由并交原则导出的检验统计量就是似然比检验导出的统计量. 关于并交原则的详细内容见参考文献 [24].

1.4.3 Bonferroni校正

Bonferroni校正是由Carlo Emilio Bonferroni[26]在1935年提出的一种较为严格的多重检验校正方法. 其内容可简述如下:

如果在同一个数据集上同时进行k个独立的假设检验, 设显著水平为α, 则用于每一个假设的检验水平α_0应为$\frac{\alpha}{k}$. 即当原假设H_0成立时, 设所有的H_i为假, 给定k个检验为T_i $(1 \le i \le k)$, 若对任意的i有

$$P(T_i\text{成立}|H_0) \le \frac{\alpha}{k},$$

则有

$$P(k个T_i同时成立|H_0) \leq \alpha.$$

其原理的本质思想为, 通过降低单次犯错误的概率, 从而降低总体犯错误的概率. 若对同一个数据集同时进行k次假设检验时, 当原假设成立时, 设A_i表示第i个检验犯第一类错误事件, 若每个检验犯错误的概率不超过$\frac{\alpha}{k}$, 则总的犯第一类错误的概率为**:**

$$P(A_1 \bigcup A_2 \bigcup \cdots \bigcup A_k) \leq P(A_1) + P(A_2) + \cdots + P(A_k) = \frac{\alpha}{k} \cdot k = \alpha.$$

从而保证了原假设的第一类犯错误概率不超过α. 关于Bonferroni校正的详细内容及研究可参考文献 [27–30].

第 2 章　高维数据正态均值的一种新检验

本章我们讨论高维数据下正态总体均值向量的假设检验问题, 当$p \geq n$时, Hotelling T^2检验中由于样本协方差矩阵S不可逆导致该检验方法失效, 为修正这一缺陷, Srivastava和Du用S的对角线元素构成新的对角阵D来代替S, 即用一个非奇异矩阵代替奇异矩阵来修正Hotelling T^2检验, 并通过模拟实验可知SD检验具有一定的优越性. 但Zhao[12]指出, SD检验方法虽然解决了样本协方差矩阵不可逆的问题, 但仍具有局限性, 这是因为该检验只考虑到样本协方差矩阵中对角线元素, 而并没有考虑到变量间的相关性问题, 这正是在变量相关性增强时SD检验效果不好的原因.

在现实生活中, 变量之间是完全独立的情况并不多见, 更多的情况是各变量之间存在相关性, 因此Zhao将“变量间具有相关性”这一问题考虑进去, 对SD检验进行了改进, 将变量两两组成一组, 即将样本协方差矩阵S分成$\frac{n}{2} \times \frac{n}{2}$块, 每一块是一个$2 \times 2$的矩阵, 用分块后对角线上的分块矩阵构成的新矩阵$D_S$来代替$S$, 通过对$S$划分方法的改进, 使得新检验统计量(记作$T_{block}$)增加了一些相关性信息, 并通过数据模拟结果显示, 该检验在变量相关性较大时检验效果优于T_{SD}.

但通过对T_{block}的数据结果观察, 我们发现当变量相关性较高时, T_{block}检验效果并不是很好, 这也就是Zhao在论文中提出的“并没有做到将S推广到用更大的分块对角阵来代替”, 也就是说, Zhao对S的划分并不是最优的划分方法, 划分后得到的矩阵D_S仍然存在变量间相关信息大量缺失的问题, 从而导致检验效果在某些条件下表现不好. 对于此缺陷, 本章对样本协方差矩阵S的划分提出了一种新的方法, 并结合Bonferroni校正的思想, 分别对高维数据下正态均值在单样本情形和两样本情形提出了新的检验统计量, 通过数据模拟可知, 当变量间相关性增强时, 新检验的检验效果与T_{block}相比更具有优良性.

Hotelling T^2检验是多元统计分析中一种经典的检验方法, 是Hotelling[4]在1931年研究正态总体分布下均值向量的检验问题时所提出的.

设$d \sim N_p(\mu, \Sigma), M \sim W_p(n, \Sigma), n \geq p, \Sigma > 0$且$d$与$M$相互独立, 则

$$\mathrm{T}^2 = nd'M^{-1}d$$

为具有自由度p和n的Hotelling T^2分布, 记为$T^2(p, n, \lambda)$, 其中$\lambda = \mu'\Sigma^{-1}\mu$为非中心参数, 当$\lambda = 0$时称为中心Hotelling T^2分布, 记作$T^2 \sim (p, n)$. 若$\mathrm{T}^2 \sim T^2(p, n, \lambda)$, 则

$$\frac{n-p+1}{np}\mathrm{T}^2 \sim F(p, n-p+1, \lambda).$$

对于假设检验问题(1.1), 经典的Hotelling T^2检验的表达式为

$$T^2_{(1)} = n(\overline{X} - \mu_0)'S^{-1}_{(1)}(\overline{X} - \mu_0),$$

其中$\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i=\frac{1}{n}X1_n,\ S_{(1)}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})'=\frac{1}{n-1}XH_nX'.$ 当

$$\frac{n-p}{(n-1)p}T_{(1)}^2\geq F_\alpha(p,n-p) \tag{2.1}$$

时, 拒绝原假设.

对于假设检验问题(1.3), 若$\Sigma_1=\Sigma_2$, 则经典的Hotelling T^2检验的表达式为

$$T_{(2)}^2=\frac{n_1n_2}{n_1+n_2}(\overline{X}-\overline{Y})'S_{(2)}^{-1}(\overline{X}-\overline{Y}),$$

其中

$$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i=\frac{1}{n_1}X1_{n_1},\ \overline{Y}=\frac{1}{n_2}\sum_{j=1}^{n_2}Y_j=\frac{1}{n_2}Y1_{n_2},$$

$$S_1=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})(X_i-\overline{X})',\ S_2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\overline{Y})(Y_j-\overline{Y})'.$$

$$S_{(2)}=\frac{1}{n_1+n_2-2}[(n_1-1)S_1+(n_2-1)S_2].$$

当

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T_{(2)}^2\geq F_\alpha(p,n_1+n_2-p-1) \tag{2.2}$$

时, 拒绝原假设.

对于假设检验问题(1.3), 若$\Sigma_1\neq\Sigma_2$, 但$n_1=n_2=n$时经典的Hotelling T^2检验的表达式为

$$T_{(3)}^2=\frac{n}{n-1}(\overline{X}-\overline{Y})'S_{(3)}^{-1}(\overline{X}-\overline{Y}),$$

其中$S_{(3)}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(Z_i-\overline{Z})(Z_i-\overline{Z})'=\frac{1}{n-1}ZH_nZ',\ Z_i=X_i-Y_i,\ \overline{Z}=\frac{1}{n}\sum\limits_{i=1}^{n}Z_i=\overline{X}-\overline{Y}.$
当

$$\frac{n-p}{p}T_{(3)}^2\geq F_\alpha(p,n-p) \tag{2.3}$$

时, 拒绝原假设.

当$\Sigma_1\neq\Sigma_2$且$n_1\neq n_2$时, 检验问题(1.3)就称作多元Behrens-Fisher问题, 此时传统方法无法给出精确的检验. 例如上述检验统计量$T_{(2)}^2$, 此时它的真实分布不再是$T^2(p,n_1+n_2-2)$, 因此需要找到$T_{(2)}^2$的近似分布, 其它传统方法基本上都是对$T_{(2)}^2$做一些修正, 相关内容可以参考文献 [34–37]; 另外还有学者用一些非传统的方法去构造精确解, 例如Gamage[31]在小样本检验中用广义p值得方法研究了两个总体的多元Behrens-Fisher问题; Gamage和Mathew等[32]又将该结果推广到了多个正态总体, 并给出了一种精确解; 范永辉[33]用广义p值和样本协方差矩阵的Bartlett分解给出Behrens-Fisher问题的精确解.

2.1 高维数据下正态均值在单样本情形下的检验

假设$X_1, X_2, \cdots, X_n$为来自p维正态分布$N_p(\mu, \Sigma)$的简单随机样本, 其中μ和Σ未知且$\Sigma > 0$, 考虑假设检验问题(1.1). 为方便起见, 我们记$X = (X_1, X_2, \cdots, X_n)$为样本阵, 那么样本均值向量$\overline{X}$和样本协方差矩阵$S$分别为

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{1}{n}X1_n,\ S = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})(X_j - \overline{X})' = \frac{1}{n-1}XH_nX'.$$

Srivastava和Du[13]所提出的方法为: 将Hotelling T^2检验中样本协方差阵S的对角线元素构成新的对角阵D来代替原样本协防差阵, 构造新的检验统计量:

$$T = n\overline{X}'D^{-1}\overline{X},$$

其中$D = \mathrm{diag}(s_{11}, s_{22}, \cdots, s_{pp}), s_{tt}$为样本协方差阵$S$中的第$i$个对角线元素, $t = 1, 2, \cdots, p$.

由此可以看出, 由矩阵D代替矩阵S, 只保留了S中对角线的元素, 而并没有考虑到变量间的相关性问题, 也就是说明当变量间为相互独立时, 该检验的检验效果良好, 若变量间存在相关性时, 该检验由于不包含变量间的相关信息导致检验效果较差, 为了改善这一点我们提出了新的划分方法, 试图划分成更大的分块, 也就是尽可能的多保留S中变量间相关性的信息.

由Bonferroni校正思想可知, 首先要将样本阵X按维数分成k个部分, 再对划分后的k个部分进行同时检验, 且当原假设成立时, 令其中任意一个检验犯第一类错误的概率均不超过$\frac{\alpha}{k}$, 由此控制原假设的第一类犯错误概率不超过α. 此外将样本阵划分这一步骤也就是“降维”的一个过程, 将原来的高维样本阵划分成了k个低维阵, 此时每个部分都可以运用Hotelling T^2这一经典的假设检验理论进行分析处理.

问题的关键在于如何将样本阵X按维数进行合理且适当地划分, 若划分的块数太多, 则会出现样本变量间相关信息大量缺失的现象, 例如T_{block}在变量间线性相关度越高时效果表现并不好; 若划分的块数较少, 则可能会出现分块后的每部分矩阵仍是高维数据阵, 即每块样本阵的维数仍大于样本容量, 没有达到降维的效果.

由此我们提出一个适当的划分方法, 令

$$k = \lfloor \frac{p}{n-1} \rfloor + 1(\lfloor x \rfloor \text{表示小于或等于} x \text{的最大整数}), \tag{2.4}$$

其中$\frac{p}{n-1}$不能整除. 按此将样本阵X划分后既保证了变量间的相关性, 又保证了划分后的每个分块均不再是高维数据阵. 为方便起见, 我们将划分后的样本阵记作$X = (X'_{(1)}, X'_{(2)}, \cdots, X'_{(k)})'$, 其中$X_{(1)}, X_{(2)}, \cdots, X_{(k-1)}$均为$(n-1) \times n$矩阵, $X_{(k)}$为$[p - (n-1)(k-1)] \times n$矩阵, 相应地, 期望向量以及样本协方差矩阵也进行相同的划分, 则可写成如下形式:

$$\mu = (\mu'_1, \mu'_2, \cdots, \mu'_k)', \mu_0 = (\mu'_{10}, \mu'_{20}, \cdots, \mu'_{k0})',$$

$$S=\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1(k-1)} & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2(k-1)} & S_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{k(k-1)} & S_{kk} \end{bmatrix}. \tag{2.5}$$

其中$\mu_1,\mu_2,\cdots,\mu_{k-1}$为$n-1$维向量, μ_k为$[p-(n-1)(k-1)]$维向量, $S_{11},S_{22},\cdots,S_{(k-1)(k-1)}$均为$(n-1)\times(n-1)$矩阵, S_{kk}为$[p-(n-1)(k-1)]\times[p-(n-1)(k-1)]$矩阵.

此时由Bonferroni校正思想可知, 假设检验问题(1.1)可看成如下k个问题同时检验:

$$H_{0i}:\mu_i=\mu_{i0}\leftrightarrow H_{1i}:\mu_i\neq\mu_{i0},i=1,2,\cdots,k. \tag{2.6}$$

在原假设成立时, 控制(2.6)中每个检验犯第一类错误的概率不超过$\frac{\alpha}{k}$, 由此将原假设(1.1)的第一类犯错误概率控制在α, 且每个部分都可以用Hotelling T^2进行检验.下面给出该方法的具体结论.

当$p>n$时, 考虑假设检验问题(1.1). 对于假设检验问题(2.6), 任意一个i, $H_{0i}:\mu_i=\mu_{i0}$的检验统计量均为一个Hotelling T^2检验, 但对于不同的i, 检验统计量的具体形式是不一样的.

每部分的Hotelling T^2形式可表示为:

$$T^2_{new1.i}=n(\overline{X}_{(i)}-\mu_{i0})'S_{ii}^{-1}(\overline{X}_{(i)}-\mu_{i0}).$$

当$i=1,2,\cdots,k-1$时, 每块样本阵的维数为$n-1$, 样本容量为n, 由(2.1)可知, 对给定的检验水平α, 当

$$\frac{1}{(n-1)^2}T^2_{new1.i}\geq F_{\frac{\alpha}{k}}(n-1,1) \tag{2.7}$$

时, 拒绝原假设H_{0i}.

当$i=k$时, 该部分样本阵的维数为$[p-(n-1)(k-1)]$, 样本容量为n, 由(2.1)可知, 对给定的检验水平α, 当

$$\frac{(n-1)k+1-p}{[p-(n-1)(k-1)](n-1)}T^2_{new1.k}\geq F_{\frac{\alpha}{k}}([p-(n-1)(k-1)],(n-1)k+1-p) \tag{2.8}$$

时, 拒绝原假设H_{0k}.

其中S_{ii}为(2.5)中矩阵S的第(i,i)位置的分块矩阵, $i=1,2,\cdots,k$. 由(2.7),(2.8) 可以得到k个检验, 结合Bonferroni校正思想, 将k个部分同时进行检验, 若其中任意一个H_{0i}被拒绝, 则检验问题(1.1)中的H_0被拒绝.

2.2 高维数据下正态均值在两样本情形下的检验

2.2.1 两个总体协方差阵相等时的情形

对于假设检验问题(1.3), 为方便起见, 记$X=(X_1,X_2,\cdots,X_{n_1}),Y=(Y_1,Y_2,\cdots,Y_{n_2})$为样本阵, 相应地, 均值向量和样本协方差矩阵可以表示为

$$\overline{X}=\frac{1}{n_1}\sum_{s=1}^{n_1}X_s=\frac{1}{n_1}X1_{n_1},\ S_1=\frac{1}{n_1-1}\sum_{s=1}^{n_1}(X_s-\overline{X})(X_s-\overline{X})'=\frac{1}{n_1-1}XH_{n_1}X',$$

$$\overline{Y}=\frac{1}{n_2}\sum_{t=1}^{n_2}Y_t=\frac{1}{n_2}Y1_{n_2},\ S_2=\frac{1}{n_2-1}\sum_{t=1}^{n_2}(Y_t-\overline{Y})(Y_t-\overline{Y})'=\frac{1}{n_2-1}YH_{n_2}Y'.$$

同样运用Bonferroni校正思想, 我们提出一个适当的划分方法, 令

$$k=\lfloor\frac{p}{n_1+n_2-2}\rfloor+1(\lfloor x\rfloor表示小于或等于x的最大整数). \tag{2.9}$$

按此将样本阵X和Y划分后既保证了变量间的相关性, 又保证了划分后的每个分块均不再是高维数据阵. 为方便起见, 将划分后的样本阵分别记为$X=(X'_{(1)},X'_{(2)},\cdots,X'_{(k)})$和$Y=(Y'_{(1)},Y'_{(2)},\cdots,Y'_{(k)})'$, 其中$X_{(1)},X_{(2)},\cdots,X_{(k-1)}$为$(n_1+n_2-2)\times n_1$ 矩阵, $Y_{(1)},Y_{(2)},\cdots,Y_{(k-1)}$为$(n_1+n_2-2)\times n_2$矩阵, $X_{(k)}$为$[p-(n_1+n_2-2)(k-1)]\times n_1$矩阵, $Y_{(k)}$ 为$[p-(n_1+n_2-2)(k-1)]\times n_2$矩阵, 相应地, 期望向量以及样本协方差矩阵也进行相同的划分, 则可写成如下形式:

$$\mu_1=(\mu'_{11},\mu'_{12},\cdots,\mu'_{1k})',\mu_2=(\mu'_{21},\mu'_{22},\cdots,\mu'_{2k})',$$

$$S_1=\begin{bmatrix}S_{1.11}&S_{1.12}&\cdots&S_{1.1(k-1)}&S_{1.1k}\\S_{1.21}&S_{1.22}&\cdots&S_{1.2(k-1)}&S_{1.2k}\\\vdots&\vdots&\ddots&\vdots&\vdots\\S_{1.k1}&S_{1.k2}&\cdots&S_{1.k(k-1)}&S_{1.kk}\end{bmatrix}, \tag{2.10}$$

$$S_2=\begin{bmatrix}S_{2.11}&S_{2.12}&\cdots&S_{2.1(k-1)}&S_{2.1k}\\S_{2.21}&S_{2.22}&\cdots&S_{2.2(k-1)}&S_{2.2k}\\\vdots&\vdots&\ddots&\vdots&\vdots\\S_{2.k1}&S_{2.k2}&\cdots&S_{2.k(k-1)}&S_{2.kk}\end{bmatrix}. \tag{2.11}$$

其中$\mu_{11},\mu_{12},\cdots,\mu_{1(k-1)}$和$\mu_{21},\mu_{22},\cdots,\mu_{2(k-1)}$为$n_1+n_2-2$维向量, μ_{1k}和μ_{2k}为$[p-(n_1+n_2-2)(k-1)]$维向量, $S_{1.11},S_{1.22},\cdots,S_{1.(k-1)(k-1)}$和$S_{2.11},S_{2.22},\cdots,S_{2.(k-1)(k-1)}$均为$[n_1+n_2-2]\times[n_1+n_2-2]$ 矩阵, $S_{1.kk}$和$S_{2.kk}$为$[p-(n_1+n_2-2)(k-1)]\times[p-(n_1+n_2-2)(k-1)]$ 矩阵.

由Bonferroni校正思想可知, 假设检验问题(1.3)可看成如下k个问题同时检验:

$$H_{0i}:\mu_{1i}=\mu_{2i}\leftrightarrow H_{1i}:\mu_{1i}\neq\mu_{2i}. \tag{2.12}$$

在原假设成立时, 控制(2.12)中每个检验犯第一类错误的概率不超过$\frac{\alpha}{k}$, 由此将原假设(1.3)的第一类犯错误概率控制在α, 且每个部分都可以用Hotelling T^2进行检验. 下面给出该方法

的具体结论.

当$p > n_1 + n_2 - 2$时, 考虑假设检验问题(1.3). 当$\Sigma_1 = \Sigma_2$时, 对于假设检验问题(2.12), 任意一个i, $H_{0i} : \mu_i = \mu_{i0}$的检验统计量是一个Hotelling T^2检验, 但对于不同的i, 检验统计量的具体形式是不一样的.

每部分的Hotelling T^2的形式可表示为

$$T^2_{new2.i} = \frac{n_1 n_2}{n_1 + n_2}(\overline{X}_{(i)} - \overline{Y}_{(i)})' S_i^{-1} (\overline{X}_{(i)} - \overline{Y}_{(i)}).$$

当$i = 1, 2, \cdots, k-1$时, 每块样本阵的维数为$n_1 + n_2 - 2$, 由(2.2)可知, 对给定的检验水平α, 当

$$\frac{1}{(n_1 + n_2 - 2)^2} T^2_{new2.i} \geq F_{\frac{\alpha}{k}}(n_1 + n_2 - 2, 1) \tag{2.13}$$

时, 拒绝原假设H_{0i}.

当$i = k$时, 该部分样本阵的维数为$[p - (n_1 + n_2 - 2)(k-1)]$, 由(2.2)可知, 对给定的检验水平$\alpha$, 当

$$\frac{(n_1 + n_2)k - 2k - p + 1}{(n_1 + n_2 - 2)[p - (n_1 + n_2 - 2)(k-1)]} T^2_{new2.k} \geq F_{\frac{\alpha}{k}}(p-(n_1+n_2-2)(k-1), (n_1+n_2)k-2k-p+1) \tag{2.14}$$

时, 拒绝原假设H_{0k}.

其中$S_i = \frac{1}{n_1+n_2-2}[(n_1 - 1)S_{1.ii} + (n_2 - 1)S_{2.ii}]$, $S_{1.ii}$为(2.10)中矩阵S_1的第(i, i)位置的分块矩阵, $S_{2.ii}$为(2.11)中矩阵S_2的第(i, i)位置的分块矩阵, $i = 1, 2, \cdots, k$. 由(2.13), (2.14)可以得到k个检验, 结合Bonferroni校正思想, 将k个部分同时进行检验, 若其中任意一个H_{0i}被拒绝, 则检验问题(1.3)中的H_0被拒绝.

2.2.2 两总体协方差阵不等时的情形

对于假设检验问题(1.3), 当$\Sigma_1 \neq \Sigma_2$时, 对于特殊情形$n_1 = n_2 = n$时, 可将样本进行(2.4)的方法进行划分, 结合Bonferroni校正思想, 用配对检验即可得到精确检验. 其思想原理同上述两节, 此时检验问题(1.3)可看成k个(2.12)问题同时检验, 同样控制(2.12)中每个检验犯第一类错误的概率不超过$\frac{\alpha}{k}$, 由此将原假设(1.3)的第一类犯错误概率控制在α, 且每个部分都可以用Hotelling T^2进行检验.下面给出该方法的具体结论.

当$\Sigma_1 \neq \Sigma_2$, $p > n_1 + n_2 - 2$且$n_1 = n_2 = n$时, 考虑假设检验问题(1.3). 对于假设检验问题(2.12), 任意一个i, $H_{0i} : \mu_i = \mu_{i0}$的检验统计量是一个Hotelling T^2检验, 但对于不同的i, 检验统计量的具体形式是不一样的:

每部分的Hotelling T^2形式为

$$T^2_{new3.i} = \frac{n}{n-1}(\overline{X}_{(i)} - \overline{Y}_{(i)})' S_i^{-1} (\overline{X}_{(i)} - \overline{Y}_{(i)}),$$

当$i=1,2,\cdots,k-1$时, 每块样本阵的维数为$n-1$, 样本容量为n, 由(2.3)可知, 对给定的检验水平α, 当

$$\frac{1}{n-1}T^2_{new3.i}\geq F_{\frac{\alpha}{k}}(n-1,1) \tag{2.15}$$

时, 拒绝原假设H_{0i}.

当$i=k$时, 该部分样本阵的维数为$[p-(n-1)(k-1)]$, 样本容量为n, 由(2.3)可知, 对给定的检验水平α, 当

$$\frac{nk+1-k-p}{p-(n-1)(k-1)}T^2_{new3.k}\geq F_{\frac{\alpha}{k}}([p-(n-1)(k-1)],nk+1-k-p) \tag{2.16}$$

时, 拒绝原假设H_{0k}.

其中$S_i=\frac{1}{n-1}\sum\limits_{i=1}^{n}(Z_{(i)}-\overline{Z}_{(i)})(Z_{(i)}-\overline{Z}_{(i)})', Z_{(i)}=X_{(i)}-Y_{(i)}, \overline{Z}_{(i)}=\frac{1}{n}\sum\limits_{i=1}^{n}Z_{(i)}=\overline{X}_{(i)}-\overline{Y}_{(i)}, i=1,2,\cdots,k$. 由(2.15)(2.16)可以得到$k$个检验, 结合Bonferroni 校正思想, 将k个部分同时进行检验, 若其中任意一个H_{0i}被拒绝, 则检验问题(1.3)中的H_0被拒绝.

对于假设检验问题(1.3), 当$\Sigma_1\neq\Sigma_2$且$n_1\neq n_2$时, 该检验称为Behrens-Fisher问题, 此时经典的Hotelling T^2统计量无法给出精确的检验. 广义p值方法给出了Behrens-Fisher问题的一个精确解. 广义p值的概念是由Tusi和Weerahandi[44]提出, 其主要目的是在小样本中克服在检验中多余参数带来的困难. 为解决上述问题, 我们仍用Bonferroni校正思想, 将样本按(2.9)的形式进行划分, 并得到检验问题(2.12)的k个检验, 划分后的每个部分均用广义p检验法代替Hotelling T^2 检验法进行检验.

Gamage[31]对两样本的多元Behrens-Fisher问题提出了广义p值的定义如下:

$$P(F\times(\frac{k_1}{B}+\frac{k_2}{1-B})\geq\frac{\delta^4}{p}),$$

对于给定的检验水平α, 若上式小于α, 则拒绝原假设. 其中$F\sim F(p,n_1+n_2-2), B\sim\text{Beta}((n_1-1)/2,(n_2-1)/2), F$和$B$均为独立的, $k_1=(n_1-1)s_1^2/n_1(n_1+n_2-2), k_2=(n_2-1)s_2^2/n_2(n_1+n_2-2)$. 设$d=[1/n_1S_1+1/n_2S_2]^{-1}(\overline{X}-\overline{Y})$, d为$d$的观测值, $\sigma_1^2=\text{d}'\Sigma_1\text{d}, \sigma_2^2=\text{d}'\Sigma_2\text{d}, S_1^2=\text{d}'S_1\text{d}, S_2^2=\text{d}'S_2\text{d}, s_1^2$和$s_2^2$为$S_1^2$和$S_2^2$ 的观测值, $\delta^2=(\overline{x}-\overline{y})'[1/n_1s_1+1/n_2s_2]^{-1}(\overline{x}-\overline{y}), \overline{x}$和$\overline{y}$分别为$\overline{X}$和$\overline{Y}$ 的观测值.

经过上述划分可得到k个检验问题, 结合Bonferroni校正思想, 将k个部分同时进行检验, 若其中任意一个H_{0i}被拒绝, 则检验问题(1.3)中的H_0被拒绝.

2.3 数值模拟

本小节运用R语言对提出的三个新检验的检验水平分别进行数据模拟. 我们取检验水平$\alpha=0.05$, 具体操作步骤如下：

(1)在总体中抽取样本, 得到样本阵X, 并将X按前$k-1$个部分和第k个部分进行划分;

(2)计算出每个部分的统计量$T_i(i=1,2,\cdots,k)$, 令k个检验同时进行, 并计算出这k个检验中拒绝原假设H_{0i}的次数$b(0\le b\le k)$;

(3)若$b\ge 1$则拒绝原假设H_0;

(4)将步骤(1)-(3)重复1000次, 记其中拒绝原假设H_0的次数为c, 用$\alpha_0=\frac{c}{1000}$来估计实际的检验水平.

模拟1 将新检验T_{new1}与与Zhao所提出的检验T_{block}进行对比. 为了实验的准确性, 我们运用了Zhao在实验中总体协方差的表达形式, 即用选取的方差阵$\mathcal{D}_\sigma$和总体相关系数阵$\mathcal{R}$确定总体协方差矩阵Σ, 其表达式为:

$$\Sigma=\mathcal{D}_\sigma^{\frac{1}{2}}\mathcal{R}\mathcal{D}_\sigma^{\frac{1}{2}}.$$

并选取如下不同的总体相关阵和方差阵:

$$\begin{aligned}
&\mathcal{R}_1=I_p,\\
&\mathcal{R}_2=(\rho_{ij}):\rho_{ii}=1,\rho_{ij}=0.25,i\neq j,\\
&\mathcal{R}_3=(\rho_{ij}):\rho_{ii}=1,\rho_{ij}=0.5,i\neq j;\\
&\mathcal{D}_{\sigma,1}=I_p,\\
&\mathcal{D}_{\sigma,2}:\sigma_{11}^{\frac{1}{2}},\cdots,\sigma_{pp}^{\frac{1}{2}}\sim Unif(2,3),\\
&\mathcal{D}_{\sigma,3}:\sigma_{11},\cdots,\sigma_{pp}\sim\chi^2(3).
\end{aligned}$$

模拟2 为了检验所提出的新统计量T_{new2}和T_{new3}的优良性, 我们考虑以下四个不同的协方差矩阵形式, 以便观察变量间相关性大小对检验水平的影响.

$$\begin{aligned}
&W_1=I_p;\\
&W_2=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.25,i\neq j;\\
&W_3=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.5,i\neq j;\\
&W_4=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.75,i\neq j.
\end{aligned}$$

表1为多元正态分布下T_{block}与T_{new1}的检验水平; 表2为多元正态分布下, 两样本协方差矩阵相等时T_{new2}的检验水平, 不失一般性地, 我们令$\mu_1=\mu_2=\mathbf{0}$; 表3为多元正态分布下, 两样本协方差矩阵不相等, 样本容量相等时T_{new3}的检验水平;

由表1可看出当$\mathcal{R}=\mathcal{R}_1$, $\mathcal{R}=\mathcal{R}_2$时, T_{block}与T_{new1}的检验水平均能较好的维持在0.05附近, 但当$\mathcal{R}=\mathcal{R}_3$时, T_{block}检验水平没有维持在0.05左右, 而T_{new1}的检验水平仍能很好的维持在0.05左右, 这说明较T_{block}相比, 我们所提出的T_{new1}更具有优良性.

由表2我们可以看出, 当n_1和n_2相差不多的时候, 新检验T_{new2}的检验水平能较好的维持在0.05左右, 但随着两样本的样本量差值的增大, 检验效果相应变差.

表3给出了高维数据下$\Sigma_1\neq\Sigma_2$且$n_1=n_2=n$时新检验T_{new3}的检验水平, 并由模拟结果显示检验水平能很好的维持在0.05左右, 这说明我们所提出的新检验具有一定的优良性.

2.4 小结与展望

本章在高维数据正态总体的条件下, 对于单样本问题和两样本问题共提出了T_{new1}, T_{new2}和T_{new3}三个新检验统计量, 并通过数据模拟可知新检验统计量在某种情况下具有优良性. 新统计量的提出想法主要就是弥补Hotelling T^2检验中样本协方差矩阵不可逆的缺陷, 随着变量间相关关系的逐渐增强, 需要用到更多样本协方差的非对角线元素来代替样本协方差, 本节划分方法的提出较好的解决了这一问题, 将样本协方差阵划分成更大的分块, 也就是利用了变量间更多的相关信息, 这也正是新检验效果优于SD检验和T_{block}的最根本原因.

但本章仍有不足之处, 第一, 由于没有找到较为合适的对比对象, 在T_{new2}和T_{new3}的数据模拟中没有提出对比检验; 并且没有对三个检验的检验功效进行数值模拟, 我们的检验结果只保证了犯第一类错误概率足够小, 但当检验水平能很好的控制下并没有计算出犯第二类错误的情况; 第二, 由于本章的出发点在于在高维数据下修正Hotelling T^2检验, 因此对于两样本协方差不等时, 只对两个样本容量相等的情况进行了数值模拟, 而对于样本不相等的情况, 在本章只给出了对样本的划分方法及检验方法, 具体的检验统计量形式及数值模拟等仍需要进一步研究讨论.

在我们提出的方法中, 若k太大的话, 相对于每一部分做检验时的检验水平$\frac{\alpha}{k}$可能太小, 所以下面我们进一步思考: 能否将k个检验统计量合并成在一起. T_{BS}检验统计量的分子$(\overline{x}_1-\overline{x}_2)'(\overline{x}_1-\overline{x}_2)-\tau\mathrm{tr}(S_n)$, 它实际上可以写为$\sum\limits_{j=1}^{p}[(\overline{x}_{1j}-\overline{x}_{2j})^2-\tau\mathrm{tr}(S_{n,jj})]$, 其中$\overline{x}_{ij}$表示$\overline{x}_i$的第$j$个分量, $S_{n,jj}$是S_n的第j 个对角元, $\tau\mathrm{tr}(S_{n,jj})$是$E(\overline{x}_{ij}-\overline{x}_{2j})^2$的一个估计, 所以我们考虑令

$$K=\sum_{j=1}^{k}(T^2_{new3.j}-a_j),$$

其中a_j是在原假设成立下的$E(T^2_{new.j})$, 是一个与j有关的常数, 类似于T_{BS}, 适当的对K做修正, 得到一个适合的检验统计量.

表 1: 多元正态分布下两种检验的水平比较

		$\mathcal{D}=\mathcal{D}_{\sigma,1}$		$\mathcal{D}=\mathcal{D}_{\sigma,2}$		$\mathcal{D}=\mathcal{D}_{\sigma,3}$	
	(p,n)	T_{block}	T_{new1}	T_{block}	T_{new1}	T_{block}	T_{new1}
$\mathcal{R}=\mathcal{R}_1$	(100,40)	0.049	0.052	0.058	0.048	0.057	0.056
	(100,60)	0.055	0.045	0.058	0.052	0.055	0.049
	(200,40)	0.055	0.044	0.053	0.051	0.056	0.05
	(200,60)	0.051	0.048	0.052	0.048	0.054	0.055
$\mathcal{R}=\mathcal{R}_2$	(100,40)	0.063	0.05	0.058	0.052	0.063	0.048
	(100,60)	0.061	0.047	0.055	0.048	0.060	0.049
	(200,40)	0.055	0.055	0.059	0.052	0.053	0.048
	(200,60)	0.054	0.046	0.057	0.045	0.051	0.048
$\mathcal{R}=\mathcal{R}_3$	(100,40)	0.035	0.049	0.035	0.053	0.035	0.049
	(100,60)	0.033	0.048	0.038	0.046	0.037	0.047
	(200,40)	0.032	0.045	0.03	0.044	0.031	0.049
	(200,60)	0.035	0.052	0.033	0.045	0.034	0.05

表 2: $\Sigma_1 = \Sigma_2$时, T_{new2}的检验水平

	(n_1, n_2)	W_1	W_2	W_3	W_4
p=60	$(15, 15)$	0.054	0.053	0.048	0.053
	$(20, 20)$	0.048	0.046	0.044	0.05
	$(25, 25)$	0.054	0.043	0.05	0.044
	$(12, 15)$	0.057	0.064	0.053	0.055
	$(15, 18)$	0.052	0.057	0.055	0.058
	$(20, 24)$	0.061	0.068	0.07	0.056
	$(13, 26)$	0.103	0.104	0.107	0.104
p=100	$(20, 20)$	0.047	0.054	0.055	0.04
	$(45, 45)$	0.053	0.044	0.049	0.048
	$(16, 20)$	0.051	0.06	0.061	0.041
	$(20, 24)$	0.063	0.054	0.052	0.052
	$(25, 30)$	0.044	0.047	0.062	0.041
	$(45, 48)$	0.053	0.048	0.049	0.045
	$(20, 48)$	0.139	0.168	0.111	0.135
p=150	$(35, 35)$	0.052	0.047	0.047	0.045
	$(70, 70)$	0.051	0.049	0.05	0.049
	$(30, 35)$	0.058	0.053	0.043	0.045
	$(30, 45)$	0.068	0.057	0.067	0.062
	$(40, 60)$	0.089	0.116	0.099	0.124
	$(40, 80)$	0.124	0.142	0.144	0.145
	$(20, 90)$	0.164	0.172	0.174	0.151

表 3: $\Sigma_1 \neq \Sigma_2$时, T_{new3}的检验水平

		(W_1, W_2)	(W_1, W_3)	(W_1, W_4)	(W_2, W_3)	(W_2, W_4)	$W_3, W_4)$
p=60	$n = 15$	0.041	0.045	0.042	0.047	0.051	0.049
	$n = 20$	0.048	0.046	0.05	0.044	0.045	0.053
	$n = 25$	0.048	0.047	0.048	0.05	0.044	0.048
	$n = 40$	0.053	0.046	0.053	0.046	0.051	0.051
p=100	$n = 20$	0.05	0.04	0.051	0.051	0.044	0.056
	$n = 45$	0.044	0.056	0.047	0.055	0.049	0.05
	$n = 60$	0.05	0.049	0.044	0.054	0.055	0.055
	$n = 80$	0.044	0.042	0.042	0.046	0.05	0.057
p=150	$n = 35$	0.049	0.048	0.041	0.056	0.047	0.055
	$n = 70$	0.055	0.045	0.052	0.048	0.048	0.052
	$n = 95$	0.046	0.055	0.056	0.047	0.055	0.049
	$n = 130$	0.049	0.051	0.045	0.057	0.051	0.044

第 3 章　高维数据下的方差分析

本章我们讨论高维数据下方差分析中的检验问题(1.4). 当$n \geq p+k$时, 这就是经典的多元方差分析问题, 用似然比方法得到的检验统计量服从Wilks[38]Λ分布; 当$n < p+k$时, 由于高维数据下样本协方差矩阵不可逆, 这导致经典的似然比检验失效. 当$k \leq 2$时, Zhao首次结合并交原则分别对单样本($k=1$)和两样本($k=2$)的总体均值检验问题提出了具有优势的广义似然比检验. 本章我们以Zhao提出的新检验方法为基础, 继续研究, 将样本个数k推广到$k \geq 2$, 给出了高维数据下方差分析问题的广义似然比检验, 并通过蒙特卡洛方法模拟了检验水平, 结果显示该方法能很好的控制检验水平.

3.1 高维数据下的方差分析

对于假设检验问题(1.4), 为方便起见我们记$X_i = (X_{i1}, X_{i2}, \cdots, X_{in_i})$ 为从第i个总体X$_i$中抽取的简单随机样本构成的数据矩阵, $X = (X_1, X_2, \cdots, X_k)$, 其中$n = \sum_{i=1}^{k} n_i$为总样本量. 由并交原则可知假设检验问题(1.4)中的原假设H_0可以写成如下检验问题：

$$H_{0a}: a'\mu_1 = a'\mu_2 = \cdots = a'\mu_k \leftrightarrow H_{1a}: a'\mu_1, a'\mu_2, \cdots, a'\mu_k \text{不全相等}. \tag{3.1}$$

中的原假设H_{0a}之交, 即$H_0 = \bigcap_{a'a=1} H_{0a}$, 其中$a \in R^p$. 若拒绝某一个$H_{0a}$则应拒绝$H_0$, 所有的$H_{0a}$不被拒绝则不拒绝$H_0$. 因此我们的任务就是在所有的$H_{0a}$中找到最容易被拒绝的$H_{0a^*}$, 若$H_{0a^*}$被拒绝, 则$H_0$应被拒绝; 若$H_{0a^*}$不被拒绝, 则不拒绝$H_0$.

容易得知$a'X_1, a'X_2, \cdots, a'X_k$相互独立且有$a'X_i \sim N(a'\mu_i, a'\Sigma a), i = 1, 2, \cdots, k$. 对给定的$a$, 检验问题(3.1)就是一元方差分析问题, 似然比检验统计量为

$$\lambda_a = \frac{L_H(a)}{L(a)}.$$

其中

$$L_H(a) = \sup_{\mu_1=\mu_2=\cdots=\mu_k,\Sigma} L(a'\mu_1 = a'\mu_2 = \cdots = a'\mu_k) = (2\pi)^{-\frac{n}{2}}[\frac{1}{n}a'(\text{SSW}+\text{SSB})a]^{-\frac{n}{2}} \cdot \exp\{-\frac{n}{2}\},$$

$$L(a) = \sup_{\mu_1,\mu_2,\cdots,\mu_k,\Sigma} L(a'\mu_1 = a'\mu_2 = \cdots = a'\mu_k) = (2\pi)^{-\frac{n}{2}}(\frac{1}{n}a'\text{SSW}a)^{-\frac{n}{2}} \cdot \exp\{-\frac{n}{2}\}.$$

其中$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'$ 是组内离差阵, $\text{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})(\bar{X}_i-\bar{X})'$ 是组间离差阵.

结合并交原则可知, 较小的λ_a对于拒绝H_{0a}提供了有力的保障, 因此我们需要找到一个a^*, 使得λ_{a^*}在所有的λ_a中是最小的, 即

$$\lambda_{a^*}=\min_{a\neq 0}\lambda_a=\min_{a\neq 0}\frac{L_H(a)}{L(a)}=\min_{a\neq 0}\frac{[a'(\mathrm{SSW}+\mathrm{SSB})a]^{-\frac{n}{2}}}{(a'\mathrm{SSW}a)^{-\frac{n}{2}}}. \tag{3.2}$$

若令

$$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij},\ \bar{X}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{X}_i,$$

$$S=\frac{1}{n}XH_nX'=\frac{1}{n}(n_1S_1+n_2S_2+\cdots+n_kS_k),\ S_i=\frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'=\frac{1}{n_i}X_iH_{n_i}X_i',$$

且有$T=nS=\mathrm{SSW}+\mathrm{SSB}, C=\mathrm{diag}\{H_{n_1},H_{n_2},\cdots,H_{n_k}\}$, 则SSW和SSB也可以分别表示为

$$\mathrm{SSW}=\sum_{i=1}^{k}n_iS_i=XCX',\quad \mathrm{SSB}=T-\mathrm{SSW}=X(H_n-C)X'.$$

对于(3.2), 当$n<p+k$时, 存在某些a, 使得$a'\mathrm{SSW}a=0$以及$L(a)=\infty$, 此时$\lambda_a=\frac{L_H(a)}{L(a)}$不能再作为检验$H_{0a}$的统计量. 我们考虑当$L(a)=\infty$时, $L_H(a)$越小越倾向于拒绝H_{0a}. 由于H_{0a^*}是H_{0a}中最容易被拒绝的, 则应找到一个a^*, 其中$a^*\in A=\{a|a\in R^p, a'a=1, L(a)=\infty\}$并使得$L_H(a)$达到最小, 此时$H_{0a^*}$的似然比统计量应为$L_H(a^*)=\min_{a\in A}L_H(a)$, 并将其作为$H_0=\bigcap_{a'a=1}H_{0a}$的检验统计量进行讨论.

定义3.1 当$p\geq n$时, 假设检验问题(1.4)的广义似然比检验统计量定义为:

$$Z(X)=L_H(a^*)=\min_{a\in A}L_H(a).$$

其中$a^*\in A$且使得$L_H(a)$达到最小值, $Z(X)$越小越倾向于拒绝原假设H_0.

引理3.1 若在条件$a'\mathrm{SSW}a=0$和$a'a=1$下使得$a'\mathrm{SSB}a$达到最大值, 则必有$a\in\Re(X)$.

证明 由于C是对称幂等矩阵, $\mathrm{SSW}=XCX'=XCC'X'$, 则有

$$a'\mathrm{SSW}a=0\Leftrightarrow a'XCC'Xa=0\Leftrightarrow (C'X'a)'(C'X'a)=0\Leftrightarrow C'X'a=0.$$

令$J=\mathrm{diag}(1_{n_1},1_{n_2},\cdots,1_{n_k})$, 由于$\Re(J)$是$\Re(C)$的正交补空间, $C'X'a=0$意味着$X'a\in\Re^{\perp}(C)=\Re(J)$, 故存在$k\times 1$的向量$b$使得$X'a=Jb$. 由于$X'$是行满秩矩阵, 故对任意的$b$, $X'a=Jb$总有解. 由相容线性方程组通解[39]可得:

$$a=(X')^{+}Jb+[I_p-(X')^{+}X']\beta=(X')^{+}Jb+(I_p-P_X)\beta, \text{ 其中}\beta\text{是}p\text{维任意向量}. \tag{3.3}$$

由$X'a=Jb$可得

$$a'\mathrm{SSB}a=(a'X)(H-C)(X'a)=b'J'(H-C)Jb. \tag{3.4}$$

由于$J'(H-C)J$是半正定矩阵, 故$\sup_{a'a=1,a'\mathrm{SSW}a=0}a'\mathrm{SSB}a>0$. 又由(3.3)式"="最右侧中两项的正交性可知

$$1=\|a\|^2=\|(X')^{+}Jb\|^2+\|(I_p-P_X\beta)\|^2. \tag{3.5}$$

下面用反证法证明$a \in \Re(X)$. 若$a \notin \Re(X)$, 则存在β使得$(I_p - P_X)\beta \neq 0$. 记

$$f = \|(I_p - P_X)\beta\|^2 = 1 - \|(X')^+ Jb\|^2.$$

由$a'a = 1$可知$0 < f \leq 1$. 当$f = 1$时, 由(3.5)可知$a = (I_p - P_X)\beta$, 代入到(3.4)得$a'\mathrm{SSB}a = 0$, 这样的a不能使$a'\mathrm{SSB}a$达到最大值, 故$0 < f < 1$,令

$$a(t) = \sqrt{\frac{1 - t^2 f}{1 - f}} \cdot (X')^+ Jb + t \cdot (I_p - P_X)\beta,$$

那么$a(t)'a(t) = 1$, 且当$t \in [0, 1)$时有

$$a(t)'\mathrm{SSB}a(t) = \frac{1 - t^2 f}{1 - f} b'J'(H - C)Jb > a'\mathrm{SSB}a. \tag{3.6}$$

这与假设相矛盾, 因此要使$a'\mathrm{SSB}a$取得最大值必有$(I_p - P_X)\beta = 0$, 即$a \in \Re(X)$.

定理3.1 在定义3.1的条件下, 假设检验问题(1.4)的广义似然比检验统计量为**:**

$$Z(X) = (2\pi)^{-\frac{n}{2}} [T(X)]^{-\frac{n}{2}} \exp\{-\frac{n}{2}\}.$$

其中$T(X) = \lambda_1[F'[J'(X'X)^{-1}J]^{-1}F]$, F是列满秩矩阵, 且有$J'(H - C)J = FF'$.

由于$Z(X)$是$T(X)$的单调函数, 故可用$T(X)$作为检验统计量, $T(X)$越大越倾向于拒绝H_0.

证明 由引理3.1可知$a \in \Re(X)$, 即(3.3)中的$\beta = 0$, 则有

$$a = (X')^+ Jb. \tag{3.7}$$

由于$X'X$是满秩的, 即X'是行满秩矩阵, X'的Moore-Penrose广义逆$(X')^+ = (X^+)' = X(X'X)^{-1}$, 将其代入到(3.7)中有

$$a = X(X'X)^{-1}Jb. \tag{3.8}$$

且由$a'a = 1$, 可得到

$$a'a = b'J'(X'X)^{-1}X'X(X'X)^{-1}Jb = b'J'(X'X)^{-1}Jb = 1.$$

令$c = [J'(X'X)^{-1}J]^{\frac{1}{2}}b$, 则$b = [J'(X'X)^{-1}J]^{-\frac{1}{2}}c$,且$c'c = 1$, 有

$$\begin{aligned} a'\mathrm{SSB}a &= a'X(H - C)X'a = b'J'(H - C)Jb \\ &= c'[J'(X'X)^{-1}J]^{-\frac{1}{2}} J'(H - C)J[J'(X'X)^{-1}J]^{-\frac{1}{2}} c. \end{aligned} \tag{3.9}$$

记$J'(H - C)J = FF'$, F为$k \times (k - 1)$的列满秩矩阵, 则(3.9)可以写成

$$\begin{aligned} a'\mathrm{SSB}a &= c'[J'(X'X)^{-1}J]^{-\frac{1}{2}} FF' [J'(X'X)^{-1}J]^{-\frac{1}{2}} c \\ &\leq \lambda_1[[J'(X'X)^{-1}J]^{-\frac{1}{2}} FF' [J'(X'X)^{-1}J]^{-\frac{1}{2}}] = \lambda_1[F'[J'(X'X)^{-1}J]^{-1}F]. \end{aligned} \tag{3.10}$$

记$A = [J'(X'X)^{-1}J]^{-\frac{1}{2}} FF' [J'(X'X)^{-1}J]^{-\frac{1}{2}}$, 当$c$为$A$的属于特征值$\lambda_1(A)$的标准特征向量$\varphi$时, (3.10)式中的不等号可取到等号, 即$a^* = (X')^+ J[J'(X'X)^{-1}J]^{-\frac{1}{2}}\varphi$时$a^{*'}\mathrm{SSB}a^* = \lambda_1(A)$, 则

$$\max_{a'a=1, a'\mathrm{SSW}a=0} a'\mathrm{SSB}a = a^{*'}\mathrm{SSB}a^* = \lambda_1[F'[J'(X'X)^{-1}J]^{-1}F].$$

令$T(X)=\lambda_1[F'[J'(X'X)^{-1}J]^{-1}F$, 可得到

$$Z(X)=(2\pi)^{-\frac{n}{2}}[T(X)]^{-\frac{n}{2}}\exp\{-\frac{n}{2}\}.$$

由于$Z(X)$是$T(X)$的单调函数, 故直接用$T(X)$作为检验统计量即可. 特别地, 当$k=2$时, F 可取作

$$F_2=\sqrt{\frac{n_1n_2}{n}}\begin{bmatrix}1\\-1\end{bmatrix},$$

则有

$$T(X)=\frac{n_1n_2}{n}\begin{bmatrix}1&-1\end{bmatrix}[J'(X'X)^{-1}J]^{-1}\begin{bmatrix}1\\-1\end{bmatrix}.$$

这与Zhao[12]所提出的检验统计量只相差了一个常数$\frac{n_1n_2}{n}$, 说明这两个统计量没有本质的差别; 当$k=3$时,

$$J'(H-C)J=\frac{1}{n}\begin{bmatrix}n_1(n-n_1)&-n_1n_2&-n_1n_3\\-n_2n_1&n_2(n-n_2)&-n_2n_3\\-n_3n_1&-n_3n_2&n_3(n-n_3)\end{bmatrix},$$

此时可取

$$F'=F_3'=\frac{1}{\sqrt{n}}\begin{bmatrix}\sqrt{n_1(n-n_1)}&-\frac{n_2\sqrt{n_1}}{\sqrt{n-n_1}}&-\frac{n_2\sqrt{n_1}}{\sqrt{n-n_1}}\\0&-\sqrt{\frac{nn_2n_3}{n-n_1}}&\sqrt{\frac{nn_2n_3}{n-n_1}}\end{bmatrix}.$$

3.2 数值模拟

针对本章节所提出的检验方法, 我们运用R语言对$k=3$的情况进行数据模拟, 取显著水平$\alpha=0.05$, 运用蒙特卡洛方法模拟出实际的检验水平α_0, 具体步骤如下：

(1)分别在总体$\mathrm{X}_1,\mathrm{X}_2,\mathrm{X}_3$中抽取样本, 得到样本阵$X$, 计算出$T=\lambda_1[F'[J'(X'X)^{-1}J]^{-1}F]$;

(2)对样本阵X的分量进行随机置换, 得到X_b, 计算出$T_b=\lambda_1[F'[J'(X_b'X_b)^{-1}J]^{-1}F]$;

(3)将步骤(2)重复1000次, 得到$T_b,b=1,2,\cdots,1000$, 并计算检验的p值: $p=\frac{1}{1000}\sum\limits_{b=1}^{1000}I_{\{T_b\geq T\}}$, 当$p\leq\alpha$ 时拒绝原假设;

(4)将步骤(1)-(3)重复2000次, 记第j次重复时得到的p值为p_j, 用$\alpha_0=\frac{1}{2000}\sum\limits_{j=1}^{2000}I_{\{p_j\leq 0.05\}}$来估计实际的检验水平.

为了检验本文所提出的新统计量的优良性, 我们考虑以下四个不同的协方差矩阵, 以便观察变量间相关性大小对检验水平的影响.

$$\begin{aligned}&\Sigma_1=I_p;\\&\Sigma_2=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.25,i\neq j;\\&\Sigma_3=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.5,i\neq j;\\&\Sigma_4=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.75,i\neq j.\end{aligned}$$

表4给出了多元正态分布下$k = 3$时新统计量$Z(X)$的检验水平, 不失一般性, 令$\mu_i = 0$, $i = 1, 2, 3$, 由表4可知新统计量的检验水平能很好地控制在显著水平0.05左右, 这说明我们提出的统计量能很好的控制检验水平, 具有优良性.

表 4: 多元正态分布下k=3时Z(X)的检验水平

(p, n_1, n_2, n_3)	Σ_1	Σ_2	Σ_3	Σ_4
(8,2,2,3)	0.05	0.0465	0.05	0.051
(14,3,4,5)	0.0435	0.052	0.047	0.047
(20,3,5,5,5)	0.0495	0.052	0.0525	0.049
(25,6,7,8)	0.052	0.049	0.0515	0.0485
(37,9,13,9)	0.0525	0.0435	0.0495	0.0495
(50,13,10,11)	0.0485	0.053	0.0515	0.0525
(100,25,15,30)	0.051	0.0515	0.049	0.0515
(100,25,25,25)	0.0475	0.05	0.046	0.049
(150,26,35,42)	0.05	0.046	0.044	0.0465

3.3 小结

本章节我们在似然比的基础上, 对高维数据下的方差分析检验问题提出了一种新的广义似然比统计量$Z(X)$, 并给出了该统计量的一些性质定理及其证明, 并通过数据模拟可知该检验具有一定的优良性. 并交原则的使用将多元方差分析问题转化成较为简单的一元方差分析问题, 在数据模拟部分, 由于本章节对新检验的渐近分布情况没有进行讨论, 在计算实际检验水平的过程中采用了蒙特卡洛的方法计算出检验统计量的p值, 从而弥补了没有讨论渐近分布的不足.

第 4 章　高维数据下关于正态均值线性约束的检验

本章我们讨论高维数据下总体均值受约束的检验问题(1.5). 在实际的检验问题中, 常常需要根据一定的要求对参数施加一定的约束条件, 例如特殊的球性检验、均值向量的子向量等于某已知向量的检验等等, 在低维情形下, 此类问题均可以用经典的检验理论进行分析, 本节我们以Zhao结合并交原则为基础, 针对高维数据下的线性约束问题提出相应的广义似然比检验.

4.1 高维数据下正态均值受线性约束的似然比检验

下面考虑假设检验问题(1.5). 为方便起见, 记$X = (X_1, X_2, \cdots, X_n)$为样本阵, 根据并交原则, 假设检验问题(1.5)中的原假设H_0可以写成如下检验问题

$$H_{0a} : a'R\mu = a'r \leftrightarrow H_{1a} : a'R\mu \neq a'r. \tag{4.1}$$

中的原假设H_{0a}之交, 即$H_0 = \bigcap_{a'a=1} H_{0a}$, 其中$\forall a \in R^q$且$a'a = 1$. 若拒绝某一个$H_{0a}$, 则应拒绝$H_0$. 所有的$H_{0a}$不被拒绝则不拒绝$H_0$. 因此我们的任务就是在所有的$H_{0a}$中找到最容易被拒绝的$H_{0a^*}$, 若$H_{0a^*}$被拒绝, 则$H_0$应被拒绝; 若$H_{0a^*}$不被拒绝, 则不拒绝$H_0$.

容易得知$a'RX_1, a'RX_2, \cdots, a'RX_n$独立同分布于$N(a'R\mu, a'R\Sigma R'a)$, 对给定的$a$, 假设检验问题(4.1)的似然比检验统计量为:

$$\lambda_a = \frac{L_{HR}(a)}{L_R(a)}.$$

其中

$$\begin{aligned} L_{HR}(a) &= \sup_{R\Sigma R'} L(a'r, a'R\Sigma R'a) = L(a'r, \frac{n-1}{n}a'RSR'a + a'(R\overline{X} - r)(R\overline{X} - r)'a) \\ &= (2\pi)^{-\frac{n}{2}}(\frac{n-1}{n}a'RSR'a + a'(R\overline{X} - r)(R\overline{X} - r)'a)^{-\frac{n}{2}} \cdot \exp\{-\frac{n}{2}\}, \\ L_R(a) &= \sup_{R\mu, R\Sigma R'} L(a'R\mu, a'R\Sigma R'a) = L(a'R\overline{X}, a'RSR'a) \\ &= (2\pi)^{-\frac{n}{2}}(\frac{n-1}{n}a'RSR'a)^{-\frac{n}{2}} \cdot \exp\{-\frac{n}{2}\}. \end{aligned}$$

结合并交原则可知, 较小的λ_a对于拒绝H_{0a}提供了有力的保障, 因此我们要找到一个a^*, 使得λ_{a^*} 在所有λ_a中是最小的. 即

$$\lambda_{a^*} = \min \lambda_a = \min \frac{(\frac{n-1}{n}a'RSR'a)^{\frac{n}{2}}}{[\frac{n-1}{n}a'RSR'a + a'(R\overline{X} - r)(R\overline{X} - r)'a]^{\frac{n}{2}}}. \tag{4.2}$$

当$p > n$时, 样本协方差矩阵S是奇异的, 由于此时矩阵R是行满秩的, 因此矩阵是RSR'否为奇异矩阵尚不能确定, 接下来我们对RSR'是否为奇异矩阵进行分类讨论.

4.1.1 RSR'是奇异矩阵时

若RSR'是奇异的, 则存在一些a使得$a'RSR'a = 0$, 即$L_R(a) = \infty$, 若$a'(R\overline{X} - r) \neq 0$, 会导致$\lambda_a = 0$, 此时$\lambda_a = \frac{L_{HR}(a)}{L_R(a)}$不能再作为检验$H_{0a}$的统计量, 因此我们考虑当$L_R(a) = \infty$时, $L_{HR}(a)$越小越倾向于拒绝H_{0a}. 由于H_{0a^*}是H_{0a}中最容易被拒绝的, 则应找到一个a^*, 令$a^* \in A = \{a|a \in R^q, a'a = 1, L_R(a) = \infty\}$, 并使得$L_H(a)$达到最小, 此时$H_{0a^*}$的似然比统计量为$L_{HR}(a^*) = \min\limits_{a\in A} L_{HR}(a)$, 并将其作为$H_0 = \bigcap\limits_{a'a=1} H_{0a}$的检验统计量进行讨论.

定义4.1 对于假设检验问题(1.5), 若RSR' 为奇异矩阵且$p > q \geq n$时, 则$H_0 = \bigcap\limits_{a'a=1} H_{0a}$的广义似然比统计量为:

$$F(X) = L_H(a^*) = \min_{L(a)=\infty} L_H(a),$$

其中$a \in R^q$且$a'a = 1$. 当$F(X)$足够小时, 我们拒绝$H_0 : R\mu = r$. 不失一般性, 对数据稍作变换, 假设问题(1.5)可写为

$$H_0 : R\mu = \mathbf{0} \leftrightarrow H_1 : R\mu \neq \mathbf{0} \tag{4.3}$$

即$r = \mathbf{0}$的情形.

定理4.1 在定义4.1条件下, $H_0 : R\mu = \mathbf{0}$的广义似然比统计量为

$$F(X) = \frac{1}{(2\pi)^{\frac{n}{2}}[U(X)]^{\frac{n}{2}}} \cdot \exp\{-\frac{n}{2}\},$$

其中$U(X) = \frac{1}{1_n'(X'R'RX)^{-1}1_n}$.

由于$F(X)$是$U(X)$的单调函数, 故可用$U(X)$作为检验统计量, $U(X)$越大越倾向于拒绝$H_0 : R\mu = \mathbf{0}$.

证明 证明方法类似于文献 [11]中定理1的证明方法.

引理4.1 设$a^* = \arg\max\limits_{a'a=1, a'RSR'a=0} a'R\overline{X}\ \overline{X}'R'a$, 则$a^*$一定在$\Re(RX)$中.

证明 由于

$$\begin{aligned} & a'RSR'a = \frac{1}{n-1}a'RX(I_n - \frac{1}{n}1_n1_n'X'R'a) = 0 \\ \Leftrightarrow & (I_n - \frac{1}{n}1_n1_n')'X'R'a = 0, \end{aligned}$$

则$X'R'a$属于$(I_n - \frac{1}{n}1_n1_n')$列向量张成的正交补空间, 即$X'R'a \in \Re^{\perp}(I_n - \frac{1}{n}1_n1_n')$. 又由于$(I_n - \frac{1}{n}1_n1_n') \cdot 1_n = 0$, 则有$c \in \Re(1_n)$, 即$X'R'a = c \cdot 1_n$($c$为任意常数). 由相容线性方程组通解 [24]可得

$$a = c(X'R')^+ \cdot 1_n + [I_q - (X'R')^+X'R']\beta, \text{其中}\beta\text{是任意向量.} \tag{4.4}$$

由于$X'R'RX$是满秩的, $X'R'$ 的“$+$”广义逆为$(RX)(X'R'RX)^{-1}$, 代入(4.4)可得：

$$a = c(RX)(X'R'RX)^{-1} \cdot 1_n + (I_q - P_{RX})\beta. \tag{4.5}$$

将(4.5)代入到$a'a=1$可推出

$$c^2=\frac{1-\beta'(I_q-P_{RX})\beta}{1_n'(X'R'RX)^{-1}1_n}.$$

且有

$$\begin{aligned}a'R\overline{X}\ \overline{X}'R'a=&\frac{1}{n^2}\{[\beta'(I_q-P_{RX})+c1_n'(X'R'RX)^{-1}(RX)']RX1_n1_n'X'R'\cdot\\&[c(RX)(X'R'RX)^{-1}1_n+(I_q-P_{RX})\beta]\}=c^2.\end{aligned}$$

下面运用反证法证明$a\in\Re(RX)$. 若$a\notin\Re(RX)$, 则(4.5)中的$(I_q-P_{RX})\beta\neq 0$, 即$\beta'(I_q-P_{RX})\beta>0$. 令

$$a^{(1)'}R\overline{X}\ \overline{X}'R'a^{(1)}=\frac{1}{1_n'(X'R'RX)^{-1}1_n},$$

则此时有$a^{(1)'}R\overline{X}\ \overline{X}'R'a^{(1)}>a'R\overline{X}\ \overline{X}'R'a$, 即当$a\notin\Re(RX)$时,$a'R\overline{X}\ \overline{X}'R'a$不是最大值, 假设不成立.

引理4.2 $a\in\Re(RX)$, $a'a=1$且$a'RSR'a=0\Leftrightarrow a=\pm\frac{1}{\sqrt{1_n'(X'R'RX)^{-1}1_n}}(RX)(X'R'RX)^{-1}1_n$.

证明 先证充分性. 由于$a\in\Re(RX)$, 则$(I_q-P_{RX})\beta=0$, 代入(4.5)有

$$a=c(RX)(X'R'RX)^{-1}\cdot 1_n. \tag{4.6}$$

由$a'a=1$可推出

$$c=\pm\frac{1}{\sqrt{1_n'(X'R'RX)^{-1}1_n}}. \tag{4.7}$$

将(4.7)代入到(4.6)中, 可证$a=\pm\frac{1}{\sqrt{1_n'(X'R'RX)^{-1}1_n}}(RX)(X'R'RX)^{-1}1_n$.

再证必要性. 实际上, 只需要将a的值直接代入到$a'a=1$和$a'RSR'a=0$ 中即可.

下面介绍分块矩阵的逆矩阵求法:

引理[24]**4.3** 设

$$A=\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}$$

可逆. 若$|A_{22}|\neq 0$, 则

$$A^{-1}=\begin{bmatrix}A_{11.2}^{-1}&-A_{11.2}^{-1}A_{12}A_{22}^{-1}\\-A_{22}^{-1}A_{21}A_{11.2}^{-1}&A_{22}^{-1}+A_{22}^{-1}A_{21}A_{11.2}^{-1}A_{12}A_{22}^{-1}\end{bmatrix},$$

其中$A_{11.2}=A_{11}-A_{12}A_{22}^{-1}A_{21}$.

定理4.2 设$p>q\geq n$, $P_{RSR'}$为RSR'列空间的正交投影, 则有

$$T(X)=\frac{1}{1_n'(X'R'RX)^{-1}1_n}=\|(I-P_{RSR'}R\overline{X})\|^2=\overline{X}'R'(I-P_{RSR'})R\overline{X}.$$

证明 由于$\|(I-P_{RSR'}R\overline{X})\|^2=\overline{X}'R'(I-P_{RSR'})R\overline{X}$ 容易得到, 因此我们只需要证明

$$\frac{1}{1_n'(X'R'RX)^{-1}1_n}=\overline{X}'R'(I-P_{RSR'})R\overline{X}. \tag{4.8}$$

成立即可, 又因为(4.8)等价于

$$\frac{1_n'(X'R'RX)^{-1}1_n}{n}=\frac{1}{n\overline{X}'R'(I-P_{RSR'})R\overline{X}}, \tag{4.9}$$

所以我们只需验证(4.9)成立. 我们令$Q=\begin{bmatrix}\frac{1_n'}{\sqrt{n}}\\ Q_1\end{bmatrix}$, 且易知$Q$是正交矩阵, 则$Q^{-1}=Q'$, 有

$$Q'Q=\begin{bmatrix}\frac{1_n}{\sqrt{n}} & Q_1'\end{bmatrix}\begin{bmatrix}\frac{1_n'}{\sqrt{n}}\\ Q_1\end{bmatrix}=\frac{1_n1_n'}{n}+Q_1'Q_1=I,$$

及

$$Q(X'R'RX)^{-1}Q'=\begin{bmatrix}\frac{1_n'(X'R'RX)^{-1}1_n}{n} & \frac{1_n'(X'R'RX)^{-1}Q_1'}{\sqrt{n}}\\ \frac{Q_1(X'R'RX)^{-1}1_n}{\sqrt{n}} & Q_1(X'R'RX)^{-1}Q_1'\end{bmatrix}$$

由于$Q(X'R'RX)^{-1}Q'=[(Q(X'R'RX)^{-1}Q')^{-1}]^{-1}=[Q(X'R'RX)Q']^{-1}$, 则有

$$\begin{aligned}[Q(X'R'RX)Q']^{-1}&=\begin{bmatrix}\frac{1_n'(X'R'RX)1_n}{n} & \frac{1_n'(X'R'RX)Q_1'}{\sqrt{n}}\\ \frac{Q_1(X'R'RX)1_n}{\sqrt{n}} & Q_1(X'R'RX)Q_1'\end{bmatrix}^{-1}\\&=\begin{bmatrix}\frac{1_n'(X'R'RX)^{-1}1_n}{n} & \frac{1_n'(X'R'RX)^{-1}Q_1'}{\sqrt{n}}\\ \frac{Q_1(X'R'RX)^{-1}1_n}{\sqrt{n}} & Q_1(X'R'RX)^{-1}Q_1'\end{bmatrix}.\end{aligned}$$

运用引理4.3分块矩阵求逆公式可证得(4.9)式, 又因为$RSR'=\frac{1}{n-1}RXQ_1'Q_1X'R'$, 即$\Re(RSR')=\Re(RXQ_1'Q_1X'R')=\Re(RXQ_1')$, 则有$P_{RSR'}=P_{RXQ_1'}$, 证毕.

4.1.2 RSR'是非奇异矩阵时

定理4.3 对于假设检验问题(1.5), 若RSR'为非奇异矩阵且$q<n-1<p$时, 则$H_0=\bigcap\limits_{a'a=1}H_{0a}$的广义似然比统计量服从$F$分布:

$$D\sim\frac{q}{n-q}F(q,n-q),$$

其中$D=\frac{n}{n-1}\cdot(R\overline{X}-r)'(RSR')^{-1}(R\overline{X}-r)$, 对于给定显著水平$\alpha$, 当$\frac{n-q}{q}\cdot D\geq F_\alpha(q,n-q)$时, 拒绝原假设.

证明 由(4.2)知

$$\begin{aligned}\lambda_{a^*}=\min\lambda_a=&\min\frac{(\frac{n-1}{n}a'RSR'a)^{\frac{n}{2}}}{[\frac{n-1}{n}a'RSR'a+a'(R\overline{X}-r)(R\overline{X}-r)'a]^{\frac{n}{2}}}\\ \Leftrightarrow&\max\frac{n}{n-1}\cdot\frac{a'(R\overline{X}-r)(R\overline{X}-r)'a}{a'RSR'a}.\end{aligned}$$

将$\max\frac{n}{n-1}\cdot\frac{a'(R\overline{X}-r)(R\overline{X}-r)'a}{a'RSR'a}$ 定义为D, 令$b=(RSR')^{\frac{1}{2}}\cdot a$, 则有

$$\begin{aligned}D=&\max\frac{n}{n-1}\cdot\frac{b'(RSR')^{-\frac{1}{2}}(R\overline{X}-r)(R\overline{X}-r)'(RSR')^{-\frac{1}{2}}b}{b'b}\\=&\frac{n}{n-1}\cdot(R\overline{X}-r)'(RSR')^{-1}(R\overline{X}-r).\end{aligned}$$

由于$R\overline{X}-r$与$\frac{n-1}{n}RSR'$相互独立, 则有$(n-1)(R\overline{X}-r)'(\frac{n-1}{n}RSR')^{-1}(R\overline{X}-r)\sim T^2(q,n-1)$, 由文献[24]可知$D\sim\frac{q}{n-q}F(q,n-q)$, 证毕.

4.2 数值模拟

Bai和Saranadasa[10]对于假设检验问题(1.1)提出了一种方法, 考虑RSR'为奇异矩阵时, 我们针对问题(1.5)将Bai和Saranadasa的方法进行相应的推导并得到如下检验统计量:

$$T_{BSC}=\frac{n(R\overline{X})'(R\overline{X})-\mathrm{tr}(RSR')}{[\frac{2(n-1)n}{(n-2)(n+1)}\cdot(\mathrm{tr}(RSR')^2-\frac{(\mathrm{tr}(RSR')^2}{n-1})]^{\frac{1}{2}}},$$

且T_{BSC}的拒绝域为:

$$C_{BSC}=\{X:\frac{n(R\overline{X})'(R\overline{X})-\mathrm{tr}(RSR')}{[\frac{2(n-1)n}{(n-2)(n+1)}\cdot(\mathrm{tr}(RSR')^2-\frac{(\mathrm{tr}(RSR')^2}{n-1})]^{\frac{1}{2}}}\geq z_\alpha\},$$

其中z_α为标准正态α的分位点.

对于本文提出的RSR'为奇异矩阵情况下的新统计量F_{new},为了研究新统计量F_{new}在不同协方差结构的情形下的表现, 我们采用蒙特卡洛方法模拟出实际的检验的水平α_0, 具体步骤如下:

(1)在总体中抽取样本, 得到样本阵X, 计算出$F=\frac{1}{1_n'(X'R'RX)^{-1}1_n}$;

(2)令$X_b=(X_{b_1},X_{b_2},\cdots,X_{b_n})$, 其中$X_{b_i}=(-1)^{\eta_i}X_i$, $\eta_1,\eta_2,\cdots,\eta_n\sim B(\frac{1}{2},1)$, 并计算出$F_b=\frac{1}{1_n'(X_b'R'RX_b)^{-1}1_n}$;

(3)将步骤(2)重复1000次, 得到$\{F_b,b=1,2,\cdots,1000\}$, 并计算出$p$值: $p=\frac{1}{1000}\sum\limits_{b=1}^{1000}I_{\{F_b\geq T\}}$, 当$p\leq\alpha$ 时拒绝原假设;

(4)将步骤(1)-(3)重复2000次, 记第j次重复得到的p值为p_j, 用$\alpha_0=\frac{1}{2000}\sum\limits_{j=1}^{2000}I_{\{p_j\leq 0.05\}}$来估计实际的检验水平.

为了比较T_{BSC}和F_{new}这两种检验统计的优良性, 我们考虑以下四个不同的协方差矩阵, 以便观察变量间相关性大小对于这两种检验水平的影响.

$$\begin{aligned}
&W_1=I_p;\\
&W_2=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.25,i\neq j;\\
&W_3=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.5,i\neq j;\\
&W_4=(\rho_{ii}):\rho_{ii}=1,\rho_{ij}=0.75,i\neq j.
\end{aligned}$$

表5给出了在多元正态分布下F_{new}和T_{BSC}的检验水平, 其中显著水平$\alpha=0.05$, 均进行2000次实验. 由表1可知, 新检验F_{new}的水平能更好的维持在显著水平0.05左右, 特别是在变量间的相关性变强时, 更能体现出新检验的优良性.

表 5: 多元正态分布下两种检验的水平比较

		$n=10$		$n=20$		$n=40$		$n=50$	
		T_{BSC}	F_{new}	T_{BSC}	F_{new}	T_{BSC}	F_{new}	T_{BSC}	F_{new}
q=60	W_1	0.0705	0.0525	0.069	0.0605	0.0595	0.0495	0.054	0.0595
	W_2	0.0885	0.0545	0.083	0.058	0.067	0.0625	0.0745	0.062
	W_3	0.099	0.046	0.0835	0.061	0.079	0.058	0.0785	0.053
	W_4	0.1075	0.0555	0.088	0.061	0.073	0.055	0.0735	0.0565
q=100	W_1	0.0585	0.056	0.061	0.0585	0.057	0.047	0.0565	0.053
	W_2	0.1025	0.0585	0.0815	0.0605	0.0735	0.0585	0.065	0.0575
	W_3	0.102	0.053	0.0815	0.0485	0.0775	0.0615	0.0825	0.06
	W_4	0.119	0.0545	0.0815	0.0585	0.0835	0.061	0.07	0.051
q=150	W_1	0.063	0.057	0.0565	0.059	0.06	0.061	0.063	0.052
	W_2	0.11	0.057	0.0795	0.0585	0.0655	0.051	0.07	0.0565
	W_3	0.1095	0.0495	0.085	0.0595	0.0785	0.0545	0.078	0.0495
	W_4	0.1045	0.0555	0.0865	0.0565	0.081	0.058	0.075	0.058
q=150	W_1	0.0575	0.059	0.061	0.051	0.053	0.0615	0.056	0.056
	W_2	0.0955	0.055	0.082	0.0535	0.0785	0.0625	0.0825	0.0545
	W_3	0.107	0.051	0.0915	0.061	0.0725	0.0595	0.0778	0.0575
	W_4	0.106	0.0525	0.0835	0.0635	0.0755	0.058	0.076	0.053

4.3 小结

本章节在高维数据正态总体的条件下, 结合并交原则对均值线性约束检验问题提出了一种新的广义似然比统计量F_{new}, 并给出了该统计量的一些相关性质, 并通过数据模拟可知该检验与Bai和Saranadasa所提出检验相比具有优良性.

第 5 章　高维数据下随机变量的独立性检验

本章我们讨论高维数据下随机变量的独立性检验问题, 依然沿用并交原则, 对该问题提出了一个新的检验统计量, 并给出了相关定理及证明, 其中要特别说明地是, Zhao[12] 曾解释道, 在高维背景下, 存在某些a使得$a'Sa = 0$, 及$\sup L(a'\mu, a'\Sigma a) = \infty$的原因是相对于维数来说, 样本容量过小导致无法对于每个a都提供有限的似然. 在本章中, 我们不再检验总体的期望, 而是检验协方差, 因此以$a'S_{11}a \neq 0, b'S_{22}b \neq 0$为条件对该检验问题进行研究.

5.1 高维数据下随机变量独立性问题的似然比检验

引理[40]5.1　对于假设检验问题(1.6), 当$p < n$时, 其似然比检验统计量为:

$$\lambda = \frac{|A|^{\frac{n}{2}}}{(\prod_{i=1}^{q} |A_{ii}|)^{\frac{n}{2}}}, \tag{5.1}$$

其中$A = XH_nX' = (n-1)S_{(1)}, A_{ii}$为$A$的第$(i,i)$子块, λ越小越容易拒绝原假设.

引理5.1的证明可见Muirhead[40]中定理11.2.1.

引理[41]5.2　检验统计量(5.1)具有如下渐进分布:

$$-s\log V \xrightarrow{L} \chi^2_m,$$

其中$V = \lambda^{\frac{n}{2}} = \frac{|A|}{\prod_{i=1}^{q}|A_{ii}|}, m = \frac{1}{2}(p^2 - \sum_{i=1}^{q} p_i^2), s = n - 3 - (p^3 - \sum_{i=1}^{q} p_i^3)/[3(p^2 - \sum_{i=1}^{q} p_i^2)]$, 符号$\xrightarrow{L}$表示依分布收敛.

当$p > n$时, 问题转化成高维数据下随机变量的独立性检验问题, 此时由于A不可逆, 上述引理的结论不再成立. 不失一般性地, 可讨论$q = 2$的情形. 设$X_1, X_2, \cdots, X_n$为简单随机样本, $X = (X_1, X_2, \cdots, X_n)$为样本阵, 将第$k$列向量$X_k$分块为$\begin{bmatrix} Y_k \\ Z_k \end{bmatrix}$, 其中$Y_k$为$p_1$维列向量, Z_k为p_2维列向量, $p_1 + p_2 = p$, $k = 1, 2, \cdots, n$.相应的样本阵、均值向量、总体协方差矩阵和样本协方差矩阵也可分块为:

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_Y \\ \mu_Z \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

其中μ_Y为p_1维列向量, μ_Z为p_2维列向量, Σ_{12}为$p_1 \times p_2$矩阵. 将均值、总体样本协方差作如下表达:

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k, \quad \overline{Z} = \frac{1}{n}\sum_{k=1}^{n} Z_k,$$

$$S_{12} = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \overline{Y})(Z_k - \overline{Z})' = \frac{1}{n-1}YH_nZ' = \frac{1}{n-1}YZ' - \frac{n}{n-1}\overline{Y}\overline{Z}'.$$

此时假设检验问题(1.6)可写成:

$$H_0 : \Sigma_{12} = 0 \leftrightarrow H_1 : \Sigma_{12} \neq 0. \tag{5.2}$$

对$\forall a \in R^{p_1}$和$\forall b \in R^{p_2}$,且$a \neq 0, b \neq 0$, 有

$$\begin{pmatrix} a'Y_k \\ b'Z_k \end{pmatrix} \sim N\left[\begin{pmatrix} a'\mu_Y \\ b'\mu_Z \end{pmatrix}, \begin{pmatrix} a'\Sigma_{11}a & a'\Sigma_{12}b \\ b'\Sigma_{21}a & b'\Sigma_{22}b \end{pmatrix}\right].$$

由并交原则可知, 假设检验问题(5.2)中的原假设可以写成如下假设检验问题

$$H_{0ab} : a'\Sigma_{12}b = 0 \leftrightarrow H_{1ab} : a'\Sigma_{12}b \neq 0. \tag{5.3}$$

中原假设H_{0ab}之交, 即$H_0 = \bigcap\limits_{a\neq 0, b\neq 0} H_{0ab}$, 若拒绝某一个$H_{0ab}$, 则应拒绝$H_0$, 所有的$H_{0ab}$不被拒绝则不拒绝$H_0$. 因此我们的任务就是在所有的$H_{0ab}$中找到最容易被拒绝的$H_{0ab^*}$, 若$H_{0a^*b^*}$被拒绝, 则$H_0$应被拒绝; 若$H_{0a^*b^*}$不被拒绝, 则不拒绝$H_0$.

容易知道, 当$n > p$时, 假设检验问题(5.3)的似然比统计量为

$$\lambda_{a,b} = \frac{L_H(a,b)}{L(a,b)},$$

其中

$$\begin{aligned} L_H(a,b) = &\sup_{a'\Sigma_{12}b=0} L\left[\begin{pmatrix} a'\mu_Y \\ b'\mu_Z \end{pmatrix}, \begin{pmatrix} a'\Sigma_{11}a & a'\Sigma_{12}b \\ b'\Sigma_{21}a & b'\Sigma_{22}b \end{pmatrix}\right] \\ =&L_1(a'\overline{Y}, a'S_{11}a) \cdot L_2(b'\overline{Z}, b'S_{22}b) = (2\pi)^{-n}(a'S_{11}a)^{-\frac{n}{2}}(b'S_{22}b)^{-\frac{n}{2}}\mathrm{e}^{-n}. \end{aligned}$$

$$\begin{aligned} L(a,b) = &\sup_{a'\Sigma_{12}b\neq 0} L\left[\begin{pmatrix} a'\mu_Y \\ b'\mu_Z \end{pmatrix}, \begin{pmatrix} a'\Sigma_{11}a & a'\Sigma_{12}b \\ b'\Sigma_{21}a & b'\Sigma_{22}b \end{pmatrix}\right] \\ =&(2\pi)^{-n}\left|\begin{matrix} a'S_{11}a & a'S_{12}b \\ b'S_{21}a & b'S_{22}b \end{matrix}\right|^{-\frac{n}{2}}\mathrm{e}^{-n} = (2\pi)^{-n}(a'S_{11}ab'S_{22}b - b'S_{21}aa'S_{12}b)^{-\frac{n}{2}}\mathrm{e}^{-n}. \end{aligned}$$

结合并交原则可知, 较小的$\lambda_{a^*b^*}$对于拒绝H_{0ab}提供了有力的保障, 因此我们需要找到一对a^*, b^*使得$\lambda_{a^*b^*}$在所有的λ_{ab}中是最小的, 即

$$\begin{aligned} \lambda_{a^*b^*} = \min \lambda_{ab} = \min \frac{L_H(ab)}{L(ab)} &= \min(\frac{a'S_{11}ab'S_{22}b}{a'S_{11}ab'S_{22}b - b'S_{21}aa'S_{12}b})^{-\frac{n}{2}} \\ \Leftrightarrow \max \frac{(a'S_{12}b)^2}{a'S_{11}ab'S_{22}b}&. \end{aligned} \tag{5.4}$$

当S_{11}, S_{22}为非奇异矩阵时, 由并交原则可知假设检验问题的检验统计量为

$$\lambda = \max_{a\neq 0, b\neq 0} \frac{(a'S_{12}b)^2}{a'S_{11}ab'S_{22}b} = \lambda_1(S_{21}S_{11}^{-1}S_{12}S_{22}^{-1}). \tag{5.5}$$

该表达式及证明可参见张尧庭等[42].

当$n < \min(p_1, p_2)$时, S_{11}, S_{22}为奇异矩阵时, 存在某些a, b使得$a'S_{11}a, b'S_{22}b$为0, 它们不能对检验问题(5.2)提供任何信息, 所以考虑$a'S_{11}a \neq 0, b'S_{22}b \neq 0$ 的情况.

定义5.1 当$n < \min(p_1, p_2)$且$a'S_{11}a \neq 0, b'S_{22}b \neq 0$时, 假设检验问题(5.2)的广义似然比检验统计量定义为:

$$\lambda(X) = \max_{a'S_{11}a\neq 0, b'S_{22}b\neq 0} \frac{(a'S_{12}b)^2}{a'S_{11}ab'S_{22}b}.$$

并且$\lambda(X)$越大越倾向于拒绝原假设.

引理5.3 设A是任意一个$n \times m$的实矩阵, 则

(1)$A^+ = (A'A)^+A' = A'(AA')^+$;

(2)$(A^+)'A^+ = (AA')^+$.

证明 结论(1)王桂松等[43]中的定理1.7.10.

结论(2)可根据结论(1)直接证明, 过程如下:

$$\begin{aligned}(A^+)'A^+ &= (A'(AA')^+)'(A'(AA')^+)\\ &= (AA')^+AA'(AA')^+ = (AA')^+.\end{aligned}$$

定理5.1 设A, B是两个$n \times n$的半正定矩阵, 且$\mathfrak{R}(A) \subset \mathfrak{R}(B), B \neq 0$,则有

$$\max_{\alpha'B\alpha\neq 0} \frac{\alpha'A\alpha}{\alpha'B\alpha} = \lambda_1(AB^+).$$

证明 设$\mathrm{rk}(B) = r \leq n$, 由于$B$是半正定矩阵, 故存在$r \times n$的列满秩矩阵$B_1$, 使得$B = B_1B_1^{'}$, 那么$\alpha'B\alpha = \alpha'B_1B_1^{'}\alpha$.令

$$a = B_1^{'}\alpha. \tag{5.6}$$

由于α是r维列向量且$B_1^{'}$为行满秩矩阵, 对于任意的$a, B_1^{'}\alpha = a$总有解, 且$a = B_1^{'}\alpha$是$R^n \to R^r$的满射, 即α取遍R^n中的向量时, a取遍R^r中的向量.

由(5.6)及线性相容方程组通解可知, 存在任意的r维向量c, 使得

$$\alpha = B_1^{'+}a + (I - B_1^{'+}B_1^{'})c.$$

又由于$\mathfrak{R}(A) \subset \mathfrak{R}(B)$且$A$、$B$均为半正定矩阵, 故存在矩阵$D$, 使

$$A = BD = B_1B_1^{'}D,$$

$$A = A' = (BD)' = D'B' = D'B_1B_1^{'}.$$

则有

$$\begin{aligned}\alpha'A\alpha &= [a'B_1^{+} + c'(I - B_1^{'+}B_1^{'})]B_1B_1^{'}D\alpha\\ &= a'B_1^{+}B_1B_1^{'}D\alpha = a'B_1A\alpha\\ &= a'B_1D'B_1B_1^{'}[B_1^{'+}a + (I - B_1^{'+}B_1^{'})c]\\ &= a'B_1D'B_1B_1^{'}B_1^{+}a = a'B_1^{'+}AB_1^{'+}a,\end{aligned}$$

$$\alpha' B\alpha = \alpha' B_1 B_1^{'}\alpha = a'a.$$

因为$B \neq 0$, 则$\alpha' B\alpha \neq 0 \Leftrightarrow a'a = 0$, 结合引理5.3(2)可得：

$$\begin{aligned}\max_{\alpha' B\alpha \neq 0} \frac{\alpha' A\alpha}{\alpha' B\alpha} &= \max_{a\neq 0} \frac{a' B_1^{'+} A B_1^{'+} a}{a'a} \\ &= \lambda_1(B_1^{+} A B_1^{'+}) = \lambda_1(A B_1^{'+} B_1^{+}) = \lambda_1(AB^{+}).\end{aligned}$$

对于假设检验(5.2), 由于样本协方差矩阵$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$是半正定矩阵, 有$\mathfrak{R}(S_{12}) \subset \mathfrak{R}(S_{11})$, $\mathfrak{R}(S_{21}) \subset \mathfrak{R}(S_{22})$, $\mathfrak{R}(S_{21} b b' S_{21}) \subset \mathfrak{R}(S_{11})$,则有$\mathfrak{R}(S_{21} S_{11}^{+} S_{12}) \subset \mathfrak{R}(S_{22})$. 则运用上述定理5.1可得到如下定理.

定理5.2 当$n < \min(p_1, p_2)$且$a' S_{11} a \neq 0, b' S_{22} b \neq 0$时, 假设检验问题(5.2)的广义似然比检验统计量为：

$$\lambda(X) = \max_{a' S_{11} a \neq 0, b' S_{22} b \neq 0} \frac{(a' S_{12} b)^2}{a' S_{11} a b' S_{22} b} = \lambda_1(S_{21} S_{11}^{+} S_{12} S_{22}^{+}). \tag{5.7}$$

证明 定理5.1可知

$$\begin{aligned}\lambda(X) &= \max_{a' S_{11} a \neq 0, b' S_{22} b \neq 0} \frac{(a' S_{12} b)^2}{a' S_{11} a b' S_{22} b} = \max_{b' S_{22} b \neq 0} \Big(\max_{a' S_{11} a \neq 0} \frac{a' S_{12} b b' S_{21} a}{a' S_{11} a} \cdot \frac{1}{b' S_{22} b} \Big) \\ &= \max_{b' S_{22} b \neq 0} \frac{1}{b' S_{22} b} \cdot \lambda_1(S_{21} b b' S_{21} S_{11}^{+}) = \max_{b' S_{22} b \neq 0} \frac{b' S_{21} S_{11}^{+} S_{12} b}{b' S_{22} b} \\ &= \lambda_1(S_{21} S_{11}^{+} S_{12} S_{22}^{+}).\end{aligned}$$

由上述定理5.2可知, 当S_{11}, S_{22}都可逆时, $S_{11}^{+} = S_{11}^{-1}, S_{22}^{+} = S_{22}^{-1}$, 且$\lambda(X)$正好是在$n > p$时用并交原则得到的检验统计量(5.5) 式.

设Q是$n \times (n-1)$的矩阵, 使得$Q'Q = I_{n-1}, QQ' = H_n$, 令

$$V = XQ = [v_1, v_2, \cdots, v_{n-1}], \tag{5.8}$$

则$v_1, v_2, \cdots, v_{n-1}$独立同分布于$N_p(\mathbf{0}, \Sigma), VV' = XQQ'X' = XH_nX = (n-1)S \sim W_p(n-1, \Sigma)$. 将每个$v_r$划分为$\begin{bmatrix} v_{r1} \\ v_{r2} \end{bmatrix}$, v_{r1}, v_{r2}分别为p_1维和p_2维向量, $r = 1, 2, \cdots, n-1$, 那么$(n-1)S_{st} = \sum_{r=1}^{n-1} v_{rs} v_{rt}'$. 为方便起见, 可记$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$, 其中$V_1 = [v_{11}, v_{21}, \cdots, v_{(n-1)1}], V_2 = [v_{12}, v_{22}, \cdots, v_{(n-1)2}]$, 则有$(n-1)S_{st} = V_s V_t', s, t = 1, 2$, 则定理5.2 的统计量则可以写为

$$\begin{aligned}\lambda_1(S_{21} S_{11}^{+} S_{12} S_{22}^{+}) &= \lambda_1[V_2 V_1'(V_1 V_1')^{+} V_1 V_2'(V_2 V_2')^{+}] \\ &= \lambda_1[V_1'(V_1 V_1')^{+} V_1 V_2'(V_2 V_2')^{+} V_2] = \lambda_1(P_{V_1'} P_{V_2'}) = 1.\end{aligned} \tag{5.9}$$

其中V_1是$p_1 \times (n-1)$的列满秩矩阵, $P_{V_1'}$是$n-1$阶对称幂等阵, 因此$P_{V_1'} = I_{n-1}$. 同理可得$P_{V_2'} = I_{n-1}$. 如此可知统计量$\lambda_1(S_{21} S_{11}^{+} S_{12} S_{22}^{+})$总是等于1, 即它不是一个合适的检验统计量, 造成这

个结果的原因之一就是当$n < p$时, a, b的取值范围太大, 样本不能提供足够的信息, 为把假设检验问题(5.3)中所有的H_{0ab}做出合理推断, 我们将a, b 的取值范围缩小, 记

$$\mathcal{A} = \{e_i | e_i\text{是第}i\text{个分量为1,其余分量为0的}p_1\text{维向量}, i = 1, 2, \cdots, p_1\},$$

$$\mathcal{B} = \{e_j | e_j\text{是第}j\text{个分量为1,其余分量为0的}p_2\text{维向量}, j = 1, 2, \cdots, p_2\},$$

则对于假设检验问题(5.3), 可令$a \in \mathcal{A}, b \in \mathcal{B}$, 当$a = e_i, b = e_j$ 时, 假设检验问题(5.3) 可以写成如下形式

$$H_{0ij} : \sigma_{ij} = 0 \leftrightarrow H_{1ij} : \sigma_{ij} \neq 0, i < j. \tag{5.10}$$

即假设检验问题(5.2)的原假设可以写为$H_0 = \bigcap\limits_{i<j} H_{0ij}$, 对每一对$i < j$, H_{0ij}的似然比统计量为

$$\lambda_{ij} = \frac{(e_i' S_{12} e_j)^2}{e_i' S_{11} e_i e_j' S_{22} e_j} = \frac{{S_{ij}}^2}{S_{ii} S_{jj}} = r_{ij}^2.$$

其中r_{ij}是样本阵中第i, j行的样本相关系数. 当$\sigma_{ij} = 0$ 时,$\sqrt{n-2}\frac{r_{ij}}{\sqrt{1-r_{ij}^2}} \sim t(n-2)$, 故$(n-2)\frac{\lambda_{ij}}{1-\lambda_{ij}} \sim F(1, n-2)$, 该分布与未知参数无关(参见Muirhead[40]). 根据并交原则, 检验统计量可以取为

$$T_{new4} = (n-2) \max_{i<j} \frac{{S_{ij}}^2}{S_{ii} S_{jj}}. \tag{5.11}$$

注意到若$x = (x_1, x_2, \cdots, x_m)', y = (y_1, y_2, \cdots, y_m)' \in \mathbf{R}^m$, 则

$$\|x - y\|_\infty \triangleq \max_{1 \le i \le m} |x_i - y_i|$$

是向量x, y间的一种距离(参见孙炯等[45]), 从而可以看出T_{new4}是向量$(\frac{S_{12}^2}{S_{11}S_{22}}, \cdots, \frac{S_{p_1p_2}^2}{S_{p_1p_1}S_{p_2p_2}})'$到原点的距离, 如果把点的距离改为欧式距离则有

$$\|x - y\|_2 = \left[\sum_{i=1}^{m} (x_i - y_i)^2\right]^{\frac{1}{2}}.$$

因此我们还可以得到另外一个检验统计量:

$$T_{new5} = \sum_{i<j} \frac{S_{ij}^2}{S_{ii} S_{jj}} = \sum_{1 \le i < j} (\frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}})^2.$$

$T_{new5}^{\frac{1}{2}}$就是$\text{Vec}(\text{diag}\{S_{11}^{\frac{1}{2}}, S_{22}^{\frac{1}{2}}, \cdots, S_{p_1p_1}^{\frac{1}{2}}\} S_{12} \text{diag}\{S_{(p_1+1)(p_1+1)}^{\frac{1}{2}}, S_{(p_1+2)(p_1+2)}^{\frac{1}{2}}, \cdots, S_{pp}^{\frac{1}{2}}\})$ 到原点的距离.

对于本章提出的检验统计量T_{new4}, 由于其具体分布很难求出, 因此用类似Bootstrap的方法进行检验, 具体步骤如下:

(1)从$N_p(\mathbf{0},\Sigma)$中抽取(n-1)个样本$v_1,v_2,\cdots,v_{n-1}$, 并从$v_1,v_2,\cdots,v_{n-1}$中随机放回地抽取容量为2$(n-1)$个样本$y_1,y_2,\cdots,y_{2(n-1)}=\begin{pmatrix}y_{11}\\y_{12}\end{pmatrix},\begin{pmatrix}y_{21}\\y_{22}\end{pmatrix},\cdots,\begin{pmatrix}y_{2(n-1),1}\\y_{2(n-1),2}\end{pmatrix}$, 其中$y_{11},y_{21},\cdots,y_{2(n-1),1}$为$p_1$维向量, $y_{12},y_{22},\cdots,y_{2(n-1),2}$为$p_2$ 维向量;

(2)令$z_i=\frac{1}{\sqrt{2}}\begin{pmatrix}y_{2i-1,1}+y_{2i,1}\\y_{2i-1,2}-y_{2i,2}\end{pmatrix}$, 其中$i=1,2,\cdots,n-1$, 则$z_1,z_2,\cdots,z_{n-1}$可以看作来自$N_p\left(\mathbf{0},\begin{bmatrix}\Sigma_{11}&0\\0&\Sigma_{22}\end{bmatrix}\right)$, 并用$z_1,z_2,\cdots,z_{n-1}$计算检验统计量$T_{new4}$的值并记为$T$;

(3)有放回的从$z_1,z_2,\cdots,z_{n-1}$中抽取容量为$n-1$的Bootstrap样本$x_1,x_2,\cdots,x_{n-1}$, 用$x_1,x_2,\cdots,x_{n-1}$计算出检验统计量T_{new4}的一个观测值T_c;

(4)重复步骤(2)C次, 得到$\{T_c,c=1,2,\cdots,C\}$, C是一个比较大的正整数;

(5)计算p值, $p(x)=\frac{1}{C}\sum\limits_{c=1}^{C}I_{\{T_c\geq T\}}$.

对给定的显著水平α, 若$p(x)\leq\alpha$, 则拒绝原假设.

上述步骤(2)中对z_i进行该形式定义的原因为: $v_1,v_2,\cdots,v_{n-1}$独立同分布于$N_p(\mathbf{0},\Sigma)$, 其中$\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$, 此时$\Sigma_{12}$不一定为0, 所以不能直接用$v_1,v_2,\cdots,v_{n-1}$作为样本直接计算检验统计量$T_{new4}$. 因此我们构造$z_i$来保证$z_i\sim N_p(\mathbf{0},\Sigma)$, 其中$\Sigma_{12}=0$. 设$\begin{bmatrix}\xi_1\\\xi_2\end{bmatrix}\sim N_p\left(\mathbf{0},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}\right)$, $\begin{bmatrix}\eta_1\\\eta_2\end{bmatrix}\sim N_p\left(\mathbf{0},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}\right)$, 且它们相互独立, 则有

$$\frac{1}{\sqrt{2}}\begin{bmatrix}\xi_1+\eta_1\\\xi_2-\eta_2\end{bmatrix}=\frac{1}{\sqrt{2}}\begin{bmatrix}I_{p_1}&0&I_{p_1}&0\\0&I_{p_2}&0&-I_{p_2}\end{bmatrix}\begin{bmatrix}\xi_1\\\xi_2\\\eta_1\\\eta_2\end{bmatrix}\sim N_p\left(\mathbf{0},\frac{1}{2}\begin{bmatrix}I_{p_1}&0&I_{p_1}&0\\0&I_{p_2}&0&-I_{p_2}\end{bmatrix}\begin{bmatrix}\Sigma&0\\0&\Sigma\end{bmatrix}\begin{bmatrix}I_{p_1}&0\\0&I_{p_2}\\I_{p_1}&0\\0&-I_{p_2}\end{bmatrix}\right),$$

其中

$$\begin{aligned}\begin{bmatrix}I_{p_1}&0&I_{p_1}&0\\0&I_{p_2}&0&-I_{p_2}\end{bmatrix}\begin{bmatrix}\Sigma&0\\0&\Sigma\end{bmatrix}\begin{bmatrix}I_{p_1}&0\\0&I_{p_2}\\I_{p_1}&0\\0&-I_{p_2}\end{bmatrix}&=\begin{bmatrix}I_{p_1}&0&I_{p_2}&0\\0&I_{p_2}&0&-I_{p_2}\end{bmatrix}\begin{bmatrix}\Sigma_{11}&\Sigma_{12}&0&0\\\Sigma_{21}&\Sigma_{22}&0&0\\0&0&\Sigma_{11}&\Sigma_{12}\\0&0&\Sigma_{21}&\Sigma_{22}\end{bmatrix}\begin{bmatrix}I_{p_1}&0\\0&I_{p_2}\\I_{p_1}&0\\0&-I_{p_2}\end{bmatrix}\\&=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}&\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{21}&-\Sigma_{21}&-\Sigma_{22}\end{bmatrix}\begin{bmatrix}I_{p_1}&0\\0&I_{p_2}\\I_{p_1}&0\\0&-I_{p_2}\end{bmatrix}=2\begin{bmatrix}\Sigma_{11}&0\\0&\Sigma_{22}\end{bmatrix}.\end{aligned}$$

因此

$$\frac{1}{\sqrt{2}}\begin{bmatrix}\xi_1+\eta_1\\\xi_2-\eta_2\end{bmatrix}\sim N_p\left(\mathbf{0},\begin{bmatrix}\Sigma_{11}&0\\0&\Sigma_{22}\end{bmatrix}\right).$$

5.2 小结

本章在高维数据的背景下, 结合并交原则讨论了随机变量的独立性检验问题, 不失一般性地, 我们重点讨论了$q = 2$时的情况, 特别研究了S_{11}, S_{22}为奇异矩阵时的情况, 先结合矩阵的相关性质给出了$n < \min(p_1, p_2)$(即S_{11}, S_{22}为奇异矩阵)时广义似然比统计量的形式$\lambda(X)$, 进一步缩小了a, b的取值范围, 对该统计量进行了进一步的优化, 从而得到了更为准确的检验统计量形式T_{new4}. 由于很难计算出检验统计量的具体分布, 所以在检验部分我们提出了一种类似Bootstrap的方法进行检验.

结　论

科学信息技术迅猛发展, 使得高维数据广泛存在并应用于生物、经济金融、医疗等各个领域, 对此类数据进行分析处理具有实际意义. 本文以高维数据为研究背景, 讨论了一系列有关正态总体的假设检验问题.

第二章运用Bonferroni思想研究了高维数据下正态分布的均值检验问题, 提出了新的划分方法, 对于单样本和两样本问题, 具体划分方法分别如下:

$$k=\lfloor\frac{p}{n-1}\rfloor+1,\quad k=\lfloor\frac{p}{n_1+n_2-2}\rfloor+1.$$

其中$\lfloor x\rfloor$表示小于或等于x的最大整数. 此划分方法既做到了减少变量间相关信息的缺失, 又具有降维的效果, 且划分后的每个部分都可用Hotelling T^2进行检验, 再结合Bonferroni校正思想计算实际的检验水平, 并通过数值模拟可知本章提出的三个检验统计量均在某种情况下具有优良性.

第三章以似然比以及并交原则为基础, 研究了高维数据下的方差分析问题, 并给出了一种新的广义似然比检验

$$Z(X)=(2\pi)^{-\frac{n}{2}}[T(X)]^{-\frac{n}{2}}\exp\{-\frac{n}{2}\},$$

其中$T(X)=\lambda_1[F'[J'(X'X)^{-1}J]^{-1}F]$, F是列满秩矩阵, 且有$J'(H-C)J=FF'$. 特别地, 给出了$k=2$及$k=3$时F'的值:

$$F_2'=\sqrt{\frac{n_1n_2}{n}}\begin{bmatrix}1&-1\end{bmatrix},\quad F_3'=\frac{1}{\sqrt{n}}\begin{bmatrix}\sqrt{n_1(n-n_1)}&-\frac{n_2\sqrt{n_1}}{\sqrt{n-n_1}}&-\frac{n_2\sqrt{n_1}}{\sqrt{n-n_1}}\\0&-\sqrt{\frac{nn_2n_3}{n-n_1}}&\sqrt{\frac{nn_2n_3}{n-n_1}}\end{bmatrix}.$$

通过对$k=3$的情况进行数值模拟, 可知该广义似然比统计量具有优良性.

第四章以似然比以及并交原则为基础, 研究了高维数据下的正态均值线性约束的问题, 并给出了一种新的广义似然比检验

$$F(X)=\frac{1}{(2\pi)^{\frac{n}{2}}[U(X)]^{\frac{n}{2}}}\cdot\exp\{-\frac{n}{2}\},$$

其中$U(X)=\frac{1}{1_n'(X'R'RX)^{-1}1_n}$. 通过数值模拟可知, 该检验比$T_{BS}$具有优越性.

第五章以似然比以及并交原则为基础, 结合矩阵的相关性质, 研究了高维数据下的随机变量独立性检验问题, 并重点讨论了$q=2$时的情形, 并给出了$n<\min(p_1,p_2)$ 时似然比统计量的形式

$$T_{new4}=(n-2)\max_{i<j}\frac{{S_{ij}}^2}{S_{ii}S_{jj}},$$

$$T_{new5}=\sum_{i<j}\frac{S_{ij}^2}{S_{ii}S_{jj}}=\sum_{1\le i<j}(\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}})^2.$$

并提出了一种类似Bootstrap的检验方法.

参考文献

[1] T.S. Tian. Dimensionality reduction for classification with high-dimensional data[D]. *University of Southern California*, 2009.

[2] P.Hall, J.S.Marron, A.Neeman. Geometric representation of high dimension, low sample size data[J]. *Journal of Royal Statistical Society*, 2005, 67, 427-444.

[3] T.W. Anderson. An introduction to multivariate statistical analysis[M]. *Third edition.John Wiley & Sons,New York,* 2003.

[4] H. Hotelling. The generalization of Student's ratio[J]. *The Annals of Mathematical Satistics*, 1931, 2: 360-378.

[5] A.A. Amimi, M.J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components[J]. *The annals of Statistics*, 2009, 37: 2877-2921.

[6] D. Shen, H. Shen, J.S. Marron. Consistency of sparse PCA in high dimension,low sample size contexts[J]. *Journal of Multivariate Analysis*, 2008, 115: 317-333.

[7] V.Q. Vu, J. Lei. Minimax rates of estimation for sparse PCA in high dimensions[J]. *In The Fifteenth International Conference on Artificial Intelligence and Statistics.*

[8] A.P. Dempster. A high dimensional two sample significance test[J]. *The Annals of Mathematical Statistics*, 1985, 29(4): 995-1010.

[9] A.P. Dempster. A significance test for the separation of two Sample Problem[J]. *Biometrics*, 1960, 16(1): 41-50.

[10] Z. Bai, H. Saranadasa. Effect of High Dimension:By an Example of a Two Sample Problem[J]. *Statistica Sinnica*, 1996, 6(2): 311-329.

[11] J.G. Zhao, X.Z. Xu. A generalized likelihood ratio test for normal mean when p is greater than n[J]. *Computational Statistics and Data Analysis*, 2016, 99: 91-104.

[12] 赵俊光.高维数据下总体均值的假设检验问题[D]. 北京:北京理工大学,2016.

[13] M.S. Srivastava, M. Du. A test for the mean vector with fewer observations than the dimension[J]. *Journal of Multivariate Analysis*, 2008, 99: 386-402.

[14] M.S. Srivastava. A test for the mean vector with fewer observations than the dimension under non-normality[J]. *The Annals of Statistics*, 2009, 100: 518-532.

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

[16] Z.D. Bai, D.D. Jiang, J.F. Yao, S.R. Zheng. Correction to LRT on Large Dimensional Covariance Matrix by RMT[J]. *The Annals of Statistics*, 2009, 37: 3822-3840.

[17] T.F. Jiang, F.Yang. Central Limit Theorems for Classical Likelihood Ratio Tests for High-Dimensional Normal Distributions[J]. *The Annals of Statistics*, 2013, 41: 2029-2074.

[18] J.R. Schott. Testing for complete independence in high dimensions[J]. *Biometrika Trust*, 2005, 92: 951-956.

[19] J.R. Schott. A test for independence of two sets of variables when the number of variables is large relative to the sample size[J]. *Statistics and Probability Letters*, 2008, 78: 3096-3102.

[20] D.D. Jiang, Z.D.Bai, S.R.Zheng. Tarone. Testing the independence of sets of large-dimensional variables[J]. *Science China(Mathematics)*, 2013, 56(01): 133-145.

[21] M.S. Srivastava. Some tests concerning the covariance matrix in high-dimensional data [J]. *Journal of the Japan Statistical Society*, 2005, 35: 251-272.

[22] M.S. Srivastava. Some tests criteria for the covariance matrix with fewer observations than the dimension[J]. *Acta Et Commentationes Universitatis Tartuensis De Mathematica*, 2006, 10: 77-93.

[23] G.H. Wang, C.L. Zou, Z.J. Wang. A necessary test for complete independence in high dimensions using rank-correlations[J]. *Journal of Multivariate Analysis*, 2013, 121: 224-232.

[24] 张楚润.多元统计分析[M]. 北京:科学出版社,2006.

[25] S.N. Roy. On a heuristic method of test construction and its use in multivariate analysis[J]. *The Annals of Mathematical Statistics*, 1953, 24: 220-238.

[26] C.E. Bonferroni. II Calcolo delle assicurazioni su gruppi di teste[J]. *Studi in Onore del Professore Salvatore ortu Carboni*, 1953, 13-16.

[27] W.R. Rice. Analyzing tables of statistical tests[J]. *Evolution*, 1989, 43: 223-225.

[28] S. Holm. A sample sequentially rejective multiply test procedure[J]. *Scandinavian Journal of Statistics*, 1979, 6(1): 65-70.

[29] J.M. Bland, D.G. Altman. Multiple significance tests:the Bonferroni method[J]. *The BMJ*, 1995, 310(6973): 170.

[30] R.E. Tarone. A modified Bonferroni method for discrete data[J]. *Biometrics*, 1990, 46(2): 515-522.

[31] J.K. Gamage. Generalized p-value and Multivariate Behrens-Fisher Problem[J]. *Linear algebra and its applications*, 1997, 253: 369-377.

[32] J.K. Gamage, T. Mathew, S. Weerahandi. Generalized p-values and Generalized Confidence Regions for the Multivariate Behrens-Fisher Problem and MANOVA[J]. *Journal of Multivariate Analysis*, 2004, 88: 177-189.

[33] 范永辉. 线性混合效应模型的估计与检验[D].北京工业大学, 2007.

[34] B.M. Bennett. Note on a solution of the Generalized Behrens-Fisher Problem[J]. *Annals of the Institute of Statistical Mathematics*, 1951, 2: 87-90.

[35] G.S. James. Tests of Linear Hypotheses in Univariate and Multivariate Analysis when the Ratios of the Variance are Unkown[J]. *Biometrika*, 1954, 41: 19-43.

[36] S. Johansen. The Welch-James Appoximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression[J]. *Biometrika*, 1980, 67: 85-92.

[37] S. Weerahandi. Generalized Confidence Interval[J]. *American Statist,Assoc*, 1993, 88(423): 899-905.

[38] S.S. Wilks. Certain generazetions in the analysis of variance[J]. *Biometrics*, 1932, 23: 471-494.

[39] 王桂松, 史建红, 尹素菊, 等. 线性模型引理[M]. 北京:科学出版社, 2004.

[40] R.J. Muirhead. Aspects of Multivariate Statistical Throry[M]. *A John Wilky & Sons*, 1922.

[41] G.E.P. Box. A general distribution theory for a class of likelihood criteria[J]. *Biomrtrika*, 1949, 36: 317-346.

[42] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京:科学出版社, 1982.

[43] 王桂松, 史建红, 贾忠贞. 矩阵不等式[M]. 北京:科学出版社, 2006.

[44] K.W. Tusi, S. Weerahandi. Generalized p-values in Significance Testing of Hypothesis in the Presence of Nuisance Parameters[J]. *Journal of the American Statistical Association*, 1989, 84: 602-607.

[45] 孙炯, 王万义, 郝建文. 泛函分析[M]. 高等教育出版社, 2010.

致　　谢

时光荏苒, 转眼间, 在天津师范大学攻读研究生的三年学习生活即将画上圆满的句号, 在这短暂又充实的三年里, 有太多美好又难忘的时光, 让我收获很多, 感慨很多. 借此机会, 请允许我向三年来教授我知识、帮助我成长的各位老师表示感谢, 向三年来给予我陪伴、支持和鼓励的同学和朋友表示感谢.

本论文是在导师范永辉老师的悉心关怀和精心指导下完成的. 从论文的选题、研究方法和构思、论文的撰写、修改与定稿, 导师都倾注了大量的心血, 给予了宝贵的意见、悉心的指导和帮助, 每当我在撰写论文中遇到问题与困难时, 导师总能为我指明方向, 并在百忙之中对论文进行详细的批阅. 在这三年的研究生学习和生活中, 范老师深厚的学术造诣、严谨细致的工作作风、谦逊的学者风范一直是我学习的榜样, 对我的谆谆教诲我将永远铭记于心. 在此向他致以最真挚的感谢.

感谢师大数学科学学院的各位老师, 三年来对我学业和生活上给予了关心、指导和帮助.

感谢我的舍友、同学和朋友, 这三年里在学习和生活中给予我无限帮助和陪伴, 感谢你们陪我走过这美好的三年.

最后特别感谢我的父母, 他们一直为我默默奉献, 感谢他们一直以来的支持与鼓励, 这永远是我学习和生活最坚强的后盾, 也是我不断奋力前进的动力.

攻读学位期间发表的学术论文

[1] 李熠璇, 范永辉. 高维数据下关于正态均值线性约束的检验[J]. 哈尔滨商业大学学报(自然科学版), 2021, 37(01): 73-78.