密级：公开　　中图分类号：0212

浙江工商大學

硕士学位论文

论文题目：　高维数据下的多元均值检验

作者姓名：　江梦婕

学科专业：　统计学

研究方向：　数理统计

指导教师：　江涛

提交日期：　2020　年　6　月

Dissertation Submitted to Zhejiang Gongshang University for Master Degree of Statistics

MULTIVARIATE MEAN TESTING IN HIGH DIMENSIONAL DATAS

Author: Jiang Mengjie

Major: Science Statistics

Supervisor: Prof. Jiang Tao

June. 2020

School of Statistics and Mathematics

Zhejiang Gongshang University

Hangzhou, 310018, P. R. China

高维数据下的多元均值检验

摘 要

多元均值检验是经典统计学中检验的重要组成部分。在经典统计学中，多元均值检验常用的方法是 Hotelling T^2检验。在样本量 n 大于变量维数 p 时，该检验方法能够很好的控制两类错误率并且具有良好的功效和优良的性质。随着科学技术的发展，我们所面对的数据维度越来越高，高维数据的出现给传统的多元统计带来了挑战。对多元均值检验来说，在数据的变量维数 p 大于样本量 n 的情况下，样本协方差矩阵是一个奇异阵，这使得 Hotelling T^2检验统计量从形式上失去了定义；同时，经典的统计极限理论都是在样本量 n 远大于数据维数 p 的条件下建立的，所以在该情况下，经典的极限理论也不再适用。

本文在高维数据的背景下，对于多元均值检验问题，在对角 Hotelling T^2检验的基础上进行改进提出新的检验方法。我们采用几何收缩法估计的方差所构成的对角矩阵代替 Hotelling T^2检验统计量中的样本协方差估计值，建立几何收缩对角 Hotelling T^2检验。改进的对角 Hotelling T^2检验通过使用几何收缩法估计方差提高了方差估计的稳定性，因此几何收缩对角 Hotelling T^2检验适用于变量维度 p 大于样本量 n 的数据。同时，本文在一定条件下计算得到了几何收缩对角 Hotelling T^2检验统计量的渐近分布。

在蒙特卡洛模拟中，以第一类错误率和经验功效为评价标准，比较几何收缩对角 Hotelling T^2检验和其它改进的 Hotelling T^2检验方法在多元均值检验方面的优劣性。模拟研究表明，几何收缩对角 Hotelling T^2检验在大多数情形下都优于其他检验方法。在实证分析中，本文使用各种检验方法，对长期生存脑癌病人和短期生存脑癌病人之间基因拷贝数的差别以及在两种不同治疗方法下骨髓癌病人的基因表达分别进行了检验。实证数据分析都表明，几何收缩对角 Hotelling T^2检验显著优于其它检验方法。

关键词： 多元均值检验；Hotelling T^2检验；高维数据；渐近分布；微阵列基因

MULTIVARIATE MEAN TESTING IN HIGH DIMENSIONAL DATAS

Abstract

Multivariate mean testing is an important part in the classical statistics. In classical statistics, a commonly used method in Multivariate mean testing is Hotelling T^2 test. When the sample size *n* is much larger than the number of variables *p*, this test statistics performs well in terms of controlling two types of errors, providing a high power and possessing good asymptotic properties. However, the number of variables becomes larger and larger as the development of science and technique. It is a challenge task for classical statistics to handle high dimensional data sets. When the number of variables *p* is greater than the sample size *n* in Multivariate mean testing, the sample covariance matrix becomes singular. Hence the Hotelling T^2 test statistics is not well defined in such scenario. Meanwhile, the asymptotic properties of a test statistics in classical statistics is based on the assumption that the sample size is much larger than the number of variables. Hence, the asymptotic theory is inappropriate for the large *p* small *n* cases.

For a high dimensional data set, we propose a new method based on diagonal Hotelling T^2 test to address Multivariate mean testing problems. We replace the covariance matrix estimation in Hotelling T^2 test statistics with a diagonal matrix estimated from the geometric shrinkage of the sample covariance matrix, and propose a geometric shrinkage diagonal Hotelling T^2 test. The improved diagonal Hotelling T^2 test applies geometric shrinkage estimation method to improve the stability of the covariance matrix estimation. Hence, the geometric shrinkage diagonal Hotelling T^2 test can handle data sets whose number of

variables is greater than the sample size. Furthermore, we provided the asymptotic distribution of the proposed statistics under null hypothesis.

In Monte Carlo simulation studies, we evaluate the performance of the proposed method as well as other improved Hotelling T^2 tests by comparing their type I error rate and power. From the simulation, the geometric shrinkage diagonal Hotelling T^2 test outperforms other methods in most scenarios. Then we apply all methods in real data analysis. One is to test the difference of DNA copy numbers between long survival patients and short survival patients in brain cancer study. Another is to test the difference of gene expressions of patients between different treatments in myeloma. Both results show that the performance of shrinkage diagonal Hotelling T^2 test is significantly better than other methods.

KEYWORDS: Multivariate mean testing; Hotelling T^2 test; high-dimensional data; asymptotic distribution; microarray gene

目录

第 1 章　绪论

1.1　研究的背景及其意义

近几十年，许多生物学者和医学研究者都在研究基因与各种疾病的关联性，比如全基因组关联研究。全基因组关联研究从研究对象中随机抽取大量的病例和对照样本并且通过基因芯片技术进行测序，以获得各样本个体的观测数据并整理为各突变点的基因型数据（潘东东等人，2014）。分析两组基因表达是否相同涉及到经典的多元总体均值检验问题，统计分析的结果可以反映基因与疾病的关系以及药物治疗的效果，为以后疾病的预防诊疗及药物的研究提供新的思路。

多元均值检验不仅可以用于基因数据方面的对比还可以应用于很多其它方面数据的对比研究，而且在协方差未知的情况下多元均值检验是统计学上最重要的问题之一。对于多元均值检验问题，经典的处理方法是 Hotelling T^2检验。在“样本量 n 大于样本维度 p”的经典的背景下，Hotelling T^2检验具有很多优良的性质，例如，能够很好的控制两类错误并且具有良好的检验功效，在原假设下能够得到检验统计量的精确分布，在线性交换下检验统计量是一致最有效不变检验以及检验统计量是可容许检验。

近几十年来，随着科学技术的不断发展，我们可以收集和存储的数据越来越多，所面对的数据维度也越来越高，当样本维数 p 大于样本量 n 时，这种数据就为高维数据。在基因研究中，高通量技术能够在一个基因芯片内产生上千个观测，例如，研究人员借助微阵列技术（Microarray）可以在一个实验中监测上千个基因的表达水平；下一代基因测序技术（Next generation sequencing）可以用来检查一个基因上数千个位置处的 DNA 拷贝数，而样本量 n 由于各种成本等各种限制会小于样本维数 p。高维数据的出现给传统的多元统计理论带来了极大的挑战。对于多元均值检验，在高维数据的背景下，样本协方差矩阵是一个奇异矩阵导致经典的 Hotelling T^2检验的统计量从形式上失去了定义。为了解决高维数据下的多元均值检验问题，在过去几十年中，许多学者对 Hotelling T^2检验方法进行改进，提出新的检验统计方法，这些检验方法大致可以分为四类：第一类是未收缩 Hotelling T^2检验，该检验从 Hotelling T^2检验统计量中去除协方差矩阵以避免协方差矩阵估计；第二类是正则化 Hotelling T^2检验，该检验将正则化方法应用于协方差矩阵以解决奇点问题；第三类是对角似然比检验，该检验方法结合对角矩阵形式和似然比检验方法，该检验方法的检验统计量是对数变换的 t 检验统计量之和；第四类是

对角 Hotelling T^2检验，该检验方法假设 Hotelling T^2检验统计量中的协方差为对角矩阵。

基于对角 Hotelling T^2检验的优良性质，本文假设协方差矩阵为对角矩阵。实际上，协方差矩阵为对角矩阵的假设已普遍用于高维小样本数据（Bickel 和 Levina，2004，Dudoit 等人，2002）。Bickel 和 Levina（2004）指出，如果估计的相关性都非常复杂，假设协方差矩阵为对角矩阵的效果会更好。在判别分析中，Lee 等人（2016）也证明在微阵列数据中逆广义矩阵的判别规则不如对角判别规则表现好。尽管对角 Hotelling T^2检验的应用较广，但是其本身还具有一些不足比如当样本量很小时，样本方差并非一个稳定估计，这表明有必要对对角 Hotelling T^2检验进行进一步的改进以提高其性能。

本文在高维数据下的对角 Hotelling T^2检验的基础上做进一步的改进和研究。在基因差异表达的基因微阵列研究中，样本协方差矩阵是奇异矩阵，它不再是总体协方差矩阵的一个“好”的估计。由于基因检测的成本很高，在研究中通常会在相对较少的样本上测量大量基因的表达水平，因此，基于每个基因的数据得到的方差估计量的统计推断不是很稳定，会影响 Hotelling T^2检验统计量的功效，而通过跨基因合并信息估计基因特异性方差可以提高方差估计的有效性。因此本文考虑用几何收缩法估计协方差矩阵，再在该方差的基础上改进 Hotelling T^2检验建立几何收缩对角 Hotelling T^2检验（Geometric Shrinking Diagonal Hotelling T^2 Test，GSDT）。

1.2 国内外研究概况

多元总体均值检验可以应用于很多方面数据的对比，而且在总体协方差未知的条件下多元均值检验问题是统计学上最重要的问题之一。Hotelling T^2检验是一种检验均值是否相等的检验方法，这个经典的方法最初由 Hotelling（1931）提出的。在样本量 n 大于数据维度 p 时，该方法表现良好且在经典的低维情况下进行了广泛的研究。但是，在研究高维小样本数据时，样本协方差矩阵是一个奇异矩阵，所以 Hotelling T^2检验在该情况下就会失效。对于 Hotelling T^2检验的奇点问题，许多研究人员提出了一些解决方法，比如 Dempster（1960）提出的非精确显著性检验和随机化检验，但是这两种方法对自由度和一些相关的问题进行了复杂的估计（Bai 和 Saranadasa，1996），导致这些方法在实际应用中效果并不好。在近几十年中，许多学者在改进 Hotelling T^2检验以适应于高维数据做了大量工作，他们提出了新的检验统计方法。改进的 Hotelling T^2检验主要

分为四大类：未收缩 Hotelling T^2检验、正则化 Hotelling T^2检验、似然比检验以及对角 Hotelling T^2检验，下面分别对这些研究方法进行综述。

未收缩 Hotelling T^2检验是从 Hotelling T^2检验统计量中去除协方差矩阵以避免协方差矩阵估计。Bai 和 Saranadasa（1996）首先提出这种方法，他们用$(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^T(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})$代替 Hotelling T^2检验统计量中的$(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^T\boldsymbol{S}^{-1}(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})$，其中，$\overline{\boldsymbol{X}}$和$\overline{\boldsymbol{Y}}$是样本均值向量，$\boldsymbol{S}$是样本协方差矩阵。他们证明了在数据维度 p 和样本量 n 为同阶的情况下，他们所提出的检验比传统的 Hotelling T^2检验具有更好的功效。Zhang 和 Xu（2009）将未收缩 Hotelling T^2检验应用于高维数据的 k 样本 Behrens-Fisher 问题，检验了具有不等协方差的几个高维正态总体均值向量是否相等。Chen 和 Qin（2010）将该方法应用于高维小样本数据的双样本均值检验，他们在$\|\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}}\|^2$的基础上减去了$\sum_{i=1}^{n_1}\boldsymbol{X}_i^T\boldsymbol{X}_i$和$\sum_{i=1}^{n_2}\boldsymbol{Y}_i^T\boldsymbol{Y}_i$得到新的检验统计量。

正则化 Hotelling T^2检验是将正则化方法应用于协方差矩阵以解决奇点问题，Chen 等人（2011）和 Li 等人（2016）用正则化估计量$(\boldsymbol{S}+\lambda\boldsymbol{I}_p)^{-1}$代替 Hotelling T^2检验统计量中的样本协方差估计量，其中，$\boldsymbol{I}_p$是单位矩阵，$\lambda>0$是正则化参数。Shen 等人（2011）提出类似的方法，他们用$(\lambda\boldsymbol{S}+(1-\lambda)\boldsymbol{I}_{\boldsymbol{p}})^{-1}$（$0\le\lambda<1$）代替 Hotelling T^2检验统计量中的样本协方差估计量，当$\lambda=0$时，该方法就为未收缩 Hotelling T^2检验。Lopes（2011）用$E_R^{-1}\{\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{S}\boldsymbol{R})^{-1}\boldsymbol{R}^T\}$替换 Hotelling T^2检验统计量中的$\boldsymbol{S}^{-1}$，其中$\boldsymbol{R}$是$p\times k$的随机矩阵，$E_R(.)$是分布上的期望算子。

对角似然比检验是结合对角矩阵形式和似然比检验方法，该检验方法的检验统计量是对数变换的 t 检验统计量之和。Hotelling T^2检验的起源也是来自于似然比检验，Jiang 和 Yang（2013）和 Jiang 和 Qi（2015）在样本维数 p 小于样本量 n 但是满足$p/n\to0$条件下，使用似然比检验法分布检验了正态分布的均值向量和协方差矩阵是否相等。Zhao 和 Xu（2016）提出了广义高维似然，并通过联合交叉法改进了似然比检验，用改进的方法对均值向量进行比率检验。Stadler 和 Mukherjee（2017）在高维数据的条件下，对双样本提出了似然比检验。Hu, Tong 和 Genton（2018）提出了对角似然比检验，该检验方法结合对角矩阵形式和似然比检验方法，该检验方法的检验统计量是对数变换的 t 检验统计量之和，Hu, Tong 和 Genton（2018）分别考虑了单样本和双样本条件下的均值检验。

对角 Hotelling T^2检验是假设 Hotelling T^2检验统计量中的协方差矩阵为对角矩阵，在该假设下，无论 p 是否大于 n，对角矩阵总是可逆的，所以不用考虑矩阵是否为奇异阵的问题。Wu, Genton 和 Stefanski（2006）在数据缺失的情况下，改进了 Hotelling T^2检验，提出对角 Hotelling T^2检验。Srivastava 和 Du（2008）在高维背景下分别提出了单样本问题和双样本问题下的对角 Hotelling T^2检验。他们提出用样本协方差矩阵 $\boldsymbol{S}$ 对角线上的元素构成的对角矩阵代替 $\boldsymbol{S}$ 来构建新的检验统计量：

$$T_D^2 = \frac{n_1 n_2}{n_1+n_2}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})^T\{\mathrm{diag}(\boldsymbol{S})\}^{-1}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}). \qquad (1\text{-}1)$$

Srivastava 和 Du（2008）并证明了该统计量的渐近零分布和检验功效。Srivastava（2009）在非正态性条件下的建立了对角 Hotelling T^2检验统计量。Park 和 Nag Ayyala（2013）的检验是修改了 Srivastava（2009）的检验统计量，他们也是假设协方差矩阵是对角矩阵。Srivastava 等人（2013），Feng 等人（2015）和 Gregory 等人（2015）在不等协方差矩阵条件下考虑对角 Hotelling T^2检验，构建的统计量为

$$(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})^T\{\mathrm{diag}(\boldsymbol{S}_1)/n_1 + \mathrm{diag}(\boldsymbol{S}_2)/n_2\}^{-1}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}), \qquad (1\text{-}2)$$

其中，$\boldsymbol{S_1}$和$\boldsymbol{S_2}$是两个样本协方差矩阵。Dong（2016）提出了一个基于收缩的对角 Hotelling T^2检验，用收缩法得到的方差估计量代替样本协方差矩阵中的对角线上元素构成方差矩阵。Dinu 等人（2007）基于对角 Hotelling T^2检验，提出一种新的检验方法，在每个基因特异性方差估计中添加一个常数以稳固方差估计，该检验方法称为“GAM-GS”。

除了这四类检验方法，还有 Lopes，Jacob 和 Wainwright（2011）使用随机投影方法得到检验方法，他们考虑的是将数据随机投影到一个低维空间，在低维空间上计算 Hotelling T^2检验统计量，然后对所有低维空间上的 Hotelling T^2检验统计量取平均。Thulin（2013）在 Lopes，Jacob 和 Wainwright（2011）的基础上进行了进一步的改进，使用随机子空间的方法得到了检验。Wei，Lee，Wichers，Li 和 Marron（2016）把数据投影到两个特殊方向上得到了检验。为了解决在 Hotelling T^2检验统计量中样本协方差不可逆的问题，Chen，Paula，Prenticea 和 Wang（2011）和 Shen，Lin 和 Zhu（2011）提出直接使用总体协方差矩阵的可逆估计量代替样本协方差矩阵计算 Hotelling T^2检验统计量。还有一些其他的检验方法，具体参考 Ahamd，Rosen 和 Singull（2014），Ahamd（2014），Katayamaa，Kano 和 Srivastava（2013）和 Maboudou-Tchao 和 Silva（2013）。

在多元检验问题中，选择对角 Hotelling T^2检验方法比较均值向量，检验统计量中只包含协方差矩阵对角线元素，此时，只需要估计方差矩阵而不需要估计协方差矩阵。而在高维小样本的情况下，特异基因样本方差的估计通常是不稳定的，会影响 Hotelling T^2检验统计量的功效，因此需要提高方差估计。许多学者对于改进方差估计做了大量的工作，下面对收缩法估计方差进行综述。

为了稳定方差估计的最早的一种方法是由 Tusher 等人（2001）提出，他们为了避免小方差估计的不当的影响，在他们的 SAM 检验中用$(s_j + c)/2$代替标准差σ_j，其中常数c是收缩因子。Cui 和 Churchill（2003）借用基因的信息来估计方差，他们提出使用合并样本方差估计量（个体基因方差估计的平均值）作为方差估计值。随后，Cui 等（2005）在此基础上提出了用 James-Stein 收缩方法（James 和 Stein，1961）估计方差，但是该方差估计量在样本量 *n* 较小时不是很理想。Tong 和 Wang（2007）注意到方差估计量是在 *t* 检验以及对角 Hotelling T^2检验的统计量的分母里，他们提出直接用$1/\sigma_j$的估计会比用σ_j的估计更稳健，因此他们考虑$(\sigma_j^2)^t (t \neq 0)$的收缩估计，其中$1/\sigma_j$和$\sigma_j$分别是$t = -1/2$和$t = 1/2$的特殊情况。Tong 和 Wang（2007）将对$(\sigma_j^2)^t (t \neq 0)$的一系列几何收缩估计总结为下式的形式：

$$\tilde{\sigma}_j^{2h} = (c_1 t_j^{2h})^{\alpha} (c_2 s_j^{2h})^{1-\alpha} \ , 0 \leq \alpha \leq 1, \qquad (1\text{-}3)$$

其中，当$\sigma_j^2 = \sigma^2$时，$c_2 s_j^{2h}$是σ_j^{2h}的无偏估计，$c_1 t_j^{2h}$是σ^2的无偏估计，特别地，当$h = 1$和$h = -1$时，分别对应着方差估计和其逆的估计。收缩参数α控制着基因特异性方差估计$c_2 s_j^{2h}$向校正偏差的几何均值$c_1 t_j^{2h}$转换的度。当收缩参数$\alpha = 0$时，不存在收缩；当收缩参数$\alpha = 1$时，方差估计都收缩到合并方差。Tong 等人（2012）提出了另一种 James-Stein 收缩估计方法，用于将个体样本方差收缩到算术平均值的方差。Tong 和 Wang（2007）和 Tong 等人（2012）在 Stein 损失函数和平方损失函数下得出几何平均和算术平均估计量的收缩参数，并且在选择几何平均对角矩阵作为目标矩阵下对方差进行估计。肖敏（2018）在用几何收缩估计方法估计方差过程中，选择 Log-Euclidean 平方损失函数来选择最优收缩参数，并且考虑了几何平均对角矩阵和算术平均对角矩阵这两种矩阵作为目标矩阵。

1.3 研究思路

在本文中，我针对单样本和双样本的情况提出了几何收缩对角 Hotelling T^2检验。通过用几何收缩法估计得到的方差替代对角 Hotelling T^2检验中的样本方差，构造检验统计量，同时计算检验统计量的均值与方差以及检验统计量在原假设下的渐近分布。最后，模拟和实证对几何收缩对角 Hotelling T^2检验和其他检验方法进行对比。

第一章绪论，首先介绍了本文的研究背景及其意义，然后根据国内外现有的文献，从四类改进的 Hotelling T^2检验和收缩法估计方差这两个方面进行综述，最后给出了本文的研究思路和本文研究的创新点。

第二章高维均值向量检验，首先介绍了单样本和双样本多元均值检验及其检验统计量和检验统计量的所服从的分布。然后分别从检验统计量和渐近功效两个方面介绍了 Chen 和 Qin（2010），Chen 等人（2011），Hu，Tong 和 Genton（2018）和 Stivastava 和 Du（2008）改进的 Hotelling T^2检验。

第三章几何收缩对角 Hotelling T^2检验，本章首先介绍了在单样本情形下，通过 Log-Euclidean 平方损失函数计算所得的最优收缩参数，方差的最优几何型收缩估计和结合估计所得的方差建立单样本几何收缩对角 Hotelling T^2检验统计量。然后介绍了在双样本情形下，通过 Log-Euclidean 平方损失函数计算所得的最优收缩参数，方差的最优几何型收缩估计和结合估计所得的方差建立双样本几何收缩对角 Hotelling T^2检验统计量。最后分别计算了单样本和双样本检验统计量的均值与方差以及分别证明了单样本和双样本检验统计量在原假设下的渐近分布。

第四章蒙特卡洛模拟，本章在一定的条件下，选择不同的模拟参数，得到几何收缩对角 Hotelling T^2检验和其他四种改进 Hotelling T^2检验方法在不同信号强度和稀疏程度下的第一类错误率和功效函数，并通过比较第一类错误率和功效函数来评价这五种检验方法。

第五章实证分析，首先我们通过使用各个检验统计方法来检测长期生存的脑癌病人和短期生存的脑癌病人是否存在某基因片段上的基因拷贝数的不同，分析各个检验统计方法在检验脑癌数据中的表现。为了更好的比较几何收缩对角 Hotelling T^2检验和其它方法，我们也使用各个检验统计方法来检测病人在两种不同的治疗方案下是否存在某基因片段上基因表达不同。

第六章总结和展望，本章对全文内容进行了系统的总结，并对未来的研究方向进行了展望。

1.4 创新点

本文的研究主要是基于 Srivastava 和 Du（2008）的对角 Hotelling T^2检验，Srivastava 和 Du（2008）是用样本协方差矩阵 $\boldsymbol{S}$ 对角线上的元素构成的对角矩阵代替经典 Hotelling T^2检验统计量中的样本协方差矩阵 $\boldsymbol{S}$ 来构建新的检验统计量。本文针对于高维数据下的多元均值检验，对对角 Hotelling T^2检验进行了改进，用几何收缩方差估计构成的总体方差估计矩阵代替经典 Hotelling T^2检验统计量中的样本协方差矩阵 $\boldsymbol{S}$，建立了几何收缩对角检验（GSDT）并在原假设下证明了 GSDT 统计量的渐近卡方分布。本文通过对角矩阵的引入减少了检验统计量种奇异阵的影响，同时通过使用几何收缩估计的方法提高了方差的稳定估计。

第 2 章　高维均值向量检验

在多元统计问题中，总体均值检验是一个十分重要的问题。对于单样本检验问题，设$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T, i = 1, \ldots, n$，是来自服从多元正态分布$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$的随机样本，其中，$\boldsymbol{\mu}$是总体均值向量，$\boldsymbol{\Sigma}$是总体协方差矩阵。考虑下面的假设

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{v.s.} \quad H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{2-1}$$

其中，$\boldsymbol{\mu}_0$是给定的常量。对于假设（2-1），单样本的 Hotelling T^2检验统计量为

$$T_1^2 = n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0), \tag{2-2}$$

其中，样本均值$\overline{\boldsymbol{X}}$和样本协方差$\boldsymbol{S}$定义为

$$\overline{\boldsymbol{X}} = \sum_{i=1}^n \boldsymbol{X}_i / n, \tag{2-3}$$

$$\boldsymbol{S} = \sum_{i=1}^n (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T / (n-1). \tag{2-4}$$

当$p \leq n-1$且满足原假设H_0时，检验统计量$\{(n-p)/(p(n-1))\}\, T_1^2$服从$F(p, n-p)$分布。

对于双样本检验问题，设$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$，$i = 1, \ldots, n_1$和$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T$，$i = 1, \ldots, n_2$为分别来自服从多元正态分布$N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$和$N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$的随机样本。考虑下面的假设

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{v.s.} \quad H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{2-5}$$

对于假设（2-5），双样本 Hotelling T^2检验统计量定义为

$$T_2^2 = \{n_1 n_2 / (n_1 + n_2))\} (\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})^T \boldsymbol{S}_{pool}^{-1} (\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}), \tag{2-6}$$

其中样本均值$\overline{\boldsymbol{X}}$，$\overline{\boldsymbol{Y}}$和合并样本协方差$\boldsymbol{S}_{pool}$定义为

$$\bar{\boldsymbol{X}} = \sum_{i=1}^{n_1} \boldsymbol{X}_i / n_1, \tag{2-7}$$

$$\bar{\boldsymbol{Y}} = \sum_{i=1}^{n_2} \boldsymbol{Y}_i / n_2, \tag{2-8}$$

$$\boldsymbol{S}_{pool} = \left\{\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T + \sum_{i=1}^{n_2} (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T\right\} / (n_1 + n_2 - 2). \tag{2-9}$$

当$p \le n$且满足原假设H_0时，检验统计量$\{(n_1 + n_2 - p - 1)/(p(n_1 + n_2 - 2))\}\, T_2^2$服从$F(p, n_1 + n_2 - p - 1)$分布。Bai 和 Saranadasa（1996）推导出了 Hotelling T^2检验的渐近功效函数：若满足$y_n = \frac{p}{n} \to y \in (0,1)$， $n_1/(n_1 + n_2) \to \kappa \in (0,1)$ 和$\|\delta\|^2 = o(1)$，则

$$\beta_{\boldsymbol{H}}(\delta) - \boldsymbol{\Phi}(-\xi_\alpha + \sqrt{\frac{n(1-y)}{2y}}\kappa(1-\kappa)\|\delta\|^2) \to 0, \tag{2-10}$$

其中，ξ_α是标准正态分布的$1-\alpha$分位数，$\beta_{\boldsymbol{H}}(\delta)$是 Hotelling T^2检验的功效函数。

在数据维数 p 小于样本量 n 的条件下，经典 Hotelling T^2检验具有优良的性质，例如，在原假设下能够得到统计量的精确分布，在线性交换下检验统计量是一致最有效不变检验以及检验统计量是可容许检验。但是，当数据维数p大于样本数n时，即在高维数据背景下，样本协方差矩阵是不可逆矩阵，因此在该条件下 Hotelling T^2检验统计量没有定义。许多学者对 Hotelling T^2检验进行了改进，使得其检验方法在高维数据下具有定义且具有优良的检验效果，下面分别具体介绍四种改进的检验方法。

2.1 CQ 未缩放检验

为了改进 Hotelling T^2检验，Dempster（1958，1960）提出了非精确检验，但是 Dempster 的非精确检验强烈依赖于正态性假设。此外，Dempster 的非精确检验统计涉及到复杂的卡方近似值的“自由度”的估计。Bai 和 Saranadasa （1996）在非精确检验的基础上提出了一个新的检验方法—渐近正态分布检验，简称为 BS 检验，该检验方法不依赖于正态性假设，但是渐近正态分布检验依赖 4 阶矩的存在，具有一定程度的不稳定性。Chen 和 Qin（2010）在渐近正态分布检验的基础上对双样本均值检验进行了更深入的研究，提出了新的检验方法，简称为 CQ 未缩放检验。CQ 未缩放检验需假设：

（1）$\boldsymbol{X}_i = \boldsymbol{\Gamma Z}_i + \boldsymbol{\mu}_1$，$i = 1, \ldots, n_1$，$\boldsymbol{Y}_i = \boldsymbol{\Gamma Z}_i + \boldsymbol{\mu}_2$，$i = 1, \ldots, n_2$，其中$\boldsymbol{\Gamma}$是$p \times m$维矩阵$(m \leq \infty)$，$\boldsymbol{\Gamma\Gamma^T} = \boldsymbol{\Sigma}$，$\boldsymbol{Z}_i$是$m \times 1$维的独立同分布随机向量，且满足$E\boldsymbol{Z}_i = \boldsymbol{0}$，$\mathrm{V}ar(\boldsymbol{Z}_i) = \boldsymbol{I}_m$，$Ez_{ij}^4 = 3 + \Delta < \infty$ 和

$$E\left(z_{il_1}^{\alpha_1} z_{il_2}^{\alpha_2} \cdots z_{il_q}^{\alpha_q}\right) = E\left(z_{il_1}^{\alpha_1}\right) E\left(z_{il_2}^{\alpha_2}\right) \cdot \cdots \cdot E\left(z_{il_q}^{\alpha_q}\right), \tag{2-11}$$

其中，q是正整数，使得$\sum_{l=1}^{q} \alpha_l \leq 8$，$l_1 \neq l_2 \neq \cdots \neq l_q$，$\Delta$表示$z_{ij}$的四阶矩和$N(0,1)$之间的不同，$n_1/n_1 + n_2 \to \kappa \in (0,1)$；

（2）$tr\left(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_l\right) = \mathrm{o}\{tr^2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2\}$，其中 *i, j ,k, l*=1 或者 2；

（3）$\boldsymbol{\mu}'\boldsymbol{\Sigma}_i\boldsymbol{\mu} = \mathrm{o}\{\tau tr(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2\}$，$i = 1,2$，其中，$\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$，$\tau = 1/(n_1 + n_2)$。

BS 检验统计量中的$\|\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}\|^2$是两个均值向量差的模的平方的一个估计，但不是无偏估计，Bai 和 Saranadasa （1996）在构建检验统计量时，在$\|\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}\|^2$的基础上减去了$\tau tr\boldsymbol{S_n}$，而 Chen 和 Qin（2010）在$\|\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}\|^2$的基础上减去了$\sum_{i=1}^{n_1} \boldsymbol{X}_i^T \boldsymbol{X}_i, \sum_{i=1}^{n_2} \boldsymbol{Y}_i^T \boldsymbol{Y}_i$，得到新的检验统计量为

$$T_{CQ}^2 = \frac{\sum_{i \neq j}^{n_1} \boldsymbol{X}_i^T \boldsymbol{X}_j}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} \boldsymbol{Y}_i^T \boldsymbol{Y}_j}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \boldsymbol{X}_i^T \boldsymbol{Y}_j}{n_1 n_2}. \tag{2-12}$$

在条件（1）-（3）下，Chen 和 Qin（2010）证明了，当$n \to \infty$和$p \to \infty$时，在原假设下有

$$\frac{T_{CQ}^2 - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\sqrt{\frac{2}{n_1(n_1-1)} tr(\boldsymbol{\Sigma_1})^2 + \frac{2}{n_2(n_2-1)} tr(\boldsymbol{\Sigma_2})^2 + \frac{4}{n_1 n_2} tr\boldsymbol{\Sigma_1\Sigma_2}}} \to N(0,1), \tag{2-13}$$

其中，$tr(\boldsymbol{\Sigma_1})^2$，$tr(\boldsymbol{\Sigma_1})^2$和 $tr\boldsymbol{\Sigma_1\Sigma_2}$ 的比例相合估计分别为

$$\widehat{tr(\boldsymbol{\Sigma_1})^2} = \{n_1(n_1-1)\}^{-1} tr\left\{\sum_{j \neq k}^{n_1} \left(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_{(j,k)}\right)\boldsymbol{X}_j^T \left(\boldsymbol{X}_k - \overline{\boldsymbol{X}}_{(j,k)}\right)\boldsymbol{X}_k^T\right\}, \tag{2-14}$$

$$\widehat{tr(\boldsymbol{\Sigma_2})^2} = \{n_2(n_2-1)\}^{-1} tr\left\{\sum_{j \neq k}^{n_2} \left(\boldsymbol{Y}_j - \overline{\boldsymbol{Y}}_{(j,k)}\right)\boldsymbol{Y}_j^T \left(\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_{(j,k)}\right)\boldsymbol{Y}_k^T\right\}, \tag{2-15}$$

$$\widehat{tr\boldsymbol{\Sigma_1\Sigma_2}} = \{n_1 n_2\}^{-1} tr\left\{\sum_{l=1}^{n_1}\sum_{k=1}^{n_2} \left(\boldsymbol{X}_l - \overline{\boldsymbol{X}}_{(l)}\right)\boldsymbol{X}_l^T \left(\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_{(k)}\right)\boldsymbol{Y}_k^T\right\}, \tag{2-16}$$

其中，$\overline{\boldsymbol{X}}_{(j,k)}$是除去$\boldsymbol{X}_j$和$\boldsymbol{X}_k$之后的样本均值，$\overline{\boldsymbol{Y}}_{(j,k)}$是除去$\boldsymbol{Y}_j$和$\boldsymbol{Y}_k$之后的样本均值，$\overline{\boldsymbol{X}}_{(l)}$是除去$\boldsymbol{X}_l$之后的样本均值和$\overline{\boldsymbol{Y}}_{(l)}$是除去$\boldsymbol{Y}_l$之后的样本均值。将式（2-14），式（2-15）和式（2-16）代入式（2-13）得

$$M_{CQ}=\frac{T_{CQ}^2-\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\sqrt{\frac{2}{n_1(n_1-1)}\widehat{tr(\boldsymbol{\Sigma}_1)^2}+\frac{2}{n_2(n_2-1)}\widehat{tr(\boldsymbol{\Sigma}_2)^2}+\frac{4}{n_1n_2}\widehat{tr\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2}}}\to N(0,1). \tag{2-17}$$

故当$M_{CQ}>\xi_\alpha$时，拒绝原假设，其中ξ_α是标准正态分布的$1-\alpha$分位数。同时在（1）-（3）条件下，检验的渐近功效为

$$\beta_{CQ}(\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|)-\boldsymbol{\Phi}(-\xi_\alpha+\frac{n\kappa(1-\kappa)\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\sqrt{2tr(\tilde{\Sigma}(\kappa)^2)}})\to 0, \tag{2-18}$$

其中，$\tilde{\Sigma}(\kappa)=(1-\kappa)\boldsymbol{\Sigma}_1+\kappa\boldsymbol{\Sigma}_2$。

2.2 CPPW 正则化检验

Chen 等人（2011）考虑在样本量较小和数据具有较高的缺失率的情况下，检验统计量的零分布可能与其近似渐近分布有很大的不同，故他们提出使用分参数方法来推导所提出的检验统计量的显著性水平。Chen 等人（2011）在样本维数 p 大于样本量 n 的条件下，提出了正则化 Hotelling T^2检验，简称为 CPPW 正则化检验。该检验方法是将正则化方法应用于协方差矩阵以解决奇点问题，在双样本条件下其检验统计量为

$$T_{CPPW}^2=\{n_1n_2/(n_1+n_2))\}(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^T(\boldsymbol{S}_{pool}+\lambda\boldsymbol{I}_p)^{-1}(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}}), \tag{2-19}$$

其中，$\overline{\boldsymbol{X}}$和$\overline{\boldsymbol{Y}}$为样本均值，$\boldsymbol{S}_{pool}$是样本协方差矩阵，$\boldsymbol{I}_p$是单位矩阵，$\lambda>0$是正则化参数。CPPW 检验需假设：

（1） $\boldsymbol{\Sigma}$是一个正定矩阵；

（2） 当$n\to\infty$和$p\to\infty$时，$p/n\to\gamma\in(0,\infty)$，$\sqrt{n}|p/n-\gamma|\to 0$；

（3） 设$\boldsymbol{\Sigma}$的特征值为$\tau_{1,p} \geq \cdots \geq \tau_{p,p} > 0$。$\boldsymbol{\Sigma}$的经验谱分布（ESD）为$H_p(\tau) = \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{[\tau_{j,p},\infty)}(\tau)$，当$p \to \infty$时，在 H 的每一个连续有 $H_p(\tau) \to H(\tau)$，其中 $\sup(H) \in$ 紧集$[h_1, h_2]$，$0 < h_1 \leq h_2 < \infty$；

（4） $\limsup_{p\to\infty}\tau_{1,p} < \infty$，$\liminf_{p\to\infty}\tau_{p,p} > 0$。

定义 $m_F(z)$，$z \in \mathbb{C}$是$\boldsymbol{S}_{pool}$的极限 ESD 的 Stieltje*s* 变换（Marcenko 和 Pastur，1967），即$F_{n,p}(x) = \frac{1}{p}\sum_{j=1}^{p} \mathbf{1}_{[d_j,\infty)}(x)$，其中$d_1 \geq \cdots \geq d_p$是$\boldsymbol{S}_{pool}$的特征值。故

$$m_{F_{n,p}(z)} = \frac{1}{p}\sum_{j=1}^{p}\frac{1}{d_j - z} = \frac{1}{p} tr\left(\left(\boldsymbol{S}_{pool} - z\boldsymbol{I}_p\right)^{-1}\right), \tag{2-20}$$

$$m_F(z) = \int \frac{dH(\tau)}{\tau(1-\gamma-\gamma z m_F(z))-z}. \tag{2-21}$$

在满足原假设和假设（1）-（4）下，Chen 等人（2011）证明了

$$M_{CPPW} = \frac{\sqrt{p}\left((1/p)T^2_{CPPW} - \Theta_1(\lambda,\gamma)\right)}{\left(2\Theta_2(\lambda,\gamma)\right)^{1/2}} \to N(0,1), \tag{2-22}$$

其中，$\Theta_1(\lambda,\gamma) = \frac{1-\lambda m_F(-\lambda)}{1-\gamma(1-\lambda m_F(-\lambda))}$，$\Theta_2(\lambda,\gamma) = \frac{1-\lambda m_F(-\lambda)}{1-\gamma(1-\lambda m_F(-\lambda))^3} - \lambda\frac{m_F(-\lambda)-\lambda m_F'(-\lambda)}{(1-\lambda+\gamma\lambda m_F(-\lambda))^4}$。

因此，当$M_{CPPW} > \xi_\alpha$时，拒绝原假设，其中ξ_α是标准正态分布的$1-\alpha$分位数。同时在（1）-（4）条件下，检验的渐近功效为

$$\beta_{CPPW}(\delta) - \boldsymbol{\Phi}(-\xi_\alpha + \frac{c^2\Theta_1(\lambda,\gamma)}{\sqrt{2\gamma\Theta_2(\lambda,\gamma)}}) \to 0, n \to \infty. \tag{2-23}$$

2.3 HTG 的对角似然检验

为了改进对角 Hotelling T^2检验，Hu，Tong 和 Genton（2018）提出了对角似然比检验，简称为 HTG 对角似然检验。该检验方法结合对角矩阵形式和似然比检验方法，检验统计量是对数变换的 *t* 检验统计量之和。Hu，Tong 和 Genton（2018）分别考虑了单样本和双样本条件下的均值检验，本文主要介绍说明双样本条件下的 HTG 检验。

设$\boldsymbol{X}_i=(X_{i1},\dots,X_{ip})^T$，$i=1,\dots,n_1$和$\boldsymbol{Y}_i=(Y_{i1},\dots,Y_{ip})^T$，$i=1,\dots,n_2$是分别来自服从多元正态分布$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$和$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$的随机样本。令$N=n_1+n_2$，$\overline{\boldsymbol{X}}$和$\overline{\boldsymbol{Y}}$是样本均值，$\boldsymbol{S}_{pool}$是合并样本协方差矩阵。

考虑下面的假设：

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \quad \text{v.s.} \quad H_1:\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2, \tag{2-24}$$

Hu，Tong 和 Genton（2018）提出的统计量为

$$T_{HTG}^2=N\sum_{j=1}^{p}\log\left\{1+\frac{n_1n_2}{N(N-2)}\frac{(\overline{X}_j-\overline{Y}_j)^2}{s_j^2}\right\}=N\sum_{j=1}^{p}\log\left\{1+\frac{t_{Nj}^2}{\nu}\right\}, \tag{2-25}$$

其中，$t_{Nj}=\sqrt{n_1n_2/N}\left(\overline{\boldsymbol{X}}_j-\overline{\boldsymbol{Y}}_j\right)/s_j$是双样本条件下自由度为$\nu=N-2$的标准 t 检验统计量。记 $V_{Nj}=N\log\left\{1+\frac{t_{Nj}^2}{\nu}\right\}$，$\Gamma(x)$为 gamma 函数，$\Psi(x)=\Gamma'(x)/\Gamma(x)$为 digamma 函数，$D(x)=\Psi\{(x+1)/2\}-\Psi(x/2)$，$G_1=N\mathrm{D}(\nu)$和$G_2=N^2\{D^2(\nu)-2D'(\nu)\}$。HTG 检验需假设：

（1） 设$\alpha(r)=\sup\{\alpha(\mathcal{F}_1^k,\mathcal{F}_{k+r}^p):1\leq k\leq p-r\}$，其中$\mathcal{F}_a^b=\mathcal{F}_{a,n}^b=\sigma\{V_{Nj}:a\leq j\leq b\}$。假定平稳序列$\{V_{Nj}\}$满足强混合条件，使得当$r\to\infty$时，$\alpha(r)\downarrow 0$，其中↓表示单调递减收敛；

（2） 设对于部分$\delta>0$，有$\sum_{r=1}^{\infty}\alpha(r)^{\delta/(2+\delta)}<\infty$和对于任意的$k\geq 0$，存在$\lim_{p\to\infty}\sum_{j=1}^{p-k}Cov\left(V_{Nj},V_{N,j+k}\right)/(p-k)=\gamma(k)$。

Hu，Tong 和 Genton（2018）证明了若序列$\{V_{Nj}\}$是平稳的且满足条件（1）和（2），则在零假设下，对于任意固定的$N\geq 4$，有

$$M_{HTG}=\frac{T_{DLRT}^2-pG_1}{\tau\sqrt{p}}\xrightarrow{\mathcal{D}}N(0,1),p\to\infty, \tag{2-26}$$

其中，$\tau^2=G_2-G_1^2+2\sum_{k=1}^{\infty}\gamma(k)$。故当$M_{DLRT}>\xi_\alpha$时，拒绝原假设，其中$\xi_\alpha$是标准正态分布的$1-\alpha$分位数。

当$\mu_1\neq\mu_2$时，Hu，Tong 和 Genton（2018）考虑了下面的备择假设：

$$\boldsymbol{\mu_1}-\boldsymbol{\mu_2}=\sqrt{\frac{N}{n_1 n_2}}\boldsymbol{\delta}, \tag{2-27}$$

其中，$\boldsymbol{\delta}=\left(\delta_1,\ldots,\delta_p\right)^T$。设$\Delta=\left(\Delta_1,\ldots,\Delta_p\right)^T=\left(\delta_1/\sigma_{11},\ldots,\delta_p/\sigma_{pp}\right)^T$，且满足

$$\Delta_j^2\leq M_1,\ j=1,2,\ldots,p, \tag{2-28}$$

其中，σ_{jj}^2是$\boldsymbol{\Sigma}$对角线上的元素，M_1是独立于 N 和 p 的常数。

若序列$\{V_{Nj}\}$是平稳的且满足条件（1）和（2），则在备择假设（2-27）和条件（2-28）下，检验的渐近功效为

$$\beta_{HTG}(T_{HTG}^2)=1-\boldsymbol{\Phi}(\xi_\alpha-\frac{\Delta^T\Delta/\sqrt{p}}{\sqrt{\tau^2}}),\ (N,p)\to\infty, \tag{2-29}$$

因此，当$\sqrt{p}=o\left(\sum_{j=1}^p\delta_j^2/\sigma_{jj}^2\right)$时，$\beta_{DLRT}(T_{DLRT}^2)\to 1$。

2.4 SD 对角 Hotelling T^2检验

Srivastava 和 Du（2008）在高维背景下分别提出了单样本和双样本条件下的对角 Hotelling T^2检验，本文主要介绍双样本问题下的对角 Hotelling T^2检验。

设$\boldsymbol{X}_i=(X_{i1},\ldots,X_{ip})^T$，$i=1,\ldots,n_1$和$\boldsymbol{Y}_i=(Y_{i1},\ldots,Y_{ip})^T$，$i=1,\ldots,n_2$分别是来自服从多元正态分布$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$和$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$的随机样本，其中，$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$是总体均值向量，$\boldsymbol{\Sigma}$是总体协方差矩阵。记$\boldsymbol{\Sigma}=\left(\sigma_{ij}\right)$，$\boldsymbol{D}_\sigma=\mathrm{diag}\left(\sigma_{11},\sigma_{22},\cdots,\sigma_{pp}\right)$，$\boldsymbol{R}=\boldsymbol{D}_\sigma^{-1/2}\boldsymbol{\Sigma}\boldsymbol{D}_\sigma^{-1/2}=\left(\rho_{ij}\right)$，且满足

$$0<\lim_{p\to\infty}\frac{trR^i}{p}<\infty\ \text{和}\ \lim_{p\to\infty}\max_{1\leq i\leq p}\frac{\lambda_i}{\sqrt{p}}=0, \tag{2-30}$$

其中，λ_i，$i=1,2,\cdots,p$，是总体相关系数矩阵 $\boldsymbol{R}$ 的特征值，$n=O\left(p^\zeta\right)$，$\frac{1}{2}<\zeta\leq 1$。Srivastava 和 Du（2008）提出用样本协方差矩阵 $\boldsymbol{S}$ 对角线上的元素构成的对角矩阵代替样本协方差矩阵 $\boldsymbol{S}$ 来构建新的检验统计量，其检验统计量为

$$T_{SD}^2 = \frac{n_1 n_2}{n_1+n_2}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})^T \boldsymbol{D}_S^{-1}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}), \quad (2\text{-}31)$$

其中，$\boldsymbol{D}_S = \mathrm{diag}(s_{11}, s_{22}, \cdots, s_{pp})$，$s_{ii}(i = 1,2,\cdots,p)$是样本协方差矩阵对角线上的元素，$\boldsymbol{S} = \{\sum_{i=1}^{n_1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T + \sum_{i=1}^{n_2}(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})^T\}/n_1 + n_2 - 2$是样本协方差矩阵。在一定的条件下，Srivastava 和 Du（2006）证明了

$$M_{SD} = \frac{\frac{n_1 n_2}{n_1+n_2}(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}})^T \boldsymbol{D}_S^{-1}(\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}}) - \frac{np}{n-2}}{\sqrt{2\left(tr\boldsymbol{R}^2 - \frac{p^2}{n}\right)c_{p.n}}} \to N(0,1), \quad (2\text{-}32)$$

其中，$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$，$c_{p.n} = 1 + \frac{tr\boldsymbol{R}^2}{p^{3/2}}$。因此，当$M_{SD} > \xi_\alpha$时，拒绝原假设。对于检验功效，取备择假设$\mu_1 - \mu_2 = \left(\frac{n_1+n_2}{n n_1 n_2}\right)^{\frac{1}{2}}\boldsymbol{\delta}$，其中，$\boldsymbol{\delta}$是一个常向量且满足$\frac{\boldsymbol{\delta}^T D_\sigma^{-1}\boldsymbol{\delta}}{p} \le M$，$M$是一个不依赖与$p$的常数，SD 的检验功效表示为

$$\beta_{SD}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) - \boldsymbol{\Phi}(-\xi_{1-\alpha} + \frac{n_1 n_2}{n_1+n_2}\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{D}_\sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{2tr(\mathbf{R}^2)}}) \to 0. \quad (2\text{-}33)$$

第 3 章　几何收缩对角 Hotelling T^2检验

Srivastava 和 Du（2008）建立的对角 Hotelling T^2检验统计量是用样本方差的对角矩阵来估计协方差矩阵的。然而，Srivastava 等人（2013）模拟研究证明当观测数量n有限时，例如当观测数少于 10 时，样本方差不再是可靠的估计，这时对角 Hotelling T^2检验也不再是可靠的检验。因此，为了改进对角 Hotelling T^2检验，需要改进对总体方差的估计，Dinu 等人（2007）对此做过类似的改进。在本文中，我考虑采用几何收缩估计法对总体方差进行估计，在几何平均对角矩阵为目标矩阵的条件下计算总体方差的估计量并通过 Log-Euclidean 平方损失函数来计算确定最优收缩参数（肖敏，2017）。

设$\mathbf{\Sigma}$是总体协方差矩阵，$\mathbf{\Sigma}$对角上的元素为$\mathrm{diag}(\mathbf{\Sigma}) = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2)$。在本章中，主要考虑的对角协方差矩阵，为记号方便，从本节开始用$\mathbf{\Sigma}$表示$\mathrm{diag}(\mathbf{\Sigma})$。肖敏（2017）对$\mathbf{\Sigma}^h = \{\mathrm{diag}(\mathbf{\Sigma})\}^h = \mathrm{diag}(\sigma_1^{2h}, \cdots, \sigma_p^{2h})$，$h \neq 0$进行了估计，其几何型收缩估计为

$$\widehat{\mathbf{\Sigma}}_G^h = (C_1 \boldsymbol{T}^h)^\alpha (C_2 \boldsymbol{S}^h)^{1-\alpha}, 0 \leq \alpha \leq 1, \qquad (3\text{-}1)$$

其中，$\boldsymbol{S}$是对角样本协方差矩阵，$\boldsymbol{T}$是目标矩阵，其形式为对角矩阵，对角上的元素均为对角样本协方差矩阵对角线上元素的几何平均值。$\mathbf{\Sigma}_G^h$对角线上对应元素的估计为

$$\widehat{\sigma}_j^{2h} = (c_1(h) t_j^{2h})^\alpha (c_2(h) s_j^{2h})^{1-\alpha}, 0 \leq \alpha \leq 1, \qquad (3\text{-}2)$$

其中，α是收缩参数，$c_i(h)$（$i = 1,2$）是调整参数，特别地，$h = 1$和$h = -1$分别是对角型总体协方差矩阵及其逆地估计。

肖敏（2017）在 Log-Euclidean 平方损失函数 $L(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \left\| \log \widehat{\mathbf{\Sigma}} - \log \mathbf{\Sigma} \right\|_F^2 = \sum_{j=1}^p \left\{ \log(\sigma_j^{2h}) - \log \sigma_j^{2h} \right\}^2$框架下分别在算术平均对角目标矩阵（AMT）和几何目标矩阵（GMT）条件下，给出了α的最优参数估计α^*以及$\sigma_j^{2h}(\alpha)$的最优参数估计$\widehat{\sigma}_j^{2h}(\alpha^*)$。当$h = 1$时，$\widehat{\sigma}_j^{2h}(\alpha^*)$为方差的最优收缩估计；当$h = -1$时，$\widehat{\sigma}_j^{2h}(\alpha^*)$为方差倒数的最优收缩估计。

在 Wald 的统计决策理论框架下，统计推断问题取决于损失函数的选择（陈希孺2009）。在几何收缩法估计方差中，本文选择 Log-Euclidean 平方损失函数估计最优收缩参数，即

$$L\left(\boldsymbol{W},\widehat{\boldsymbol{W}}\right)=p^{-1}\left\|\log\widehat{\boldsymbol{W}}-\log\boldsymbol{W}\right\|_F^2=p^{-1}\sum_{i=1}^{p}(\log(\hat{w}_{ii})-\log(w_{ii}))^2, \tag{3-3}$$

其中，矩阵$\boldsymbol{W}=\mathrm{diag}(w_{11},w_{22},\ldots,w_{pp})$和矩阵估计值$\widehat{\boldsymbol{W}}=\mathrm{diag}(\hat{w}_{11},\hat{w}_{22},\ldots,\hat{w}_{pp})$都为正定矩阵。将$\widehat{\boldsymbol{\Sigma}}$和$\boldsymbol{\Sigma}$代入式（3-3）可得

$$L\left(\boldsymbol{\Sigma},\widehat{\boldsymbol{\Sigma}}\right)=p^{-1}\left\|\log\widehat{\boldsymbol{\Sigma}}-\log\boldsymbol{\Sigma}\right\|_F^2=p^{-1}\sum_{i=1}^{p}\left(\log\left(\hat{\sigma}_j^2\right)-\log\left(\sigma_j^2\right)\right)^2. \tag{3-4}$$

式（3-4）在$\alpha\in[0,1]$上是一个凸函数，故最优收缩参数是唯一的且可以计算得到其显示表达式。

$$\begin{aligned}\hat{\alpha}&=\arg\min_{\alpha}E\left(p^{-1}\left\|\log\widehat{\boldsymbol{\Sigma}}-\log\boldsymbol{\Sigma}\right\|_F^2\right)\\&=\arg\min_{\alpha}p^{-1}E\left(\alpha\log c_1(1)t^2+(1-\alpha)\log c_2(1)s_j^2-\log\sigma_i^2\right)\\&=\frac{\sum_{j=1}^{p}E(\log c_2(1)-\log c_1(1)+\log s_j^2-\log t_j^2)(\log c_2(1)+\log s_j^2-h\log\sigma_j^2)}{\sum_{j=1}^{p}E(\log c_2(1)-\log c_1(1)+\log s_j^2-\log t_j^2)^2}.\end{aligned} \tag{3-5}$$

因此，可得最优收缩参数为

$$\hat{\alpha}^*=\max\{0,\min\{1,\hat{\alpha}\}\}, \tag{3-6}$$

其中，$\hat{\alpha}$为式（3-5）所定义。

下面介绍在几何平均目标矩阵和 Log-Euclidean 平方损失函数条件下，单样本和双样本的总体方差的最优估计以及在该方差下建立的 GSDT 检验。

3.1 GSDT 检验：单样本情形

假设$\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$是来自$p$元正态分布的简单随机样本，其中$\boldsymbol{\mu}$是总体均值向量，$\boldsymbol{\Sigma}$是总体协方差矩阵，$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$均未知，$\boldsymbol{\Sigma}$正定，样本容量$n$小于等于维数$p$。检验如下假设：

$$H_0:\ \boldsymbol{\mu}=\boldsymbol{\mu}_0\ \ \text{v.s.}\ \ H_1:\ \boldsymbol{\mu}\neq\boldsymbol{\mu}_0, \tag{3-7}$$

其中，H_0，H_1分别是单样本原假设和备择假设，$\boldsymbol{\mu}_0$是给定的常量。样本均值$\overline{\boldsymbol{X}}$和样本协方差矩阵$\mathbb{S}$（为与双样本的样本协方差矩阵符号区别，记单样本的样本协方差矩阵为$\mathbb{S}$）分别为

$$\overline{\boldsymbol{X}}=\sum_{i=1}^{n}\boldsymbol{X}_i/n\text{和}\mathbb{S}=\sum_{i=1}^{n}(\boldsymbol{X}_i-\overline{\boldsymbol{X}})(\boldsymbol{X}_i-\overline{\boldsymbol{X}})^T/(n-1)=(\mathbb{s}_{ij}). \tag{3-8}$$

定义样本协方差矩阵的对角线上的元素构成的对角矩阵为

$$\mathbb{S}=\text{diag}(\mathbb{S})=\text{diag}\left(\mathbb{s}_1^2,\ldots,\mathbb{s}_p^2\right)=\text{diag}(\mathbb{s}_{11},..,\mathbb{s}_{pp}). \tag{3-9}$$

其中，单样本对角协方差矩阵$\mathbb{S}$对角线上的元素为$\mathbb{s}_j^2$，$j=1,\cdots,p$。

本文考虑目标矩阵为几何平均对角矩阵的情况，即取样本协方差矩阵对角元素的几何平均为中心值，几何平均目标矩阵为

$$\mathbb{T}=\text{diag}(\mathbb{t},\mathbb{t},\ldots,\mathbb{t}), \tag{3-10}$$

其中$\mathbb{t}=\left(\prod_{j=1}^{p}\mathbb{s}_j^2\right)^{1/p}$。定义调整参数为

$$c_1(h)=\left(\frac{v}{2}\right)^h\left(\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2}+\frac{h}{p}\right)}\right)^p,c_2(h)=\left(\frac{v}{2}\right)^h\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2}+h\right)}, \tag{3-11}$$

其中，$v=n-1$，$\Gamma(\bullet)$是伽马函数。

将（3-11）式代入（3-5）式中，在 Log-Euclidean 平方损失函数下，计算可得最优收缩参数为

$$\hat{\alpha}_1^*=\max\{0,\min\{1,\hat{\alpha}\}\}, \tag{3-12}$$

其中，$\hat{\alpha}=\frac{\sum_{j=1}^{p}E(A+hB)(logc_2(h)+hD)}{\sum_{j=1}^{p}E(A+hB)^2}$，$A=\log c_2(h)-\log c_1(h)$，$B=\log \mathbb{s}_j^2-\log \mathbb{t}$，$D=\log \mathbb{s}_j^2-\log \sigma_j^2$。此时，对应的$\sigma_j^{2h}$的最优几何型收缩估计为

$$\hat{\sigma}_j^{2h}=(c_1(h)\mathbb{t}^{2h})^{\hat{\alpha}_1^*}(c_2(h)\mathbb{s}_j^{2h})^{1-\hat{\alpha}_1^*}. \quad (3\text{-}13)$$

在单样本情形下，针对假设（3-7）提出的 GSDT 检验统计量为

$$T_{GSDT1}^2(\hat{\alpha}_1^*)=n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0)^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{1}}(\hat{\alpha}_1^*)(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_0)$$

$$=n\sum_{j=1}^{p}(\bar{X}_j-\mu_{0j})^2\,\hat{\sigma}_j^{-2}(\hat{\alpha}_1^*), \quad (3\text{-}14)$$

其中，$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{1}}(\alpha^*)=\mathrm{diag}(\hat{\sigma}_1^{-2}(\hat{\alpha}_1^*),\dots,\hat{\sigma}_p^{-2}(\hat{\alpha}_1^*))$。对于该检验统计量的渐近分布在第 3.3 节中进行分析。

3.2 GSDT 检验：双样本情形

假设$\boldsymbol{X}_i=(X_{i1},\dots,X_{ip})^T$，$i=1,\dots,n_1$和$\boldsymbol{Y}_i=(Y_{i1},\dots,Y_{ip})^T$，$i=1,\dots,n_2$分别是来自服从多元正态分布$N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$和$N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$的随机样本，其中，$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$分别是是总体 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的均值向量，$\boldsymbol{\Sigma}$是总体协方差矩阵。$\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$和 $\boldsymbol{\Sigma}$ 均未知，$\boldsymbol{\Sigma}$正定，样本容量n_1和n_2均小于等于维数p。考虑下面的假设：

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \quad v.s. \quad H_1:\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2, \quad (3\text{-}15)$$

其中，H_0，H_1分别是双样本原假设和备择假设。样本均值$\overline{\boldsymbol{X}}$，$\overline{\boldsymbol{Y}}$和样本协方差矩阵$\boldsymbol{S}$分别为

$$\overline{\boldsymbol{X}}=\textstyle\sum_{i=1}^{n_1}\boldsymbol{X}_i/n_1,\ \ \overline{\boldsymbol{X}}=\sum_{i=1}^{n_2}\boldsymbol{Y}_i/n_2,$$

$$\boldsymbol{S}=\left\{\textstyle\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\overline{\boldsymbol{X}})(\boldsymbol{X}_i-\overline{\boldsymbol{X}})^T+\sum_{i=1}^{n_2}(\boldsymbol{Y}_i-\overline{\boldsymbol{Y}})(\boldsymbol{Y}_i-\overline{\boldsymbol{Y}})^T\right\}/(n_1+n_2-2)\ =(s_{ij}). \quad (3\text{-}16)$$

定义样本协方差矩阵的对角线上的元素为

$$\boldsymbol{S} = \mathrm{diag}(\boldsymbol{S}) = \mathrm{diag}(s_1^2, \dots, s_p^2). \tag{3-17}$$

类似于单样本 GSDT 对角 Hotelling T^2检验，定义双样本条件下几何平均目标矩阵为

$$\boldsymbol{T} = \mathrm{diag}(t^2, t^2, \dots, t^2), \tag{3-18}$$

其中$t^2 = \left(\prod_{j=1}^{p} s_j^2\right)^{1/p}$。定义调整参数为

$$c_1(h) = \left(\frac{v}{2}\right)^h \left(\frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2}+\frac{h}{p}\right)}\right)^p,\ c_2(h) = \left(\frac{v}{2}\right)^h \frac{\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v}{2}+h\right)}, \tag{3-19}$$

其中，$v = n_1 + n_2 - 2$，$\Gamma(\bullet)$是伽马函数。将（3-19）式代入（3-5）式中，在 Log-Euclidean 平方损失函数下，计算可得最优收缩参数为

$$\hat{\alpha}_2^* = \max\{0, \min\{1, \hat{\alpha}\}\}, \tag{3-20}$$

其中，$\hat{\alpha} = \frac{\sum_{j=1}^{p} E(A+hB)(logc_2(h)+hD)}{\sum_{j=1}^{p} E(A+hB)^2}$，$A = \log c_2(h) - \log c_1(h)$，$B = \log s_j^2 - \log t$，$D = \log s_j^2 - \log \sigma_j^2$。此时，对应的$\sigma_j^{2h}$的最优几何型收缩估计为

$$\hat{\sigma}_j^{2h} = (c_1(h)t^{2h})^{\hat{\alpha}_2^*}(c_2(h)s_j^{2h})^{1-\hat{\alpha}_2^*}. \tag{3-21}$$

在双样本情形下，针对假设（3-15）提出的 GSDT 检验统计量为

$$\begin{aligned} T_{GSDT2}^2(\hat{\alpha}_2^*) &= \{n_1 n_2/(n_1+n_2)\}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{2}}(\hat{\alpha}_2^*)(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}) \\ &= \{n_1 n_2/(n_1+n_2)\} \sum_{j=1}^{p} (\bar{X}_j - \bar{Y}_j)^2 \hat{\sigma}_j^{-2}(\hat{\alpha}_2^*), \end{aligned} \tag{3-22}$$

其中，$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{2}}(\hat{\alpha}_2^*) = \mathrm{diag}(\hat{\sigma}_1^{-2}(\hat{\alpha}_2^*), \dots, \hat{\sigma}_p^{-2}(\hat{\alpha}_2^*))$。对于该检验统计量的渐近分布在第 3.3 节中进行分析。

在多元均值检验中，方差估计量在 GSDT 检验统计量的分母中，因此本文主要估计$\sigma_j^{-2}=1/\sigma_j^2$，即接下来考虑$h=-1$时$\hat{\sigma}_j^{2h}$的值（$j=1,2,\ldots,p$）。

3.3 GSDT 统计量的零分布

当样本量n较大时，GSDT 统计量的近似零分布可能是卡方分布。Wu, Genton 和 Stefanski（2006）分别研究了$p<n$和$p\geq n$条件下，对角 Hotelling T^2检验统计量的近似零分布为卡方分布的情况。本文考虑当样本量n较小时，在$p\to\infty$条件下，近似零分布为卡方分布情况。

为了得到 GSDT 检验统计量的近似零分布，需要先考虑$T_{GSDT1}^2(\alpha)$和$T_{GSDT2}^2(\alpha)$的均值和方差。

推论 3.3.1 对于任意的$v=n-1>4$和$\alpha\in(0,1]$，在H_0条件下，$T_{GSDT1}^2(\alpha)$的均值和方差分别为

$$E\{T_{GSDT1}^2(\alpha)\}=D_1\sigma_{1,pool}^{-\alpha}\sum_{j=1}^{p}{\sigma_j}^{2\alpha}, \tag{3-23}$$

$$Var\{T_{GSDT1}^2(\alpha)\}=(3D_2-D_3)\sigma_{1,pool}^{-2\alpha}\sum_{j=1}^{p}{\sigma_j}^{4\alpha}+(D_3-D_1^2)\sigma_{1,pool}^{-2\alpha}\left(\sum_{j=1}^{p}{\sigma_j}^{2\alpha}\right)^2, \tag{3-24}$$

其中，$D_1=\frac{c_1^{\alpha}(-1)\,c_2^{1-\alpha}(-1)}{c_2^{p-1}(-\alpha/p)c_2\{-\alpha/p-(1-\alpha)\}}$,

$D_2=\frac{c_1^{2\alpha}(-1)\,c_2^{2(1-\alpha)}(-1)}{c_2^{p-1}(-2\alpha/p)c_2\{-2\alpha/p-2(1-\alpha)\}}$,

$D_3=\frac{c_1^{2\alpha}(-1)\,c_2^{2(1-\alpha)}(-1)}{c_2^{p-2}(-2\alpha/p)\,c_2^2\{-2\alpha/p-(1-\alpha)\}}$,

$\sigma_{1,pool}=\left(\prod_{j=1}^{p}\sigma_j^2\right)^{1/p}$.

证明：对任意的非零$t>-v/2$，由 Tong 和 Wang（2007）可知

$$\mathrm{E}\left(\mathbb{s}_j^{2h}\right)=\sigma_j^{2h}/c_2(h),\ j=1,2,\ldots,p,$$

则，$E\{\hat{\sigma}_j^{-2}(\alpha)\}=E\left[\{c_1(-1)\mathbb{t}^{-2}\}^{\alpha}\{c_2(-1)\mathbb{s}_j^{-2}\}^{1-\alpha}\right]$

$$=c_1^{\alpha}(-1)\,c_2^{1-\alpha}(-1)E\left\{\mathbb{s}_1^{-2\alpha/p}\cdot\ldots\cdot\mathbb{s}_j^{-2\alpha/p-2(1-\alpha)}\cdot\ldots\cdot\mathbb{s}_p^{-2\alpha/p}\right\}$$

$$=c_1^{\alpha}(-1)\,c_2^{1-\alpha}(-1)E\left(\mathbb{s}_1^{-2\alpha/p}\right)\cdot\ldots\cdot E\left(\mathbb{s}_j^{-2\alpha/p-2(1-\alpha)}\right)\cdot\ldots\cdot E\left(\mathbb{s}_p^{-2\alpha/p}\right)$$

$$= c_1^{\alpha}(-1)\, c_2^{1-\alpha}(-1)\frac{\sigma_1^{-2\alpha/p}}{c_2(-\alpha/p)}\cdot\ldots\cdot\frac{\sigma_j^{-2\alpha/p-2(1-\alpha)}}{c_2\left(-\alpha/p-(1-\alpha)\right)}\cdot\ldots\cdot\frac{\sigma_p^{-2\alpha/p}}{c_2(-\alpha/p)}$$

$$= D_1\sigma_{1,pool}^{-\alpha}\sigma_j^{-2(1-\alpha)}.$$

又因为$\bar{X}_j$和σ_j^2是相互独立的，所以有

$$E\{T_{GSDT1}^2(\alpha)\} = nE\left\{\sum_{j=1}^{p}\left(\bar{X}_j-\mu_{0j}\right)^2\hat{\sigma}_j^{-2}(\alpha)\right\}$$

$$= n\sum_{j=1}^{p}\frac{\sigma_j^2}{n}D_1\sigma_{1,pool}^{-\alpha}\sigma_j^{-2(1-\alpha)}$$

$$= D_1\sigma_{1,pool}^{-\alpha}\sum_{j=1}^{p}\sigma_j^{2\alpha}.$$

若要计算$T_{GSDT1}^2(\alpha)$的方差，需要先计算$E\{T_{GSDT1}^2(\alpha)\}^2$，下面对$E\{T_{GSDT1}^2(\alpha)\}^2$进行计算，对于任意的$i\neq k$，有

$E\left\{\hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\right\}$

$= E\left[\{c_1(-1)\mathbb{t}^{-2}\}^{\alpha}\{c_2(-1)\mathbb{s}_j^{-2}\}^{1-\alpha}\{c_1(-1)\mathbb{t}^{-2}\}^{\alpha}\{c_2(-1)\mathbb{s}_k^{-2}\}^{1-\alpha}\right]$

$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left\{\mathbb{s}_1^{-4\alpha/p}\cdot\ldots\cdot\mathbb{s}_j^{-4\alpha/p-2(1-\alpha)}\cdot\ldots\cdot\mathbb{s}_k^{-4\alpha/p-2(1-\alpha)}\cdot\ldots\cdot\mathbb{s}_p^{-4\alpha/p}\right\}$

$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\, E\left(\mathbb{s}_1^{-4\alpha/p}\right)\cdot\ldots\cdot E\left(\mathbb{s}_j^{-4\alpha/p-2(1-\alpha)}\right)$

$\cdot\ldots\cdot E\left(\mathbb{s}_k^{-4\alpha/p-2(1-\alpha)}\right)\cdot\ldots\cdot E\left(\mathbb{s}_p^{-4\alpha/p}\right)$

$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\frac{\sigma_1^{-4\alpha/p}}{c_2(-2\,\alpha/p)}\cdot\ldots\cdot\frac{\sigma_j^{-4\alpha/p-2(1-\alpha)}}{c_2\left(-2\alpha/p-(1-\alpha)\right)}$

$\cdot\ldots\cdot\frac{\sigma_k^{-4\alpha/p-2(1-\alpha)}}{c_2\left(-2\alpha/p-(1-\alpha)\right)}\cdot\ldots\cdot\frac{\sigma_p^{-4\alpha/p}}{c_2(-2\,\alpha/p)}$

$= D_3\sigma_{1,pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)}\sigma_k^{-2(1-\alpha)}$,

$E\left\{\hat{\sigma}_j^{-4}(\alpha)\right\} = E\left[\{c_1(-1)\mathbb{t}^{-2}\}^{2\alpha}\{c_2(-1)\mathbb{s}_j^{-2}\}^{2(1-\alpha)}\right]$

$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left\{\mathbb{s}_1^{-4\alpha/p}\cdot\ldots\cdot\mathbb{s}_j^{-4\alpha/p-4(1-\alpha)}\cdot\ldots\cdot\mathbb{s}_p^{-4\alpha/p}\right\}$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left(\mathbb{s}_1^{-4\alpha/p}\right)\cdot\cdots\cdot E\left(\mathbb{s}_j^{-4\alpha/p-4(1-\alpha)}\right)\cdot\cdots\cdot E\left(\mathbb{s}_p^{-4\alpha/p}\right)$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\frac{\sigma_1^{-4\alpha/p}}{c_2(-2\alpha/p)}\cdot\cdots\cdot\frac{\sigma_j^{-4\alpha/p-4(1-\alpha)}}{c_2\left(-2\alpha/p-2(1-\alpha)\right)}$$

$$\cdot\cdots\cdot\frac{\sigma_p^{-4\alpha/p}}{c_2(-2\alpha/p)}$$

$$= D_2\sigma_{1,pool}^{-2\alpha}\sigma_j^{-4(1-\alpha)}.$$

又知，$E\left(\overline{\boldsymbol{X}}_j-\boldsymbol{\mu}_{0j}\right)^4=3\sigma_j^4/n^2$，则

$$\mathrm{E}\{T_{GSDT1}^2(\alpha)\}^2 = n^2E\left\{\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\bar{X}_j-\mu_{0j}\right)^2\ (\bar{X}_k-\mu_{0k})^2\ \hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\right\}$$

$$= n^2\sum_{j=1}^{p}E\left(\bar{X}_j-\mu_{0j}\right)^4\ E\left\{\hat{\sigma}_j^{-2}(\alpha)\right\}$$

$$+n^2\sum_{j\neq k}E\left(\bar{X}_j-\mu_{0j}\right)^2E(\bar{X}_k-\mu_{0k})^2E\left\{\hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\right\}$$

$$= 3D_2\sigma_{1,pool}^{-2\alpha}\sum_{j=1}^{p}{\sigma_j}^{4\alpha}+D_3\sigma_{1,pool}^{-2\alpha}\sum_{j\neq k}{\sigma_j}^{2\alpha}{\sigma_k}^{2\alpha}.$$

故可得$T_{GSDT1}^2(\alpha)$的方差为

$$Var\{T_{GSDT1}^2(\alpha)\} = E\{T_{GSDT1}^2(\alpha)\}^2 - \left[E\{T_{GSDT1}^2(\alpha)\}\right]^2$$

$$= (3D_2-D_3)\sigma_{1,pool}^{-2\alpha}\sum_{j=1}^{p}{\sigma_j}^{4\alpha}+(D_3-D_1^2)\sigma_{1,pool}^{-2\alpha}\left(\sum_{j=1}^{p}{\sigma_j}^{2\alpha}\right)^2.$$

这就完成了推论 3.3.1 的证明。

推论 3.3.2 对于任意的$v=n_1+n_2-2>4$和$\alpha\in(0,1]$，在H_0条件下，$T_{GSDT2}^2(\alpha)$的均值和方差分别为

$$E\{T_{GSDT2}^2(\alpha)\}=D_1\sigma_{2,pool}^{-2\alpha}\textstyle\sum_{j=1}^{p}{\sigma_j}^{2\alpha},\qquad(3\text{-}25)$$

$$Var\{T^2_{GSDT2}(\alpha)\} = (3D_2 - D_3)\sigma^{-4\alpha}_{2,pool}\sum_{j=1}^{p}{\sigma_j}^{4\alpha} + (D_3 - D_1^2)\sigma^{-4\alpha}_{2,pool}\left(\sum_{j=1}^{p}{\sigma_j}^{2\alpha}\right)^2, \quad (3\text{-}26)$$

其中，$D_1 = \frac{c_1^{\alpha}(-1)\, c_2^{1-\alpha}(-1)}{c_2^{p-1}(-\alpha/p)c_2\{-\alpha/p-(1-\alpha)\}}$,

$D_2 = \frac{c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)}{c_2^{p-1}(-2\alpha/p)c_2\{-2\alpha/p-2(1-\alpha)\}}$,

$D_3 = \frac{c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)}{c_2^{p-2}(-2\alpha/p)\, c_2^2\{-2\alpha/p-(1-\alpha)\}}$,

$\sigma_{2,pool} = \left(\prod_{j=1}^{p}{\sigma_j}^2\right)^{1/p}$.

证明：对任意的非零$t > -v/2$，由 Tong 和 Wang（2007）可知

$$\mathrm{E}\left(s_j^{2h}\right) = \sigma_j^{2h}/c_2(h)\,,\ j = 1,2,\dots,p,$$

则，$E\{\hat{\sigma}_j^{-2}(\alpha)\} = E\left[\{c_1(-1)t^{-2}\}^{\alpha}\{c_2(-1)s_j^{-2}\}^{1-\alpha}\right]$

$$= c_1^{\alpha}(-1)\, c_2^{1-\alpha}(-1)E\left\{s_1^{-2\alpha/p}\cdot\dots\cdot s_j^{-2\alpha/p-2(1-\alpha)}\cdot\dots\cdot s_p^{-2\alpha/p}\right\}$$

$$= c_1^{\alpha}(-1)\, c_2^{1-\alpha}(-1)E\left(s_1^{-2\alpha/p}\right)\cdot\dots\cdot E\left(s_j^{-2\alpha/p-2(1-\alpha)}\right)\cdot\dots\cdot E\left(s_p^{-2\alpha/p}\right)$$

$$= c_1^{\alpha}(-1)\, c_2^{1-\alpha}(-1)\frac{\sigma_1^{-2\alpha/p}}{c_2(-\alpha/p)}\cdot\dots\cdot\frac{\sigma_j^{-2\alpha/p-2(1-\alpha)}}{c_2\left(-\alpha/p-(1-\alpha)\right)}\cdot\dots\cdot\frac{\sigma_p^{-2\alpha/p}}{c_2(-\alpha/p)}$$

$$= D_1\sigma_{2,pool}^{-\alpha}\sigma_j^{-2(1-\alpha)}.$$

又因为$\bar{X}_j$，$\bar{Y}_j$和σ_j^2是相互独立的，所以有

$$E\{T^2_{GSDT2}(\alpha)\} = \{n_1n_2/(n_1+n_2)\}E\left\{\sum_{j=1}^{p}\left(\bar{X}_j - \bar{Y}_j\right)^2\hat{\sigma}_j^{-2}(\alpha)\right\}$$

$$= \{n_1n_2/(n_1+n_2)\}\sum\nolimits_{j=1}^{p}\frac{\sigma_j^2}{n}D_1\sigma_{2,pool}^{-\alpha}\sigma_j^{-2(1-\alpha)}$$

$$= D_1\sigma_{2,pool}^{-\alpha}\sum_{j=1}^{p}{\sigma_j}^{2\alpha}.$$

因为$\mathrm{Var}\{T^2_{GSDT2}(\alpha)\} = E\{T^2_{GSDT2}(\alpha)\}^2 - [E\{T^2_{GSDT2}(\alpha)\}]^2$，所以若要计算$T^2_{GSDT2}(\alpha)$的方差，需要先计算$E\{T^2_{GSDT2}(\alpha)\}^2$，下面对$E\{T^2_{GSDT2}(\alpha)\}^2$进行计算。对于任意的$i \neq k$，有

$$E\{\hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\}$$

$$= E\left[\{c_1(-1)t^{-2}\}^{\alpha}\{c_2(-1)s_j^{-2}\}^{1-\alpha}\{c_1(-1)t^{-2}\}^{\alpha}\{c_2(-1)s_k^{-2}\}^{1-\alpha}\right]$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left\{s_1^{-4\alpha/p}\cdot\ldots\cdot s_j^{-4\alpha/p-2(1-\alpha)}\cdot\ldots\cdot s_k^{-4\alpha/p-2(1-\alpha)}\cdot\ldots\cdot s_p^{-4\alpha/p}\right\}$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\, E\left(s_1^{-4\alpha/p}\right)$$

$$\cdot\ldots\cdot E\left(s_j^{-4\alpha/p-2(1-\alpha)}\right)\cdot\ldots\cdot E\left(s_k^{-4\alpha/p-2(1-\alpha)}\right)\cdot\ldots\cdot E\left(s_p^{-4\alpha/p}\right)$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\frac{\sigma_1^{-4\alpha/p}}{c_2(-2\,\alpha/p)}\cdot\ldots\cdot\frac{\sigma_j^{-4\alpha/p-2(1-\alpha)}}{c_2\left(-2\alpha/p-(1-\alpha)\right)}$$

$$\cdot\ldots\cdot\frac{\sigma_k^{-4\alpha/p-2(1-\alpha)}}{c_2\left(-2\alpha/p-(1-\alpha)\right)}\cdot\ldots\cdot\frac{\sigma_p^{-4\alpha/p}}{c_2(-2\,\alpha/p)}$$

$$= D_3\sigma_{2,pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)}\sigma_k^{-2(1-\alpha)}.$$

当$i=j$时，有

$$E\{\hat{\sigma}_j^{-4}(\alpha)\} = E\left[\{c_1(-1)t^{-2}\}^{2\alpha}\{c_2(-1)s_j^{-2}\}^{2(1-\alpha)}\right]$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left\{s_1^{-4\alpha/p}\cdot\ldots\cdot s_j^{-4\alpha/p-4(1-\alpha)}\cdot\ldots\cdot \mathbb{s}_p^{-4\alpha/p}\right\}$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)E\left(\mathbb{s}_1^{-4\alpha/p}\right)\cdot\ldots\cdot E\left(\mathbb{s}_j^{-4\alpha/p-4(1-\alpha)}\right)\cdot\ldots\cdot E\left(\mathbb{s}_p^{-4\alpha/p}\right)$$

$$= c_1^{2\alpha}(-1)\, c_2^{2(1-\alpha)}(-1)\frac{\sigma_1^{-4\alpha/p}}{c_2(-2\alpha/p)}\cdot\ldots\cdot\frac{\sigma_j^{-4\alpha/p-4(1-\alpha)}}{c_2\left(-2\alpha/p-2(1-\alpha)\right)}$$

$$\cdot\ldots\cdot\frac{\sigma_p^{-4\alpha/p}}{c_2(-2\alpha/p)}$$

$$= D_2\sigma_{2,pool}^{-2\alpha}\sigma_j^{-4(1-\alpha)}.$$

又因为，$E\left(\bar{X}_j-\bar{Y}_j\right)^4 = 3\sigma_j^4/n^2$，则

$$\mathrm{E}\{T_{GSDT2}^2(\alpha)\}^2 = \{n_1n_2/(n_1+n_2)\}^2E\left\{\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\bar{X}_j-\bar{Y}_j\right)^2\ (\bar{X}_k-\bar{Y}_k)^2\ \hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\right\}$$

$$= \{n_1n_2/(n_1+n_2)\}^2\sum_{j=1}^{p}E\left(\bar{X}_j-\bar{Y}_j\right)^4\ E\{\hat{\sigma}_j^{-4}(\alpha)\}$$

$$+\{n_1n_2/(n_1+n_2)\}^2\sum_{j\neq k}E\left(\bar{X}_j-\bar{Y}_j\right)^2E(\bar{X}_k-\bar{Y}_k)^2E\left\{\hat{\sigma}_j^{-2}(\alpha)\hat{\sigma}_k^{-2}(\alpha)\right\}$$

$$=3D_2\sigma_{2,pool}^{-2\alpha}\sum_{j=1}^{p}{\sigma_j}^{4\alpha}+D_3\sigma_{2,pool}^{-2\alpha}\sum_{j\neq k}{\sigma_j}^{2\alpha}{\sigma_k}^{2\alpha}.$$

故可得$T_{GSDT2}^2(\alpha)$的方差为

$$Var\{T_{GSDT2}^2(\alpha)\}=\ \mathrm{E}\{T_{GSDT2}^2(\alpha)\}^2-\ [E\{T_{GSDT2}^2(\alpha)\}]^2$$

$$=(3D_2-D_3)\sigma_{2,pool}^{-2\alpha}\sum_{j=1}^{p}{\sigma_j}^{4\alpha}+(D_3-D_1^2)\sigma_{2,pool}^{-2\alpha}\left(\sum_{j=1}^{p}{\sigma_j}^{2\alpha}\right)^2.$$

推论 3.3.2 得证。

对于较小的样本维数p，卡方分布是一个良好的近似零分布。在本节中，设所提出的检验统计量近似服从缩放的卡方分布$b\chi_d^2$，其中$b_i\chi_{di}^2$（$i=1,2$）的均值和方差分别等于$T_{GSDTi}^2(\alpha_i^*)$的均值和方差，具体表示为

$$E\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}=b_1d_1,\ Var\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}=2b_1^2d_1;\qquad(3\text{-}27)$$

$$E\{T_{GSDT2}^2(\hat{\alpha}_2^*)\}=b_2d_2,\ Var\{T_{GSDT2}^2(\hat{\alpha}_2^*)\}=2b_2^2d_2.\qquad(3\text{-}28)$$

定理 3.3.1 对于任意的$v=n-1>4$和最优收缩参数$\hat{\alpha}_1^*$，在零假设条件下，有

$$T_{GSDT1}^2(\hat{\alpha}_1^*)\sim b_1\chi_{d_1}^2,\qquad(3\text{-}29)$$

其中，$b_1=\frac{(3D_2-D_3)\sigma_{1,pool}^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}_1^*}+(D_3-D_1^2)\sigma_{1,pool}^{-2\hat{\alpha}_1^*}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}_1^*}\right)^2}{2D_1\sigma_{1,pool}^{-\hat{\alpha}_1^*}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}_1^*}}$,

$$d_1=\frac{2D_1^2\sigma_{1,pool}^{-2\hat{\alpha}_1^*}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}_1^*}\right)^2}{(3D_2-D_3)\sigma_{1,pool}^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}{\sigma_j}^{4\hat{\alpha}_1^*}+(D_3-D_1^2)\sigma_{1,pool}^{-2\hat{\alpha}_1^*}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}_1^*}\right)^2}.$$

证明：由$E\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}=b_1d_1$，$Var\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}=2b_1^2d_1$，知

$$b_1=\frac{Var\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}}{2E\{T_{GSDT1}^2(\hat{\alpha}_1^*)\}},$$

$$d_1 = \frac{2\left[E\{T^2_{GSDT1}(\hat{\alpha}^*_1)\}\right]^2}{Var\{T^2_{GSDT1}(\hat{\alpha}^*_1)\}},$$

即，$b_1 = \frac{(3D_2-D_3)\sigma_{1,pool}^{-2\hat{\alpha}^*_1}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}+(D_3-D_1^2)\sigma_{1,pool}^{-2\hat{\alpha}^*_1}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}\right)^2}{2D_1\sigma_{1,pool}^{-\hat{\alpha}^*_1}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}}$,

$$d_1 = \frac{2D_1^2\sigma_{1,pool}^{-2\hat{\alpha}^*_1}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}\right)^2}{(3D_2-D_3)\sigma_{1,pool}^{-2\hat{\alpha}^*_1}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}+(D_3-D_1^2)\sigma_{1,pool}^{-2\hat{\alpha}^*_1}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_1}\right)^2}.$$

定理 3.3.1 得证。

定理 3.3.2 对任意的$v = n_1 + n_2 - 2 > 4$和最优收缩参数$\hat{\alpha}^*_2$，在零假设条件下，有

$$T^2_{GSDT2}(\hat{\alpha}^*_1) \sim b_2\chi^2_{d_2}, \qquad (3\text{-}30)$$

其中，$b_2 = \frac{(3D_2-D_3)\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}+(D_3-D_1^2)\sigma_{2,pool}^{-2\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}{2D_1\sigma_{2,pool}^{-\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}}$,

$$d_2 = \frac{2D_1^2\sigma_{2,pool}^{-2\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}{(3D_2-D_3)\sigma_{2,pool}^{-2\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{4\hat{\alpha}^*_2}+(D_3-D_1^2)\sigma_{2,pool}^{-2\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}.$$

证明：由$E\{T^2_{GSDT2}(\hat{\alpha}^*_2)\} = b_2d_2$，$Var\{T^2_{GSDT2}(\hat{\alpha}^*_2)\} = 2b_2^2d_2$，知

$$b_2 = \frac{Var\{T^2_{GSDT2}(\hat{\alpha}^*_2)\}}{2E\{T^2_{GSDT2}(\hat{\alpha}^*_2)\}},$$

$$d_2 = \frac{2\left[E\{T^2_{GSDT2}(\hat{\alpha}^*_2)\}\right]^2}{Var\{T^2_{GSDT2}(\hat{\alpha}^*_2)\}},$$

即，$b_2 = \frac{(3D_2-D_3)\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{4\hat{\alpha}^*_2}+(D_3-D_1^2)\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}{2D_1\sigma_{2,pool}^{-2\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}}$,

$$d_2 = \frac{2D_1^2\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}{(3D_2-D_3)\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\sum_{j=1}^{p}{\sigma_j}^{4\hat{\alpha}^*_2}+(D_3-D_1^2)\sigma_{2,pool}^{-4\hat{\alpha}^*_2}\left(\sum_{j=1}^{p}{\sigma_j}^{2\hat{\alpha}^*_2}\right)^2}.$$

定理 3.3.2 得证。

定理 3.3.1 和定理 3.3.2 中的b_1，d_1，b_2和d_2中包含一些未知量，以b_1，d_1为例，$b_{11}(\boldsymbol{\sigma^2}) = \sigma_{1,pool}^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}\sigma_j{}^{2\hat{\alpha}_1^*}$和$b_{12}(\boldsymbol{\sigma^2}) = \sigma_{1,pool}^{-4\hat{\alpha}_1^*}\sum_{j=1}^{p}\sigma_j{}^{4\hat{\alpha}_1^*}$是未知的。对于这些未知量可以根据不同的情况来进行估计，

（1）对于固定的p和“大”n，$\mathbb{s}_j^2 \xrightarrow{a.s.} \sigma_j^2$，$n \to \infty$，有

$$\hat{b}_{11}(\boldsymbol{\sigma^2}) = \mathbb{t}^{-\hat{\alpha}_1^*}\sum_{j=1}^{p}\mathbb{s}_j^{2\hat{\alpha}_1^*},$$

$$\hat{b}_{12}(\boldsymbol{\sigma^2}) = \mathbb{t}^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}\mathbb{s}_j^{4\hat{\alpha}_1^*}.$$

（2）对于固定的n和“大”p，根据 Tong 和 Wang（2007）的引理 2，有

$$\check{b}_{11}(\boldsymbol{\sigma^2}) = w(\hat{\alpha}_1^*)\mathbb{t}^{-\hat{\alpha}_1^*}\sum_{j=1}^{p}\mathbb{s}_j^{2\hat{\alpha}_1^*},$$

$$\check{b}_{12}(\boldsymbol{\sigma^2}) = w(2\hat{\alpha}_1^*)\mathbb{t}^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}\mathbb{s}_j^{4\hat{\alpha}_1^*},$$

其中，$w(\alpha) = (v/2)^{-\alpha}c_2(h)exp\big(\alpha\Psi(v/2)\big)$，$\Psi(t) = \Gamma'(t)/\Gamma(t)$。

（3）对于$b_{11}(\boldsymbol{\sigma^2})$和$b_{12}(\boldsymbol{\sigma^2})$中的$\sigma_j^2$用最优收缩估计$\hat{\sigma}_j^2(\hat{\alpha}_1^*)$代替，有

$$\tilde{b}_{11}(\boldsymbol{\sigma^2}) = t_1^{-\hat{\alpha}_1^*}\sum_{j=1}^{p}\hat{\sigma}_j^{2\hat{\alpha}_1^*}(\hat{\alpha}_1^*),$$

$$\tilde{b}_{12}(\boldsymbol{\sigma^2}) = t_1^{-2\hat{\alpha}_1^*}\sum_{j=1}^{p}\hat{\sigma}_j^{4\hat{\alpha}_1^*}(\hat{\alpha}_1^*).$$

第 4 章　蒙特卡罗模拟

本章通过用 R 软件进行蒙特卡罗模拟研究来比较 SD 对角检验、CQ 未缩放的检验、CPPW 正则化检验、 HTG 对角似然检验和 GSDT 检验的检验功效。通过进行大量的模拟，给出模拟所得数据，通过数据对模拟的结果比较这些检验方法的功效表现情况。在模拟中，我们考虑了三种常用的总体相关系数矩阵用来确定总体协方差矩阵，以便对应实际数据的不同情况。

为了比较上述五种检验，我们需要定义检验第一类错误率和经验功效。令$\xi_{1-\alpha}$表示在零假设条件下统计量渐近分布的$100(1-\alpha)\%$的分位点。我们在零假设条件下进行 M 次模拟，那么经验第一类错误率为

$$\hat{\alpha} = \frac{N_1(t_a \geq \xi_{1-\alpha})}{M},$$

其中，t_a表示检验统计量在模拟中的数值，$N_1(t_a \geq \xi_{1-\alpha})$表示$t_a \geq \xi_{1-\alpha}$的总次数，M 表示模拟总次数。为了比较经验功效需要选择功效的评判点，我们选择$M\alpha$值作为经验评判点，记为$\hat{\xi}_{1-\alpha}$。我们在备择假设条件下进行 M 次模拟，那么经验功效为

$$\hat{\beta} = \frac{N_2(t_b \geq \xi_{1-\alpha})}{M},$$

其中，t_a表示在备择假设下检验统计量在模拟中的数值，$N_2(t_b \geq \xi_{1-\alpha})$表示$t_b \geq \xi_{1-\alpha}$的总次数，M 表示模拟总次数。

通过计算第一类错误率和功效函数，我们可以分析各个检验统计量的优劣性。另外，我们绘制出各个检验统计量的接受者操作特征曲线（ROC）并计算曲线下面的面积的值(AUC)。接受者操作特征曲线（ROC）描述真阳性率（TPR）和假阳性率（FPR）的变化情况。通过比较观察曲线以及所对应的曲线下的面积，我们分析了各种检验统计量在各种不同情况下的表现。本章分为两节，第一节介绍模拟中符号和参数的设置，第二节列出模拟数据，并通过模拟结果对几种检验方法的功效进行评价。

4.1　模拟设计

为了记号方便，SD 对角检验、CQ 未缩放检验、CPPW 正则化检验、 HTG 对角似然检验和 GSDT 检验分别记为 SD、CQ、CPPW、HTG 和 GSDT。

对于计算各检验统计量的第一类错误率，设两组样本数据均来自多元正态分布$N_p(\mathbf{0},\mathbf{\Sigma})$。对于计算各检验统计量的功效函数，设一组样本数据来自多元正态分布$N_p(\mathbf{0},\mathbf{\Sigma})$，另一组样本数据由来自$N_p(\boldsymbol{\mu},\mathbf{\Sigma})$，其中对于均值$\boldsymbol{\mu}$，其前$p_0$个非零元素由$\mu_j = c\sigma_j^2$， $j = 1,\dots,p_0$生成，其余元素为$\mathbf{0}$，其中c是效应量，通过改变效应量c的值，可以模拟信号强弱的程度，σ_j^2由缩放的卡方分布$\chi_5^2/5$随机生成，p_0用于模拟中信号的稀疏程度。

我们比较各种不同的检验统计量在各种不同结构的协方差矩阵下的表现，注意到总体协方差矩阵可以表示为$\mathbf{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{R}\boldsymbol{D}^{1/2}$，其中方差阵$\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Sigma}) = \mathrm{diag}(\sigma_1^2,..,\sigma_p^2)$，$\boldsymbol{R}$是总体相关系数矩阵，其结构为

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R_\rho} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R_{-\rho}} & \mathbf{0} & \ddots & \vdots \\ \vdots & \mathbf{0} & \boldsymbol{R_\rho} & \mathbf{0} & \vdots \\ \vdots & \ddots & \mathbf{0} & \boldsymbol{R_{-\rho}} & \ddots \\ \mathbf{0} & \cdots & \cdots & \ddots & \ddots \end{pmatrix}$$

其中，$\boldsymbol{R_\rho}$是一个$q \times q$矩阵且q ≤ p。我们选取如下三种常用的总体相关系数矩阵来确定总体协方差矩阵，

1. 单位矩阵($\mathbf{\Sigma}_{\mathrm{ID}}$)；
2. 一阶自回归相关矩阵($\mathbf{\Sigma}_{\mathrm{AR}}$)：$\boldsymbol{R_\rho} = (r_{ij})_{q\times q}$，其中$r_{ij} = \rho^{|i-j|}$， $1 \le i,j < q$；
3. 复合对称相关矩阵($\mathbf{\Sigma}_{\mathrm{CS}}$)：$\boldsymbol{R_\rho} = (r_{ij})_{q\times q}$，其中$r_{ij} = \rho$， $1 \le i \ne j < q$，$r_{ij} = 1$，$i = j$。

在模拟中，设置两组数据的样本量相同，其取值分别为$n = n_1 = n_2 =$ 5，15 和 50；效应量取值分别为 c=0.55，0.45 和 0.35。对于不同的总体相关系数矩阵 $\boldsymbol{R}$，ρ的取值分别为 0，0.2 和 0.4，其中在复合对称结构$\mathbf{\Sigma_{CS}}$中，$\rho = \mp 0.4$会导致 $\boldsymbol{R}$ 成为非正定矩阵，对于这种情况，ρ的值只取 0.4。我们设置 p=50，q=5，显著水平α =0.05。第一类错误率和效用函数均由 1000 次模拟计算得出。

4.2 模拟结果

我们首先比较各个检验统计量在不同情况下的第一类错误率，其结果总结在表 4.1。在假设检验中，人们常常控制第一类错误率在某个显著水平。从表 4.1 中可以看出，在

样本量较大的情况下，SD 检验的第一类错误率要比 GSDT 检验略低，但是在样本量较小的情况下，SD 检验并不能很好的控制第一类错误率，GSDT 检验统计量的第一类错误率明显比 SD 检验统计量的第一类错误率小。SD 检验的第一类错误率比较受样本量的影响。当样本量较小时，CPPW 检验统计量和 GSDT 检验统计量比其他方法能更好的控制第一类错误率。当相关系数较小且样本量较小时， GSDT 检验统计量的第一类错误率比 CPPW 检验统计量的第一类错误率略小，当样本量较大时，总体上看，CPPW 检验的第一类错误率比 GSDT 检验略低。随着相关系数的增大，各检验统计量的第一类错误率均有所增加。从整体来看，GSDT 检验的第一类错误率比其他检验统计量略底。根据以上综合分析，GSDT 检验可以有效的控制第一类错误率。

其次，我们比较各个统计量在不同效应量情况下的效用函数，对于效应量较小的情况（0.35），其结果总结在表 4.2，表 4.3 和表 4.4 中。在假设检验中，功效函数用于代表检验统计量检测信号能力的强弱。从表 4.2，表 4.3 和表 4.4 中均可看出，在信号较弱的情况下，GSDT 检验统计量的功效函数均比 CQ 检验，CPPW 检验和 HTG 检验的功效函数大。CQ 检验是在 Hotelling T^2检验统计量中去除协方差矩阵，因此在协方差矩阵相关性逐渐增加的情况下，CQ 检验的表现比其他方法明显较差。在样本量较小的情况下，GSDT 检验统计量的功效函数比 SD 检验的功效函数小，但 SD 检验的第一类错误概率却高出很多，而且随着样本量的增加，GSDT 检验统计量的功效函数逐渐比 SD 检验统计量的功效函数大，对于 CPPW 检验统计量，在信号稀疏的情况下，其功效函数比起其他方法有一定优势。但随着信号变得密集，其表现逐渐变得没有其他方法好。对于 HTG 检验统计量，在样本量较小时，其无法控制第一类错误率，另外随着样本量增加，其功效函数比起其他方法也没有明显的提高。随着效应量的增加，各个检验统计量的功效函数均有所增加，在各种效应量的情况下，GSDT 检验统计量的功效函数也比其他检验方法好。根据模拟结果分析，GSDT 检验统计量比起其他检验统计量能更好的检测出信号，在样本量较小时，GSDT 检验能够较好的控制第一类错误率，并且在此情况下提供更好的效应函数。

表 4.1　在零假设条件下第一类错误率

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	3	0.076	0.429	0.085	0.083	0.09
		5	0.057	0.115	0.064	0.089	0.06
		15	0.052	0.046	0.051	0.085	0.046
		30	0.053	0.037	0.048	0.097	0.06
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	3	0.077	0.388	0.089	0.071	0.094
		5	0.05	0.0107	0.067	0.077	0.064
		15	0.051	0.039	0.054	0.073	0.045
		50	0.054	0.038	0.059	0.085	0.06
	0.4	3	0.074	0.395	0.111	0.055	0.09
		5	0.052	0.108	0.08	0.074	0.067
		15	0.047	0.045	0.072	0.078	0.049
		50	0.051	0.04	0.081	0.086	0.061
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.2	3	0.069	0.408	0.107	0.075	0.091
		5	0.051	0.113	0.071	0.091	0.064
		15	0.05	0.043	0.059	0.08	0.048
		50	0.052	0.036	0.061	0.091	0.059
	0.4	3	0.08	0.371	0.124	0.074	0.087
		5	0.06	0.108	0.106	0.086	0.076
		15	0.043	0.041	0.084	0.098	0.05
		50	0.05	0.043	0.111	0.096	0.06

表 4.2　在效应量为 c=0.35 并且p_0 = 5 下的功效函数

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	3	0.099	0.406	0.077	0.084	0.093
		5	0.082	0.132	0.074	0.1	0.077
		15	0.127	0.095	0.08	0.13	0.074
		50	0.402	0.0317	0.0291	0.161	0.306
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	3	0.105	0.367	0.081	0.073	0.098
		5	0.099	0.129	0.066	0.082	0.077
		15	0.123	0.091	0.089	0.108	0.085
		50	0.421	0.309	0.272	0.144	0.234
	0.4	3	0.122	0.393	0.078	0.068	0.096
		5	0.101	0.117	0.076	0.073	0.083
		15	0.144	0.092	0.086	0.118	0.078
		50	0.398	0.261	0.264	0.13	0.197
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.2	3	0.114	0.397	0.074	0.086	0.094
		5	0.084	0.126	0.078	0.089	0.085
		15	0.139	0.101	0.087	0.101	0.078
		50	0.399	0.285	0.259	0.139	0.181
	0.4	3	0.137	0.367	0.073	0.074	0.094
		5	0.133	0.124	0.083	0.102	0.094
		15	0.162	0.09	0.094	0.097	0.075
		50	0.402	0.226	0.227	0.112	0.151

表 4.3　在效应量为 c=0.35 并且p_0 = 25 下的功效函数

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	3	0.167	0.114	0.106	0.126	0.167
		5	0.204	0.16	0.171	0.153	0.204
		15	0.594	0.463	0.467	0.355	0.594
		50	1	0.995	0.966	0.995	1
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	3	0.167	0.114	0.1	0.133	0.167
		5	0.212	0.152	0.153	0.157	0.212
		15	0.608	0.443	0.448	0.364	0.608
		50	1	0.992	0.951	0.998	1
	0.4	3	0.179	0.113	0.088	0.129	0.179
		5	0.194	0.153	0.133	0.155	0.194
		15	0.586	0.401	0.386	0.368	0.586
		50	1	0.982	0.916	0.998	1
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.2	3	0.162	0.103	0.095	0.122	0.162
		5	0.202	0.143	0.158	0.16	0.202
		15	0.566	0.409	0.417	0.429	0.566
		50	1	0.989	0.941	1	1
	0.4	3	0.188	0.119	0.087	0.127	0.188
		5	0.238	0.142	0.125	0.142	0.238
		15	0.555	0.358	0.285	0.234	0.555
		50	0.988	0.949	0.797	0.816	0.988

表 4.4　在效应量为 c=0.35 并且p_0 = 50 下的功效函数

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	3	0.287	0.181	0.155	0.198	0.259
		5	0.422	0.311	0.35	0.325	0.396
		15	0.968	0.892	0.944	0.695	0.962
		50	1	1	1	1	1
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	3	0.288	0.175	0.16	0.199	0.267
		5	0.404	0.283	0.313	0.284	0.372
		15	0.961	0.864	0.937	0.717	0.946
		50	1	1	1	1	1
	0.4	3	0.31	0.163	0.147	0.182	0.277
		5	0.414	0.262	0.303	0.28	0.386
		15	0.937	0.822	0.905	0.748	0.932
		50	1	1	1	1	1
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.2	3	0.286	0.16	0.17	0.168	0.253
		5	0.403	0.282	0.326	0.277	0.372
		15	0.973	0.844	0.957	0.863	0.961
		50	1	1	1	1	1
	0.4	3	0.295	0.171	0.128	0.18	0.273
		5	0.433	0.251	0.227	0.228	0.4
		15	0.866	0.709	0.692	0.413	0.855
		50	1	1	1	0.99	1

为了更直观的观测各检验统计量的表现，我们描绘出各统计量在不同情况下的 ROC 曲线并计算其所对应的 AUC。此文中我们仅画出样本量 n=5，p_0= 5 时的 ROC 曲线，其结果在图 4.1，图 4.2 和图 4.3 中，并计算了样本量为 5 时的 AUC 值，其它样本量结果与其相似，在本文中不再展示。在效应量为 c=0.35 时，AUC 的结果总结在表 4.5 至表 4.7 中。在效应量为 c=0.45 时，AUC 的结果总结在表 4.8 至表 4.10。在效应量为 c=0.55 时，AUC 的结果总结在表 4.11 至表 4.13。

表 4.5　效应量为 c=0.35，样本量为 n=5，p_0= 5 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.5393	0.4956	0.4931	0.4934	0.5108
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.5292	0.5103	0.5029	0.4914	0.4884
	0.4	5	0.5321	0.5068	0.5068	0.5061	0.4924
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.5234	0.5105	0.4974	0.4919	0.5061

表 4.6　效应量为 c=0.35，样本量为 n=5，p_0= 25 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.6751	0.5597	0.5765	0.5719	0.5802
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.6623	0.5391	0.5749	0.5497	0.5659
	0.4	5	0.6560	0.5488	0.5587	0.5536	0.5496
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.6282	0.5403	0.5480	0.5296	0.5465

表 4.7　效应量为 c=0.35，样本量为 n=5，p_0= 50 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.8014	0.6692	0.6867	0.6982	0.6752
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.7862	0.6409	0.6726	0.6721	0.6629
	0.4	5	0.7765	0.6449	0.6472	0.6742	0.6389
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.7323	0.6226	0.6389	0.6186	0.6139

表 4.8　效应量为 c=0.45，样本量为 n=5，p_0= 5 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.5630	0.5113	0.5143	0.4888	0.5192
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.5496	0.5147	0.5094	0.4878	0.5176
	0.4	5	0.5523	0.5133	0.5179	0.5091	0.4876
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.5396	0.5171	0.4941	0.4885	0.5100

表 4.9　效应量为 c=0.45，样本量为 n=5，p_0=25 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.7637	0.6314	0.6469	0.6407	0.6433
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.7484	0.6008	0.6446	0.6138	0.6304
	0.4	5	0.7386	0.6133	0.6199	0.6179	0.6021
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.6979	0.5907	0.5992	0.5705	0.5899

表 4.10　效应量为 c=0.45，样本量为 n=5，p_0= 50 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.9042	0.8026	0.8177	0.8279	0.7946
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.89234	0.7876	0.8063	0.8212	0.7832
	0.4	5	0.8823	0.7804	0.7862	0.8135	0.7776
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.8284	0.7263	0.7351	0.7142	0.7027

表 4.11　效应量为 c=0.55，样本量为 n=5，p_0= 5 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.5918	0.5205	0.5253	0.5167	0.5314
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.5751	0.5221	0.5209	0.5154	0.5269
	0.4	5	0.5769	0.5206	0.5285	0.5113	0.5182
$\boldsymbol{\Sigma}_{\mathbf{CS}}$	0.4	5	0.5918	0.5205	0.5253	0.5167	0.5315

表 4.12　效应量为 c=0.55，样本量为 n=5， p_0= 25 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.8469	0.7199	0.7388	0.7175	0.7238
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.8323	0.6982	0.7284	0.7001	0.7129
	0.4	5	0.8197	0.7030	0.7016	0.7019	0.6844
$\boldsymbol{\Sigma}_{\mathrm{CS}}$	0.4	5	0.7695	0.6561	0.6662	0.6238	0.6446

表 4.13　效应量为 c=0.55，样本量为 n=5， p_0= 50 条件下的 AUC

$\boldsymbol{\Sigma}$	ρ	n	GSDT	SD	CQ	CPPW	HTG
$\boldsymbol{\Sigma}_{\mathrm{ID}}$	0	5	0.9656	0.9165	0.9239	0.9285	0.9077
$\boldsymbol{\Sigma}_{\mathrm{AR}}$	0.2	5	0.9587	0.9071	0.9088	0.9246	0.8925
	0.4	5	0.9515	0.8957	0.8991	0.9196	0.8902
$\boldsymbol{\Sigma}_{\mathrm{CS}}$	0.4	5	0.9048	0.8302	0.8358	0.8193	0.7935

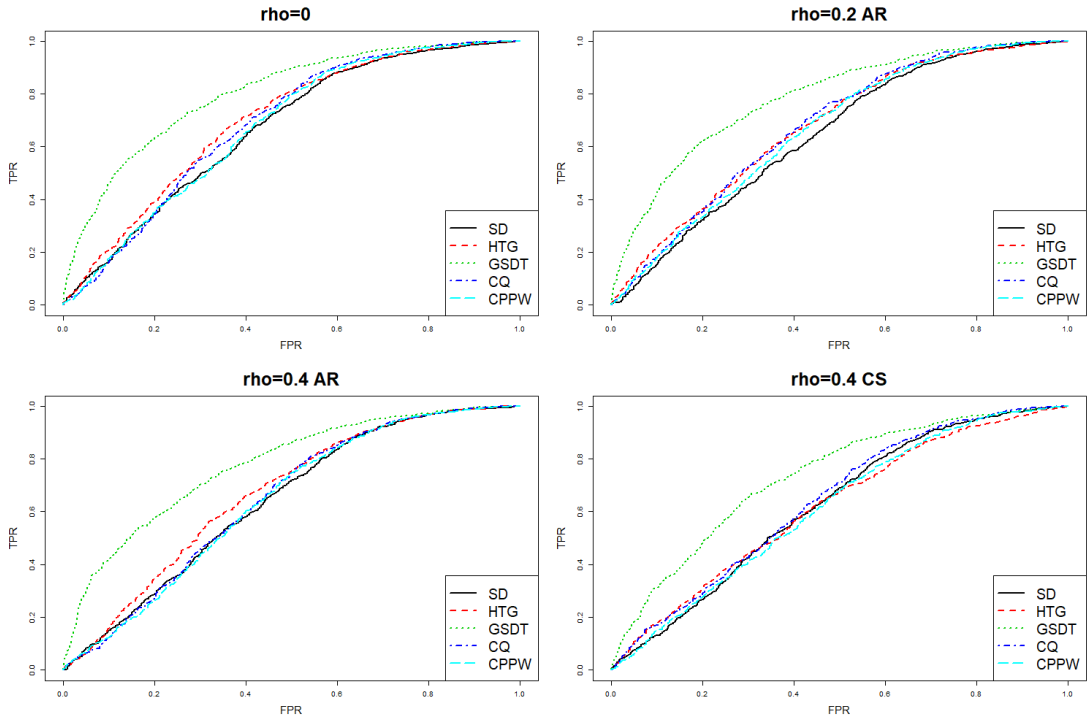

图 4.1　效应量为 c=0.35，样本量为 n=5， p_0= 50 条件下的 AUC

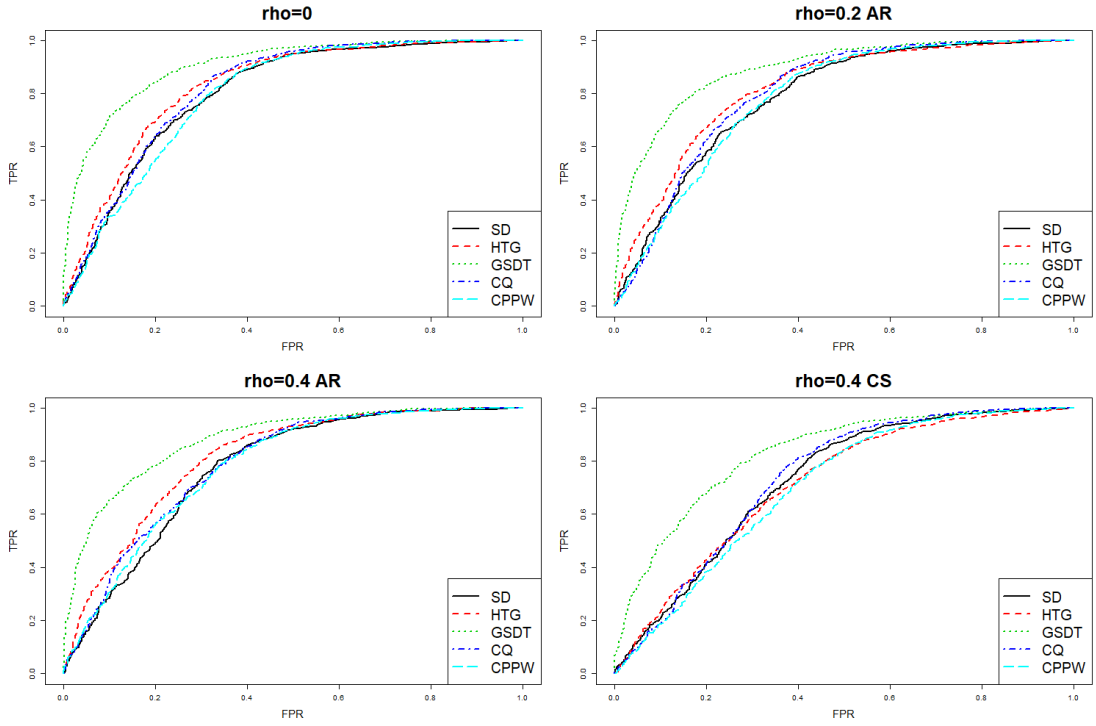

图 4.2　效应量为 c=0.45，样本量为 n=5，p_0= 50 条件下的 AUC

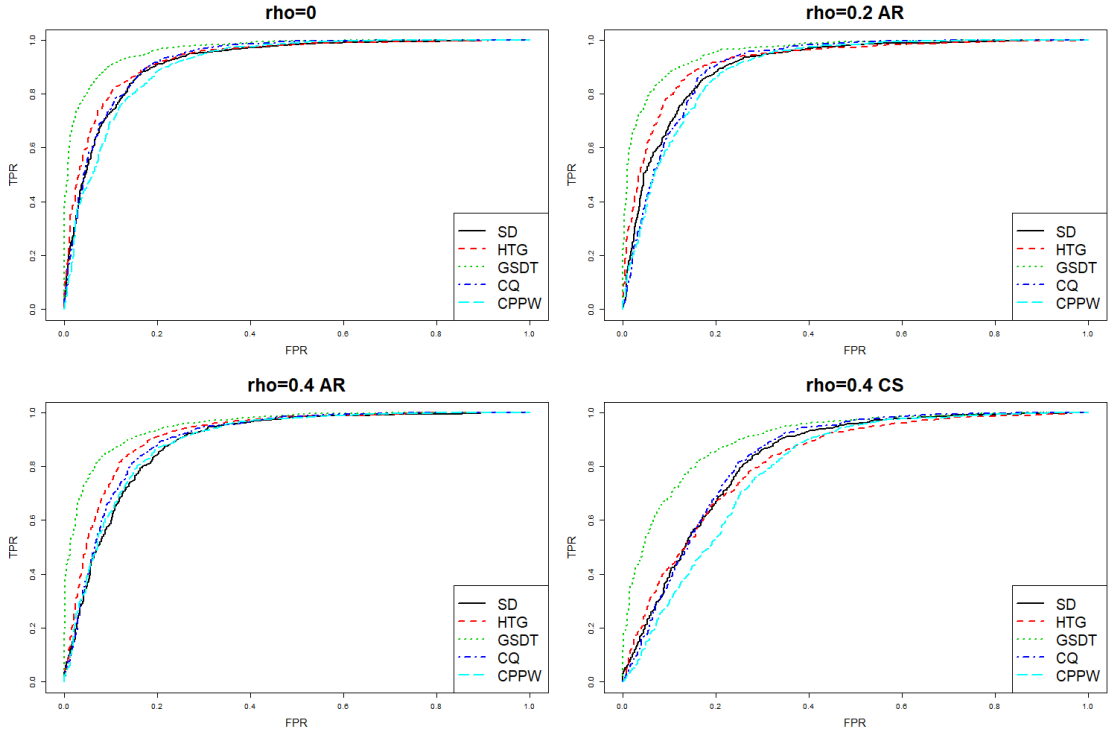

图 4.3　效应量为 c=0.55，样本量为 n=5，p_0= 50 条件下的 AUC

ROC 及其对应的 AUC 用于描述一个检验统计量真阳性率（TPR）和假阳性率（FPR）的相对应的变化情况。一个检验统计量真阳性率越高，假阳性率越低，该检验统计量越好。从图 4.1、图 4.2 和图 4.3 中我们可以看出，随着效应量 c 的增大，即随着信号强度增强，各种检验方法的检验能力也随之增强，同时各个检验方法检验能力也相差更大。特别地，在不同的效应量 c 下，GSDT 检验统计量均优于其他检验统计量。

在信号强度较弱且信号相对稀疏的情况下（表 4.5），GSDT 检验统计量的表现与其他检验统计量接近，但仍优于其他检验统计量。随着信号越来越稠密（表 4.6 和表 4.7），GSDT 检验统计量显著的优于其他统计量。在信号强度较强且信号相对稀疏的情况下（表 4.11），GSDT 检验统计量的表现明显优于其他检验统计量。随着信号越来越稠密（表 4.12 和表 4.13），GSDT 检验统计量显著的优于其他统计量。另一方面,随着信号强度的减弱，GSDT 检验统计量的优良性越发明显。综上所述，在控制第一类错误率的情况下，GSDT 检验统计量的效用函数优于其他的检验统计量。

第 5 章 实证分析

在本章中，我们研究各个检验统计方法在真实基因表达数据中的表现。通常的基因表达数据中，样本个数常常远远小于变量的个数，因此传统的 Hotelling T^2检验无法有效地进行统计检验，我们比较各个改进的 Hotelling T^2检验在此类中两种数据中的表现。

5.1 脑癌数据中的多元均值检验

首先我们研究各个检验统计方法在检验脑癌数据中的表现，该脑癌数据来自 The Cancer Genome Atlas （TCGA）数据库。多形性胶质母细胞瘤是最常见的致命性脑部肿瘤，科学家们通过研究长期生存的病人和短期生存的病人之间基因拷贝数的差别，试图寻找治疗该脑部肿瘤的有效方法。我们所研究的数据包含 92 位长期存活的病人和 138 位短期存活的病人，这里长期存活定义为该病人自第一次诊断后存活两年以上。短期存活定义为该病人自第一次诊断后存活少于两年。通过使用基因微阵列技术，该数据记录病人染色体上各基因位置上拷贝数目。Gregory 等人（2015）将整个染色体划分为长度不同的片段。我们主要研究染色体 1 上 q-arm 上的片段，这个片段包含 400 个基因位置上的探针所记录的拷贝数。Olshen 等人(2004) 和 Baladandayuthapani 等人（2010）指出，不同组病人的拷贝数的变化会发生在多个探针处，而不是单一探针处。在比较不同组病人的拷贝数变化时会在多个探针处发生较小的拷贝数变化，而不是单一探针处发生明显的拷贝数变化。因此，我们使用多元均值检验方法来检测不同组的病人是否存在某基因片段上的基因拷贝数的不同。

在运用各个检验方法之前，我们随机的从 400 个基因中筛选出 50 个基因，然后我们对这 50 个基因运用本文所举的检验统计量，比较各检验统计量的优劣性，为了计算 FPR，我们用 bootstrap 的方法从长期生存的病人组里抽取两个组，并计算各个检验统计量的值。因为两个组均来自长期生存的病人，因此可以认为零假设为真。为了计算 TPR，我们用 bootstrap 的方法从长期生存的病人组和短期生存的病人组里各取一个组并计算各个检验统计量的值。这时我们认为备择假设为真，对于每一个组，抽样样本数均为 5。

图 5.1 给出了将各检验统计量运用于脑癌数据后的 ROC 曲线，表 5.1 给出各检验统计量的 ROC 曲线对应的 AUC。由各检验统计量的 ROC 曲线以及对应的 AUC，我们可以看出，GSDT 检验显著的优于 SD 和 CQ 检验，与 CPPW 和 HTG 对比，GSDT 检验也

有一定的优势。因此，在样本量较小并且多个探针处发生较小的拷贝数变化的情况下，相较于其他检验统计量，GSDT 检验统计量能够较好地检测出两组病人基因的不同。同时，相较于其他检验方法，GSDT 检验也能更好的避免检测出同组病人基因的不同。

表 5.1 各检验统计量对脑癌数据分析的 AUC 值

检验统计量	GSDT	SD	CQ	CPPW	HTG
AUC	0.564	0.513	0.491	0.548	0.520

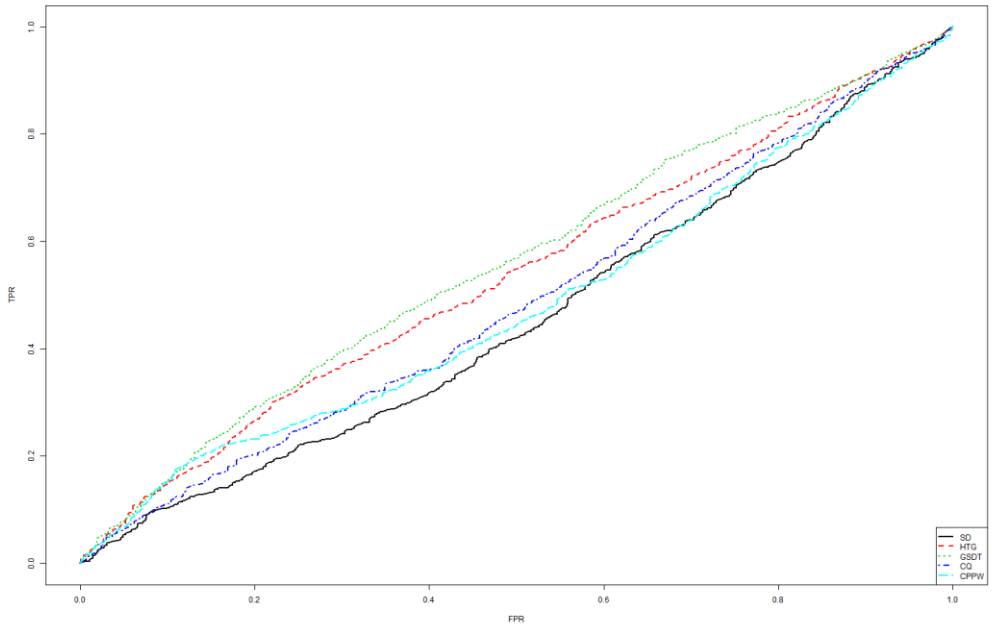

图 5.1 各检验统计量对脑癌数据分析的 ROC 曲线

5.2 骨癌数据中的多元均值检验

为了更好的比较 GSDT 检验和其他检验方法，我们也研究各个检验统计方法在检验骨髓癌数据中的表现。该骨髓癌数据来自 Gene Expression Omnibus（GEO）数据库，其序列号为 GSE2658。多发性骨髓癌以浆细胞肿瘤性增殖，并产生单克隆免疫球蛋白为特征，浆细胞在骨髓内增殖，常常导致广泛溶骨性骨质破坏，骨质减少和病理性骨折。对于多发性骨髓癌，常用的治疗方式包含高剂量治疗并伴随干细胞移植（TT2）和 3 型总治疗（TT3）。科学家们希望通过分析该两类治疗情况下病人基因表达是否不同来为将来的医疗发展提供依据。我们所研究的数据包含 351 位在 2 型总治疗下的骨髓癌病人和

208 位在 3 型总治疗下的骨髓癌病人。通过使用基因微阵列技术，记录病人染色体上的基因表达值。我们所研究的基因片段包含 54675 个探针所记录的基因表达数据。因此，我们使用各多元均值检验方法来检测不同治疗方案的病人是否存在某基因片段上的基因表达的不同。

我们使用上文所提的同样的筛选方法，随机的从 54675 个基因中筛选出 50 个基因，并且用上文所提的同样的 bootstrap 的方法，计算各个检验统计量的 FPR 和 TPR。对于每一个组，抽样样本同样为 5。

表 5.2　各检验统计量对骨髓癌数据分析的 AUC 值

检验统计量	GSDT	SD	CQ	CPPW	HTG
AUC	0.753	0.616	0.627	0.611	0.621

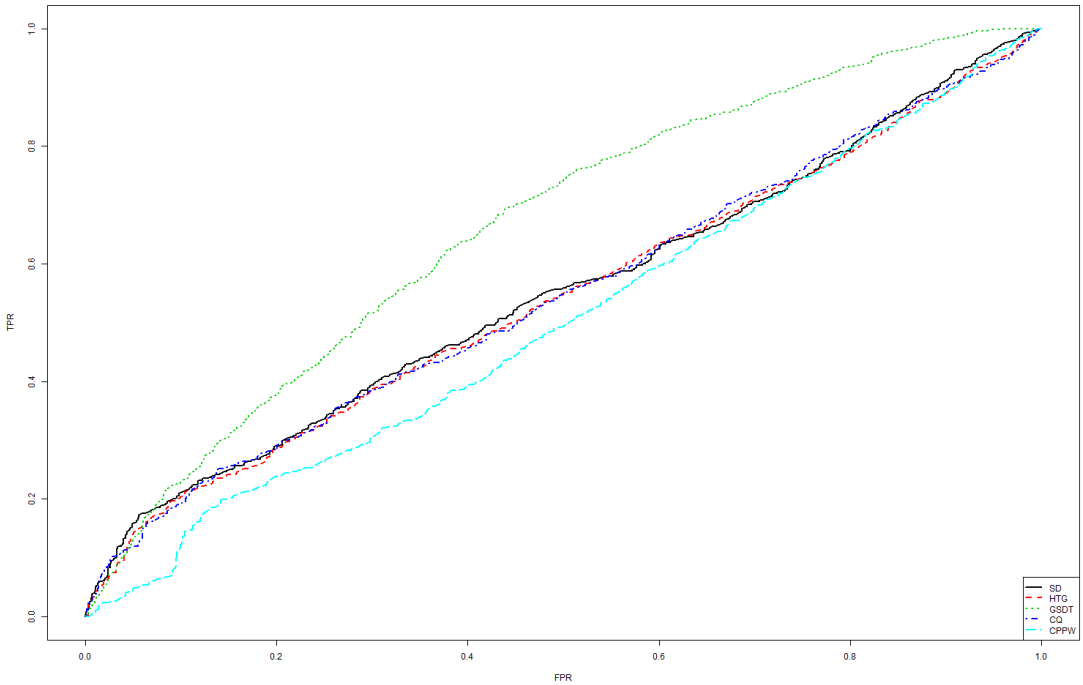

图 5.2　各检验统计量对骨髓癌数据分析的 ROC 曲线

图 5.2 给出了将各检验统计量运用于骨髓癌数据后的 ROC 曲线，表 5.2 给出各检验统计量的 ROC 曲线对应的 AUC。由各检验统计量的 ROC 曲线及对应的 AUC，我们可以看出，GSDT 检验显著的优于其他的检验。因此，相较于其他检验，GSDT 检验能够

较好地检测出两组病人基因表达的不同。同时，相较于其他检验统计量，GSDT 检验也能更好的避免检测出同组病人基因的不同。

第 6 章 总结与展望

本文研究的是关于多元总体均值假设问题，当变量维数小于样本量时，解决该问题的最优方法是 Hotelling T^2检验。高维数据给经典的 Hotelling T^2检验带来了极大的挑战：当变量维数大于样本量时，Hotelling T^2检验失去定义。本文针对于高维数据的特点，在单样本和双样本情形下分布提出了几何收缩对角 Hotelling T^2检验。通过用几何收缩法估计的方差来代替对角 Hotelling T^2检验中的样本方差来构造检验，并推导证明了改进的对角 Hotelling T^2检验的检验统计量在不同情况下的渐近分布。

本文讲述了经典的 Hotelling T^2检验，并介绍了四种对 Hotelling T^2检验改进后的检验，通过模拟分析，几何收缩对角 Hotelling T^2检验相比于其他改进的检验方法能过更好地检测出信号；在样本量较小时，几何收缩对角 Hotelling T^2检验能够较好地控制第一类错误率；在控制第一类错误率的情况下，几何收缩对角 Hotelling T^2检验的效用函数优于其他检验方法。在实证分析中，本文研究了各个检验统计方法在检验长期生存脑癌病人和短期生存脑癌病人之间基因拷贝数的差别以及研究了各个检验统计方法在两种不同治疗方法下，骨髓癌病人的基因表达是否不同。两组实证数据分析都表明，几何收缩对角 Hotelling T^2检验能更好地检验出同组病人基因的不同。几何收缩对角 Hotelling T^2检验在大多数情形下都比其他四种改进方法具有更好的检验效果。

因为本人的水平有限，这篇文章种所提出的检验方法还存在这一些不足。在模拟分析中，本文的模拟数据是从多元正态分布中随机生成的，实际的基因表达数据可能与多元正态分布有出入，这些情况需要在未来做进一步深入的研究，力求我在将来建立更加优越的检验，能够适用于不同类型的基因表达数据。

参考文献

[1] 陈希孺. 高等数理统计学[M]. 中国科学技术大学出版社, 2009.

[2] 潘东东, 李正帮, 张维和李启寨. 全基因组关联研究综述[J]. 应用概率统计, 2014, 30(1): 84-103.

[3] 肖敏. 协方差矩阵的几何型收缩估计及其应用[D]. 浙江工商大学, 2018.

[4] Ahmad MR, Rosen D and Singull M. A note on mean testing for high dimensional multivariate data under non-normality[J]. Statistica Neerlandica 2012, 67(1): 81-99.

[5] Ahmad MR. A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens-Fisher setting[J]. Annals of the Institute of Statistical Mathematics 2014, 66: 33-61.

[6] Bai Z and Saranadasa H. Effect of high dimension: by an example of a two sample problem[J]. Statistica Sinica 1996, 6(2): 311–329.

[7] Baladandayuthapani V, Ji Y, Talluri R, Nieto-Barajas LE and Morris JS. Bayesian random segmentation models to identify shared copy number aberrations for array CGH data[J]. Journal of the American Statistical Association 2010, 105(429): 1358-1375.

[8] Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations[J]. Bernoulli 2004, 10(6): 989–1010.

[9] Chen L, Paul D, Prentice R and Wang P. A regularized Hotelling's T^2 test for pathway analysis in proteomic studies[J]. Journal of the American Statistical Association 2011, 106(496): 1345–1360.

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

[11] Cui X, Hwang JTG, Qiu J, Blades NJ and Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates[J]. Biostatistics 2005, 6(1):59–75.

[12] Cui X and Churchill GA. Statistical tests for differential expression in cDNA microarray experiments[J]. Genome Biology 2003, 4(4): 210.

[13] Dempster AP. A high dimensional two sample significance test[J]. The Annals of Mathematical Statistics 1958, 29(4): 995-1010.

[14] Dempster AP. A significance test for the separation of two highly multivariate small samples[J]. Biometrics 1960, 16(1): 41–50.

[15] Dinu I, Potter JD, Mueller T, Liu Q, Adewale A, Jhangri G, Einecke G, Famulski K, Halloran P and Yasui Y. Improving gene set analysis of microarray data by SAM-GS[J]. BMC Bioinformatics 2007, 8: 242.

[16] Dong K, Pang H, Tong T and Genton MG. Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data[J]. Journal of Multivariate Analysis 2016, 143: 127–142.

[17] Dudoit S, Fridlyand J and Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data[J]. Journal of the American Statistical Association 2002, 97(457): 77–87.

[18] Feng L, Zou C, Wang Z and Zhu L. Multivariate-sign-based high-dimensional tests for the two-sample location problem[J]. Journal of the American Statistical Association 2016, 514(111): 721–735.

[19] Gregory KB, Carroll RJ, Baladandayuthapani V and Lahiri SN. A two sample test for equality of means in high dimension[J]. Journal of the American Statistical Association 2015, 110(510): 837–849.

[20] Hotelling H. The generalization of Student's ratio[J], The Annals of Mathematical Statistics 1931, 2(3): 360–378.

[21] Hu Z, Tong T and Genton MG. Diagonal likelihood ratio test for equality of mean vectors in high-dimensional data[J]. Biometrics 2018, 75: 256-267.

[22] James W and Stein C. Estimation with quadratic loss [M]. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA: University of California Press, 1961, 1: 443–460.

[23] Jiang T and Qi Y. Likelihood ratio tests for high-dimensional normal distributions[J]. Scandinavian Journal of Statistics 2015, 42(4): 988-1009.

[24] Jiang T and Yang F. Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions[J]. The Annals of Statistics 2013, 41(4): 2029-2074.

[25] Katayama S, Kano Y and Srivastava MS. Asymptotic distributions of some test criteria for the mean vector with fewer observations than the dimension[J]. Journal of Multivariate Analysis 2013, 116: 410-421.

[26] Li H, Aue A, Paul D, Peng J and Wang P. An adaptable generalization of Hotelling's T^2 test in high dimension[EB/OL]. arXiv preprint arXiv 1609.08725, 2016. http://arxiv.org/pdf/1609.08725

[27] Lopes ME, Jacob L and Wainwright MJ. A more powerful two-sample test in high dimensions using random projection[J]. Advances in Neural Information Processing Systems 2011 1(2): 1206–1214.

[28] Maboudou-Tchao EM and Silva I. Tests for mean vectors in high dimension[J].Statistical Analysis and Data Mining 2013, 6(6): 578-598.

[29] Marcenko V and Pastur L. Distribution of eigenvalues in certain sets of random matrices[J]. Mat. Sb. (N.S.). 1967, 72: 507-536.

[30] Olshen AB, Venkatraman ES, Lucito R and Wigler M. Circular binary segmentation for the analysis of array‐based DNA copy number data[J]. Biostatistics 2004, 5(4): 557-572.

[31] Park J and Ayyala DN. A test for the mean vector in large dimension and small samples[J]. Statistical Planning and Inference 2013, 143(5): 929–943.

[32] Shen Y, Lin Z and Zhu J. Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis[J]. Computational Statistics and Data Analysis 2011, 55(7): 2221–2233.

[33] Srivastava MS, Katayama S and Kano Y. A two sample test in high dimensional data[J]. Multivariate Analysis 2013, 114(1): 349–358.

[34] Srivastava MS. A test for the mean vector with fewer observations than the dimension under non-normality[J]. Multivariate Analysis 2009,100(3): 518–532.

[35] Srivastava MS and Du M. A test for the mean vector with fewer observations than the dimension[J]. Multivariate Analysis 2008, 99(3): 386–402.

[36] Stadler N and Mukherjee S. Two-sample testing in high dimensions[J]. Journal of the Royal Statistical Society: Series B 2017, 79(1): 225-246.

[37] Thulim M. A high-dimensional two-sample test for the mean using random subapaces[J]. Computational Statistics and Data Analysis 2014, 74: 26-38.

[38] Tong T, Jang H and Wang Y. James-Stein type estimators of variances[J]. Multivariate Analysis 2012, 107: 232–243.

[39] Tong T and Wang Y. Optimal shrinkage estimation of variances with applications to microarray data analysis[J]. Journal of the American Statistical Association 2007, 102(477): 113–122.

[40] Wei S, Lee C, Wichers L, Li G and Marron JS. Direction-Projection-Permutation for high dimensional hypothesis tests[J]. Journal of Computational and Graphical Statistics 2016, 2016(2): 549-569.

[41] Wu Y, Genton MG and Stefanski LA. A multivariate two-sample mean test for small sample size and missing data[J], Biometrics 2006, 62(3): 877–885.

[42] Zhang J and Xu J. On the *k*-sample Behrens-Fisher problem for high-dimensional data[J]. Science in China Series A: Mathematics 2009, 52(6): 1285-1304.

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

致谢

在毕业论文即将完成之际，心中虽然万千感慨，却依然难以下笔，或许是对自己二十几年的学业生涯的不舍，或许是经不起朋友离别时的那种悲伤，完成了致谢，也为自己的毕业论文画上了一个句号。在浙工商的两年的的时光中，从求知、求学到积累、历练，它是我人生的一个及其重要的阶段，在此期间，我的心灵得到了锤炼，学业也得到了升华，使我可以充满自信的投入人生新阶段。回顾这两年半，我所获得的各种财富都离不开我的导师江涛教授、各位老师、同学以及家人们的关心与帮助。在论文完成之际，我要在此对他们表示最诚挚的感谢。

首先，我要特别感谢我的导师江涛教授，一个才华横溢的学者，一个宽厚仁慈的长辈，我从你身上学到的东西虽然只有九牛一毛，但足以我受益终生。统计学与数学的学习方法有着很大的不同，起步阶段我用了很长的时间。江老师给予我足够的时间让我扎扎实实学好统计的基础知识，让我感受到统计的魅力。江老师总是教导我们，打好基础是做学术的第一步。学习统计的两年时间里，我磨掉了急性子，从江老师身上学会了静心和耐心的思考问题。在研究生期间，我很感谢江老师带着我们参加各种统计会议，使我有机会能接触最新的最前沿的统计学知识。在此谨向导师江涛教授表示崇高的敬意和衷心的感谢！

其次，我要感谢明瑞星老师，一个对学术具有不懈专研的精神的学者。我要特别感谢明老师给我选的第一个研究课题—因子模型。尽管我在这个领域没有做出任何好的成果，但是那些统计学大家深邃的思想带给我无穷的思考。感谢明老师在自己的工作十分的繁忙的情况下，仍然坚持给我们上讨论班，这种尽职尽责的工作态度让我感念至深。

另外我还要感谢陈振龙教授，王炳兴教授，王伟刚教授以及在研究生期间给我教授过课程的每一位老师。是你们让我在统计知识的海洋里汲取更多的营养，使得我的统计知识更加稳固扎实。每一位老师渊博的专业知识，严谨的治学态度，诲人不倦的高德师德对我影响深远。

我还要感谢我的师姐肖敏老师，我十分荣幸有这样一个师姐，在我研究的道路上给予莫大的帮助。每次有困难，肖敏师姐都耐心的指导我，帮助我开拓研究思路，精心点拨，并为我指导迷津，才使得我在面对各种问题的时候得以豁然开朗。同时感谢段会敏，陶杨明，陈雨馨，陈婷师姐，鲁亚会师姐，周港归和孟雪等同门姐妹和王文慧和张巧艳两室友，她们在学习和生活上都给与了我无私的帮助和关心。

衷心感谢我的父母在我多年求学生涯中给予我精神上、生活上无私的付出和帮助。你们含辛茹苦地培养了我二十四年，你们从不埋怨，因为有你们的支持，我才能顺利地走到今天，请相信我，我将用实际行动回报你们！

最后，感谢答辩组地老师们，感谢你们提出地宝贵意见，使我的论文质量有了很大的提高。

谨以此文献给所有关心、帮助和支持过我的人们！

独 创 性 声 明

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

学位论文作者签名：江梓婕　　导师签名：

签字日期： 2020 年 6 月 1 日　　签字日期：2020 年 6 月 1 日

关于论文使用授权的说明

本学位论文作者完全了解浙江工商大学有关保留、使用学位论文的规定：浙江工商大学有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅，可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文，并且本人电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

学位论文作者签名：江梓婕　　导师签名：

签字日期：2020 年 6 月 1 日　　签字日期：2020 年 6 月 1 日