单位代码	10475
学　　号	104753180640
分 类 号	O21

河南大學
硕士学位论文

高维数据总体双可交换协方差矩阵的似然比检验
The Likelihood Ratio Tests of Doubly Exchangeable Covariance Structure for High-Dimensional Data

学科、专业：统计学
研究方向：数理统计
申请学位类别：理学硕士
申　请　人：袁百城
指导教师：解俊山 副教授

二〇二一　年　六　月

The Likelihood Ratio Tests of Doubly Exchangeable Covariance Structure for High-Dimensional Data

A Dissertation Submitted to
the Graduate School of Henan University
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science

By
Yuan Baicheng
Supervisor: Associate Professor Xie Junshan

Date: June, 2021

关于学位论文独创声明和学术诚信承诺

本人向河南大学提出硕士学位申请。本人郑重声明：所呈交的学位论文是本人在导师的指导下独立完成的，对所研究的课题有新的见解。据我所知，除文中特别加以说明、标注和致谢的地方外，论文中不包括其他人已经发表或撰写过的研究成果，也不包括其他人为获得任何教育、科研机构的学位或证书而使用过的材料。与我一同工作的同事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了谢意。

在此本人郑重承诺：所呈交的学位论文不存在舞弊作伪行为，文责自负。

学位申请人（学位论文作者）签名：袁百城

2021 年 6 月 15 日

关于学位论文著作权使用授权书

本人经河南大学审核批准授予硕士学位。作为学位论文的作者，本人完全了解并同意河南大学有关保留、使用学位论文的要求，即河南大学有权向国家图书馆、科研信息机构、数据收集机构和本校图书馆等提供学位论文（纸质文本和电子文本）以供公众检索、查阅。本人授权河南大学出于宣扬、展览学校学术发展和进行学术交流等目的，可以采取影印、缩印、扫描和拷贝等复制手段保存、汇编学位论文（纸质文本和电子文本）。

（涉及保密内容的学位论文在解密后适用本授权书）

学位获得者（学位论文作者）签名：袁百城

2021 年 6 月 15 日

学位论文指导教师签名：

2021 年 6 月 15 日

摘　要

随着大数据时代的到来, 高维数据广泛地应用于金融、生物、物理等领域. 但是在假设检验问题中考虑高维似然比统计量时, 很多传统的统计方法往往基于样本容量 N 很大, 而样本维数 p 相对固定的假定, 这时可以通过经典卡方近似 (CHI 方法) 等方法进行有效地处理. 但是这些方法在处理样本容量 N 和样本维数 p 都很大时, 往往效果很差甚至失效. 因此我们需要探索新的方法来处理具有大 N、大 p 特征的高维数据.

本文主要研究了具有双可交换协方差矩阵结构的高维数据的假设检验问题. 首先, 在双可交换协方差矩阵结构的原假设下, 寻求似然比统计量的矩的表达形式. 然后以此作为本文的出发点, 探寻假设检验当中似然比统计量的渐近分布问题. 在本文中, 我们主要采用了以下三种方法. 第一种是高阶伽马函数展开的方法 (HGM 方法), 通过多元伽马函数将高阶矩展开来分析似然比统计量的渐近性质. 第二种是通过将似然比统计量转化为独立但非同分布的贝塔分布随机变量的乘积形式 (HBM 方法), 借用 Lindeberg 条件的方法, 以此为出发点研究渐近性质. 最后一种是高阶 Edgeworth 展开的方法 (HEM 方法), 这种方法在寻求渐近分布的同时对渐近分布的误差界进行了研究. 前两种方法具有简洁与直观的特点, 且这两种方法的期望与方差在极限状态是等价的; 后一种方法可以通过控制参数 l 来控制渐近分布的精度.

最后, 将本文种所提出的三种方法与传统的卡方近似方法进行比较. 数值模拟表明, 当样本容量 N 和样本维数 p 都很大时, 传统的卡方近似方法 (CHI 方法) 已失效, 本文中提出的似然比统计量的三种近似方法 (HGM 方法、HBM 方法和 HEM 方法) 效果较好. 同时通过实例分析, 表明双可交换协方差矩阵结构具有很好的现实意义.

关键词: 高维数据, 双可交换协方差矩阵结构, 似然比统计量, 渐近分布.

ABSTRACT

In the era of big data, high-dimensional data is widely used in finance, biology, physics and other fields. However, when considering the high-dimensional likelihood ratio statistics in the hypothesis testing, many traditional statistical methods are usually based on the assumption that the sample size N is large and the sample dimension p is relatively fixed. At this time, it can be effectively processed by classical chi-square approximate and other methods. But, when the sample size N and the sample dimension p are both large, these methods are very poor or even invalid on dealing with high-dimensional data. Therefore, we need to explore new methods to deal with high-dimensional data with large N and large p.

This paper mainly considers the hypothesis test problem with high-dimensional doubly exchangeable covariance matrix structure. Firstly, under the assumption of the doubly exchangeable covariance matrix structure on the Gaussian distribution, seeking the expression of moments under the null hypothesis. Then take this as the starting point of this paper to explore the asymptotic distribution. In this paper, we mainly adopt three different methods. The first method is the expansion of the gamma function (the HGM method), which uses the multivariate gamma function to explore the asymptotic properties of the likelihood ratio statistics. The second method is to transform the likelihood ratio statistic into the product of independent but non-identically distributed beta distribution random variables (the HBM method), and using the method of Lindeberg condition to study the asymptotic properties. The last method is the high Edgeworth expansion method (the HEM method), which studies the error bounds of asymptotic distribution while seeking the asymptotic distribution. The fist two methods are simple and intuitive, and their expectation and variance are equivalent in the limit state; the last one can control the precision of the approximate distribution by controlling the parameters l.

Finally, we compare the three methods proposed in this paper to the traditional chi-square approximation method. Numerical simulation shows that when the sample size N and the sample dimension p are both large, the traditional chi-square approximation method (the CHI method) is invalid. The three approximation methods (the HGM method, the HBM method and the HEM method) for likelihood ratio statistics

presented in this paper perform better. At the same time, through the example analysis, they show that the doubly exchangeable covariance matrix structure has a good practical significance.

KEY WORDS: High-dimensional data, Doubly exchangeable covariance matrix structure, Likelihood ratio statistic, Approximate distribution.

文中符号说明

tr(A)	矩阵 A 的迹
exp(x)	e^x
$A > 0$	矩阵 A 是正定阵
I_n	$n \times n$ 的单位矩阵
1_n	所有元素为 1 的 $n \times 1$ 向量
$X_n \xrightarrow{d} X$	随机变量序列 $\{X_n\}$ 依分布收敛于随机变量 X
$a_n = O(b_n) \quad (n \to \infty)$	$\lim_{n\to\infty} \sup \lvert \frac{a_n}{b_n} \rvert < \infty$
$a_n = o(b_n) \quad (n \to \infty)$	$\lim_{n\to\infty} \frac{a_n}{b_n} = 0$
$a_n \sim b_n \quad (n \to \infty)$	$\lim_{n\to\infty} \frac{a_n}{b_n} = 1$

目 录

第一章　绪论

§1.1　研究背景及意义

统计学是一门涉及收集、组织、分析和解释数据的学科. 目前广泛地应用于物理学, 生物学, 社会学, 金融学等各个学科, 同时也应用于企业, 政府等部门的决策当中.

在过去, 由于时代和生产力的限制, 人们能够观测和记录下数据的数据容量 N 及数据维数 p 不会很大, 因此人们在分析和解释数据时往往也局限于低维的情况, 统计模型和分析方法也仅仅针对低维的情况, 但是这些方法几乎很少考虑样本维数很大时的有效性与可行性. 如 Box (1949) [6] 基于样本维数 p 相对固定, 而样本容量 N 趋于无穷的低维情况时, 研究了基于原假设下似然比统计量的卡方近似.

如今, 随着科学技术的发展, 搜集数据、存储数据的成本降低, 统计模型和分析方法考虑的数据容量和数据维数越来越大, 此时我们称这些具有大 N、大 p 特征的数据为高维数据. 高维数据具有丰富的商业价值和科研价值. 对于电商企业, 可以通过对客户的浏览记录和购买记录的分析, 提供个性化服务, 增加顾客粘度. 对于科研机构, 在医学领域基因表示数据中, 由于基因的表示往往涉及大量的指标, 所以通过分析与解释这些基因数据, 科研人员可以提高成果转化率, 为科研带来极大的便利. 对于政府, 可以通过分析国内每年的经济数据指标, 为下一年的经济政策提供数据支撑.

随着生产力的快速发展, 大数据既会带来丰富的信息量, 但也会增加处理和分析数据的难度. 由于经典的假设检验问题主要是基于低维的情况分析统计数据和统计量的, 但基于大 N、小 p 的假定在当今海量数据面前显然是无法被满足的, 也就使得传统的方法无法有效地应用于当今的数据集中, 如在 Bai & Saranadasa (1996) [4] 中专门讨论了传统方法在解决高维数据时的有效性, 其结果表明传统理论方法在高维领域将不再具有有效性.

为此, 寻找新的有效的处理高维数据的方法已成为各个领域的迫切需求. 所以基于高维数据的假设检验问题对现代学科领域具有显著的意义, 同时能够对现实生活中的实践活动有促进作用.

§1.2 研究现状

在本文中, 我们重点研究假设检验问题中似然比统计量的相关问题. 在统计学中似然比统计量 (Likelihood Ratio Statistic) 是一种应用广泛的统计量之一, 其最早是由 Neyman & Pearson (1928) [32] [33] 提出的, 后经过不断地完善, 逐步发展为比较原假设和备择假设拟合优度中应用广泛的检验理论. 在经典的统计学问题当中, 我们一般考虑样本容量 N 很大, 而样本维数 p 固定不变或很小的情况, 如 Box (1949) [6] 提出了在样本维数 p 固定情形下的经典卡方近似方法, Bartlett (1954) [5] 总结了多种针对卡方近似结果的检验, Muirhead (1982) [31], Eaton (1983) [16], Anderson (2003) [2] 中也都总结了基于不同假设检验问题似然比统计量的卡方近似和矩的表达式, 但随着科学技术的发展, 很多方法不再适用, 因此需要开发新的统计方法和构造新的统计模型对高维数据进行区别于之前的处理.

在高维数据下, 基于高维数据的统计方法和统计模型, 假设检验中似然比统计量焕发了新的生机. 在样本容量 N 样本维数 p 都很大时, 众多研究者对其做出各项研究, 如: Schott (2001) [40], Schott (2005) [41], Schott (2007) [42] 检验了多个协方差矩阵的相等性; Ledoit & Wolf (2002) [25], Srivastava (2005) [45], Srivastava & Du (2008) [46], Chen et al. (2010) [10] 等人基于正态分布研究了样本维数略微大于样本容量的数据; Wakaki (2006) [51], Wakaki (2007) [52], Fujikoshi et al. (2010) [19] 等人通过 Edgeworth 方法研究了似然比统计量的渐近性质; Bai et al. (2009) [3], Jiang et al. (2012) [21], Jiang & Yang (2013) [22], Jiang & Qi (2015) [23] 等人通过伽马函数的展开研究了多种协方差矩阵的似然比统计量的性质. 这些研究结果都在不同的方面对假设检验问题下高维似然比统计量的近似估计进行解释说明, 构建了高维统计中渐近分布的比较完善的理论方法体系, 为分析高维数据下似然比统计量的渐近分布问题起到了启发作用.

在现实生活中, 我们往往会考虑在多个地点重复地收集多个反应变量的测量值所构成的数据集, 这一类的数据集往往被称为三级多元数据. 本文主要研究的就是一个包含三级多元向量的双可交换协方差矩阵结构, 这一问题最早是 Shults et al. (2004) [44] 在以非裔美国妇女体育活动参与习惯为动机构建广义估计方程时提出的, 同时也可以视作 Shults & Morrow (2002) [43] 中提出的二级多元数据协方差矩阵的推广.

在一定条件下, 双可交换协方差矩阵结构可以转化为更一般的复合对称结构和球型结构的形式. 因此, 许多有趣的协方差矩阵结构都可以视为它的特殊情况之一, 如: 双复合对称结构、块复合对称结构、对角块复合对称结构、复合对称结构、对角块球型结构、

球型结构等. 关于块复合对称结构, Votaw (1948) [50], Szatrowski (1976) [47], Szatrowski (1982) [48], Coelho & Roy (2017) [12] 等人解决了此结构的渐近性质和基于此结构来分析统计模型; 关于对角块球型结构, 在 Moschopoulos (1992) [30], Cardeño & Nagar (2001) [9], Marques & Coelho (2012) [29] 等文献中, 研究者们提出了解决渐近分布等问题的想法; 关于球型结构, 由于球型结构是一个比较常见的统计结构, 所以许多研究者都基于球型结构讨论了似然比统计量的统计性质, 在 Gleser (1966) [20], Wang & Yao (2013) [53], Coelho & Marques (2013) [11], Jiang & Yang (2013) [22], Jiang & Qi (2015) [23] 中针对似然比统计量的球形检验提出了渐近估计并与 John's 检验等做了对比. 这些研究工作汇总在一起, 涵盖了低维和高维情形下的多种情况, 而且基于不同的统计模型, 所以具有十分重要的应用价值.

本文主要研究了双可交换协方差矩阵结构模型的假设检验问题. Roy & Leiva (2008) [35] 研究了三级多元数据的协方差结构的可分离性. Roy & Fonseca (2012) [36] 通过三级多元数据拟合一个广义线性模型, 其中误差向量具有双可交换协方差矩阵结构. Leiva & Roy (2011) [26], Leiva & Roy (2012) [27] 使用双可交换协方差矩阵结构对三级多元数据进行分类研究. 但这些研究者都没有直接解决高维情况下双可交换协方差矩阵结构的近似分布问题, 直到近两年, Coelho & Roy (2020) [13] 才进一步研究了似然比统计量的近似精确分布, 但也没有考虑到三级变量 u, v, m 皆为高维的情形.

然而, 这些研究要么没有直接解决高维情况下双可交换协方差矩阵结构的渐近分布问题, 要么未以高维情形为出发点研究统计量的渐近分布. 因此, 针对具有双可交换协方差矩阵结构的高维统计模型的假设检验问题, 如何解决似然比检验统计量的渐近分布, 在假设检验中是一个具有现实意义的问题, 这也正是本文关注的焦点.

基于此, 本文考虑采用似然比检验方法, 并以似然比统计量的高阶矩为出发点, 分别通过高阶伽马函数展开方法, 高阶贝塔分布随机变量乘积展开方法和高阶 Edgeworth 渐近展开方法去解决高维双可交换协方差矩阵结构的假设检验问题.

§1.3 相关理论

在统计学当中, 似然比检验通过对两个统计模型的似然比来评估两个模型的拟合优度. 如果观测数据支持该约束 (即原假设), 则这两种可能性的差异不应超过采样误差, 如 Neyman & Pearson (1928) [32] [33] 中介绍的. 在解决具体问题时, 主要有显著性水平检验

和 p 值检验两种方法. 下面, 我们将介绍似然比检验的基本思路.

对于一般的情况, 我们假设一个带有参数空间 Ω 的高维统计模型, 其随机变量 $X = (X_1, X_2, \cdots, X_n)$, 且 X_i 具有密度函数 $f(x_i;\theta)$, $i = 1, \cdots, n$. 在零假设下, 往往假定参数 θ_0 在一个位于 Ω 特殊的子集 Ω_0 当中, 备择假设的参数 θ 位于子集 $\Omega_0^c = \Omega/\Omega_0$ 中. 在零假设 $H_0 : \theta_0 \in \Omega_0$ 下的似然比统计量:

$$\Lambda = \frac{\sup_{\theta_0 \in \Omega_0} \pounds(\theta_0)}{\sup_{\theta \in \Omega} \pounds(\theta)},$$

其中

$$\pounds(\theta) = \prod_{k=1}^{N} f(x_k;\theta).$$

由此对于高维样本 X, 我们由极大似然比统计量的原理知, 若 Λ 取值很小, 则在原假设 H_0 为真时观测到此样本 X 的概率远远小于在 H_0 为不真时观测到此样本 X 的概率. 故而, 在 Λ 取值很小时, 有理由认为原假设 H_0 不成立. 那么, 从似然比检验的角度考虑, 对于给定的显著性水平 α, 这将导致一种直觉决策规则,

$$\text{当 } \Lambda \le c, \text{拒绝 } H_0, \text{ 接受 } H_1.$$

其中常数 c 使得 $\alpha = P_{\theta_0}[\Lambda \le c]$. 从这里可以看出, 研究似然比检验统计量 Λ 的抽样分布是似然比检验的一个关键点, 这也正是本文关注的焦点.

接下来, 若记统计量 Λ 的抽样分布为 $g(\Lambda)$, 则可算得一个概率

$$p = g(\Lambda) = P\{\Lambda < \Lambda_\alpha\}.$$

这即是检验的 p 值. 基于此, 考虑上述假设检验问题, 可作如下判断:

(1) 当 $\alpha < p$ 时, 有 $\Lambda_\alpha < \Lambda_0$, 即 Λ_0 不属于拒绝域 $\omega = \{\Lambda < \Lambda_\alpha\}$, 应接受原假设;

(2) 当 $\alpha \ge p$ 时, 有 $\Lambda_\alpha \ge \Lambda_0$, 即 Λ_0 属于拒绝域 $\omega = \{\Lambda < \Lambda_\alpha\}$, 应拒绝原假设.

§1.4 主要创新点

本文的主要创新点是借助似然比统计量的矩的函数, 利用高阶伽马函数展开方法, 独立贝塔分布的随机变量乘积展开方法和高阶 Edgeworth 展开方法来求解假设检验下高维

样本似然比统计量的渐近分布问题. 这三种方法都是对高阶似然比统计量的矩函数的展开, 利用矩母函数的一致有界性和特征函数的逆转公式得到统计量的渐近分布. 通过数值模拟, 与传统的卡方近似方法进行了比较, 验证了在高维样本时这三种方法的有效性. 主要体现在以下三个方面.

一、高阶伽马函数展开方法在展开似然比统计量的矩的函数时, 通过采用 Jiang & Yang (2013) [22], Jiang & Qi (2015) [23] 中对多元伽马函数的展开式, 再由泰勒展开的方法合并同类项, 构造合适的矩母函数, 然后通过矩母函数的一致有界性得出该统计量服从渐近正态分布.

二、由似然比统计量转变为独立贝塔分布的随机变量乘积可以较为容易地求解统计量的期望与方差, 但是在证明似然比统计量在极限情况具有渐近正态分布时, 需要通过对 Dette & Dõrnemann (2020) [14] 中提出的衡量统计量渐近分布的五个约束条件一一求解, 最后通过经验 Size 和经验 Power 这两个指标去验证所提出的方法是合理有效的.

三、通过 Edgeworth 展开方法对似然比统计量的特征函数进行逆转运算, 可以得出统计量的渐近分布和与真实分布的渐近误差界. 同时在对渐近分布的误差界进行研究时, 经过傅里叶逆变换后需要选择合适的方法对积分进行约束, 因此可以通过控制变量来控制误差界的精确度.

§1.5 文章结构

本文的主要研究内容和结构如下:

第一章是绪论部分. 主要介绍了本文所研究问题的背景, 意义及国内外研究现状.

第二章是本文核心部分. 主要陈述了本文的模型结构和主要定理. 首先基于三级多元数据介绍了双可交换协方差矩阵结构, 然后给出了基于此高维协方差矩阵结构的似然比统计量的假设检验模型.

第三章是数值分析部分. 主要借助 Matlab 软件对第二章所提出的定理进行了随机模拟和实例分析. 首先通过随机模拟表明了在给定显著性水平下, 同传统的卡方近似进行对比, 本文提出的理论方法在高维数据检验理论中的有效性. 然后通过实例分析了第二章提出的三种高维近似的效果, 增强了本文理论的真实性.

第四章是详细证明部分. 主要对第二章所提出的定理做出了技术性证明.

最后一章是总结与展望部分. 对本文的内容进行了回顾, 也提出了今后需要继续努力的研究方向.

第二章　高维双可交换协方差矩阵结构的假设检验

§2.1　模型介绍

在过去的几十年里, 计算和分析能力的进步极大地鼓励了所有科学领域: 生物、医学、环境、社会科学和工程等的多级多元数据的收集. 而且, 对于这类数据集, 复杂的多元测试问题经常发生. 在临床试验研究中, 为了测试某些药物、饮食或行为习惯的有效性, 在多个地点重复地收集多个反应变量的测量值是很常见的. 这些被收集来的数据集往往被称为三级多元数据. 而双可交换协方差结构就是适用于这类数据的方差–协方差矩阵.

一个双可交换协方差矩阵是指基于 u 个位置, v 个时间/空间点的 m 个变量的三级多元观测值所构成的矩阵结构. 在 Roy & Leiva (2008) [35], Roy & Leiva (2011) [37], Roy & Fonseca (2012) [36], Roy et al. (2015) [38], Coelho & Roy (2020) [13] 等文献中基于双可交换协方差矩阵结构解决了分类问题、广义估计问题、渐近精确分布问题等. 假定 X 为 uvm 维实值随机向量, $\mathbf{\Theta} = \mathrm{Cov}[X]$ 是一个 $uvm \times uvm$ 的协方差矩阵. 假若 $\mathbf{\Theta}$ 满足下列结构, 我们称矩阵 $\mathbf{\Theta}$ 为双可交换协方差矩阵结构:

$$
\mathbf{\Theta} = \left[\begin{array}{cccc|cccc|c|cccc}
U_0 & U_1 & \cdots & U_1 & W & W & \cdots & W & \cdots & W & W & W & W \\
U_1 & U_0 & \cdots & U_1 & W & W & \cdots & W & \cdots & W & W & W & W \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
U_1 & U_1 & \cdots & U_0 & W & W & \cdots & W & \cdots & W & W & W & W \\
\hline
W & W & \cdots & W & U_0 & U_1 & \cdots & U_1 & \cdots & W & W & W & W \\
W & W & \cdots & W & U_1 & U_0 & \cdots & U_1 & \cdots & W & W & W & W \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
W & W & \cdots & W & U_1 & U_1 & \cdots & U_0 & \cdots & W & W & W & W \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
W & W & \cdots & W & W & W & W & W & \cdots & U_0 & U_1 & \cdots & U_1 \\
W & W & \cdots & W & W & W & W & W & \cdots & U_1 & U_0 & \cdots & U_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
W & W & \cdots & W & W & W & W & W & \cdots & U_1 & U_1 & \cdots & U_0
\end{array}\right],
$$

同理 $\boldsymbol{\Theta}$ 也可以写为

$$\begin{aligned}\boldsymbol{\Theta} &= I_{uv}\otimes U_0+[I_U\otimes(J_u-I_u)]\otimes U_1+[J_{uv}-(I_v\otimes J_U)]\otimes W\\ &= I_{uv}\otimes(U_0-U_1)+I_v\otimes J_u\otimes(U_1-W)+J_{uv}\otimes W\\ &= I_v\otimes U+(J_v-I_v)\otimes W^*,\end{aligned}\tag{2-1}$$

其中

$$\begin{aligned}U &= I_u\otimes U_0+(J_u-I_u)\otimes U_1,\\ W^* &= J_u\otimes W,\\ J_u &= 1_u1_u^{'}.\end{aligned}$$

且 U_0 是 $m\times m$ 的正定对称矩阵, U_1 和 W 是 $m\times m$ 的对称矩阵, 但 U_0, U_1, W 的具体形式不定. 也就是说, 双可交换协方差矩阵结构可以由对角块矩阵 U 和非对角块矩阵 W 组成, 而对角块矩阵 U 是由矩阵 U_0 和 U_1 组成.

因此, 三级多元向量 $X=(x_{11},\cdots,x_{1u},\cdots,x_{v1},\cdots,x_{vu})$ 若是满足

$$\mathrm{Cov}[x_{ts};x_{t^*s^*}]=\begin{cases}U_0 & \text{如果 } t=t^*,s=s^*,\\ U_1 & \text{如果 } t=t^*,s\neq s^*,\\ W & \text{如果 } t\neq t^*,\end{cases}$$

其中 $t=1,\cdots,v;\ s=1,\cdots,u$, 则称三级多元数据是双可交换的.

在 (2-1) 式中 $m\times m$ 的对角块矩阵 U_0 表示在任意给定位置和任意给定时间点的 m 个响应变量的方差–协方差矩阵. 在 (2-1) 式中 $m\times m$ 的非对角块矩阵 U_1 表示在任意两个不同给定位置和任意给定时间点的 m 个响应变量的协方差矩阵. 我们假设矩阵 U_0 是对于所有地点和时间点都是一样的. 矩阵 U_1 对于所有时间点都是一样的. $m\times m$ 的非对角块矩阵 W 表示在任意两个不同给定时间点之间的 m 个响应变量的协方差矩阵, 此时假设对于任何一对不同的时间点, 无论位置如何, 矩阵 W 都是相同的.

对于高维双可交换协方差矩阵结构模型应用的实际情形, 下面给出三个实例:

例 2.1.1 像素位置数据. 像素通常被认为是数字图像中最小的单个组成部分, 而每个像素的强度是可变的. 在颜色系统中, 每个像素通常有 3 或 4 个组成部分, 如红 *(Red)*、绿 *(Green)*、蓝 *(Blue)*, 其中每个像素是 *RGB* 空间中的一个点. 因此, *RGB* 空间是一个三级向量空间 $(u=3)$, 每个像素由红、绿、蓝坐标的有序三重坐标定义. 由于红色、绿色和蓝色坐标的有序三级组的强度之间存在相关性, 所以相邻像素 m 之间也存在相关性. 为

了获取模型任意过程或任意随时间的变化, 数据应当是随着时间 v 获取的. 在这个例子中, 不同像素和不同时间点上的变量不是独立的, 而是随机相关的. 也就是说, 不同的像素在不同的时间点之间可以互换或交换. 因此, 假设变量具有双可交换结构是合理的. 基于这些邻近像素随时间变化的样本的模型必须考虑到数据集的双可交换结构.

例 2.1.2 骨质疏松数据. 在 *Roy & Leiva (2011)*[37] 中介绍了一个关于骨密度的三级多元数据集. 研究人员选择通过使用 *GE-Lunar Prodigy* 型双 *X* 射线骨密度仪 *(DXA)* 获得患者的骨密度值 *(BMD)*. 这些测量数据来自于髋部, 测量得到的股骨径和股骨转子区 $(m = 2)$, 左和右股骨 $(u = 2)$. 由于股骨颈主要是皮质骨, 而股骨转子区主要是松质骨, 所以这两种测量结果可以看作是两个不同的随机变量. 这四个测量位置大约两年后重新测量一次 $(v = 2)$. 因此整个数据集有一个三层多元结构, 可以检验其协方差矩阵是否满足双可交换的结构.

例 2.1.3 矿物含量数据. *Johnson & Wichern (2007)*[24] 在研究骨骼矿物含量的相关问题时收集了一个三级多元数据集. 在他们的一项研究数据当中, 研究人员测量了三个位置骨骼的矿物质含量, 即桡骨, 肱骨和尺骨 $(m = 3)$. 通过光子吸收法检查特定饮食补充剂是否会增加老年妇女的骨骼矿物质含量和质量, 所有三个测量值均记录了每个女性的优势和非优势方面的身体数据 $(u = 2)$. 在首次参加实验计划后的一年, 这些两级多元测量值再次被测量. 因此整个数据集有一个三层多元结构, 考虑到所涉及的变量所代表的意义, 此时检验双可交换协方差矩阵结构可能是一个适当的目标, 可以用以确定饮食补充剂的效果.

接下来在第三章第二节实例分析中将检验例 2.1.2 和例 2.1.3 中的数据集是否满足双可交换协方差矩阵结构的假定.

§2.2 相关预备知识

设 $X_1, X_2, \cdots, X_N$ 是取自多元正态总体 $\boldsymbol{N}_{uvm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 中样本容量为 N 的随机样本, 我们可以将向量 X_i 划分为

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iv} \end{pmatrix}, \quad \text{其中 } X_{it} = \begin{pmatrix} X_{it1} \\ \vdots \\ X_{itu} \end{pmatrix}, \quad \text{其中 } X_{its} = \begin{pmatrix} X_{its1} \\ \vdots \\ X_{itsm} \end{pmatrix},$$

其中 $i=1,\ldots,N,\ t=1,\ldots,v,\ s=1,\ldots,m.$ m 维的向量 X_{its} 表示的是在第 s 位置和第 t 时间点/空间点的观测对象. 由此, 可写出样本均值和样本离差阵分别为

$$\bar{\boldsymbol{X}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i,\quad \boldsymbol{S}=\sum_{i=1}^{N}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})'.$$

接下来, 本文主要考虑多元正态总体的协方差阵 $\boldsymbol{\Sigma}$ 的假设检验问题, 即考虑

$$\boldsymbol{H}_0:\boldsymbol{\Sigma}=\boldsymbol{\Theta}\quad \textbf{vs}\quad \boldsymbol{H}_1:\boldsymbol{\Sigma}\neq\boldsymbol{\Theta}, \tag{2-2}$$

其中 $\boldsymbol{\Theta}$ 由 (2-1) 式定义.

由 Roy & Fonseca (2012)[36] 中 Lemma 3.1 可知, 我们可令 $\Gamma^*=\underset{u\times u}{C}{}'\otimes I_{mu}$ 和 $\Gamma^\bullet=I_v\otimes(\underset{u\times u}{C}{}^{*\prime}\otimes I_m)$. 其中 C 和 C^* 为第一列与 1 成比例的正交 Helmert 矩阵, 且有

$$\Gamma^\bullet\Gamma^*\boldsymbol{\Theta}\Gamma^{*\prime}\Gamma^{\bullet\prime}=\mathrm{diag}(\Delta_3,I_{u-1}\otimes\Delta_1,\Delta_2,I_{u-1}\otimes\Delta_1,\Delta_2,\cdots,I_{u-1}\otimes\Delta_1),$$

其中

$$\Delta_1=U_0-U_1,$$

$$\Delta_2=U_0+(u-1)U_1-uW=(U_0-U_1)+u(U_1-W),$$

$$\Delta_3=U_0+(u-1)U_1+u(u-1)W=(U_0-U_1)+u(U_1-W)+uvW.$$

因为 $\Gamma^\bullet$ 和 Γ^* 不是 U_0 , U_1 , W 的函数, 所以 (2-2) 式中检验 H_0 就等价于检验

$$H_0:\Sigma^*=\boldsymbol{\Delta}. \tag{2-3}$$

其中

$$\Sigma^*=\Gamma^\bullet\Gamma^*\boldsymbol{\Sigma}\Gamma^{*\prime}\Gamma^{\bullet\prime},\qquad \boldsymbol{\Delta}=\Gamma^\bullet\Gamma^*\boldsymbol{\Theta}\Gamma^{*\prime}\Gamma^{\bullet\prime}.$$

同时根据 (2-2) 式零假设 H_0 可以分解为

$$H_0=(H_{0c|a}||H_{0b|a})\circ H_{0a},$$

其中 $\circ$ 表示 ‘之后’ , $||$ 表示 ‘之前或之后’ , 即有:

$H_{0a}:\Sigma^*=\mathrm{diag}(\Sigma_i^*),\ i=1,\cdots,uv;$

假定在 H_{0a} 下有 :

$$H_{0b|a}:\underbrace{\Sigma_2^*=\cdots=\Sigma_u^*}_{u-1}=\underbrace{\Sigma_{u+2}^*=\cdots=\Sigma_{2u}^*}_{u-1}=\cdots=\underbrace{\Sigma_{(v-1)u+2}^*=\cdots=\Sigma_{vu}^*}_{u-1},$$

$$H_{0c|a}:\Sigma_{u+1}^*=\Sigma_{2u+1}^*=\cdots=\Sigma_{(v-1)u+1}^*.$$

由此, 当原假设 H_0 成立时, 在 Coelho & Roy (2020) [13] 的工作中提到了在参数空间下的极大似然估计, 并由此可得似然比检验统计量为

$$\begin{aligned}\boldsymbol{\Lambda} &= \Lambda_a\Lambda_b\Lambda_c \\ &= \left((vu-v)^{mv(u-1)}(v-1)^{m(v-1)}\frac{|A|}{|A_1||A^*|^{v(u-1)}|A^{**}|^{v-1}}\right)^{n/2}.\end{aligned} \tag{2-4}$$

其中 Λ_a, Λ_b, Λ_c 分别是定义在 H_0, $H_{0b|a}$, $H_{0c|a}$ 下的似然比统计量, 由 Anderson (2003) [2] 中第 9 章的第 2 节和第 10 章的第 2 节中的定义可知

$$\begin{aligned}\Lambda_a &= \left(\frac{|A|}{\prod_{j=1}^{uv}|A_j|}\right)^{n/2}, \\ \Lambda_b &= \left((vu-v)^{mv(u-1)}\frac{\prod_{l=1}^{v}\prod_{k=1}^{u-1}|A_{(l-1)u+1+k}|}{|A^*|^{v(u-1)}}\right)^{n/2}, \\ \Lambda_c &= \left((v-1)^{m(v-1)}\frac{\prod_{k=1}^{v-1}|A_{ku+1}|}{|A^{**}|^{v-1}}\right)^{n/2}, \\ A &= \Gamma^{\bullet}\Gamma^{*}\boldsymbol{A}^{+}\Gamma^{*\prime}\Gamma^{\bullet\prime}, \\ &A_j \text{ 是 } A \text{ 的第 } j \text{ 个维数为 } m\times m \text{ 的对角块,} \\ A^* &= \sum_{l=1}^{v}\sum_{k=1}^{u-1}A_{(l-1)u+1+k}, \quad A^{**} = \sum_{k=1}^{v-1}A_{ku+1}.\end{aligned}$$

且由 Anderson (2003) [2] 中引理 10.3.1 可知, Λ_a, Λ_b, Λ_c 相互独立.

当样本维数 uvm 相对固定, 在原假设 H_0 成立时, 可由 Box (1949) [6] 中介绍的卡方近似方法和 Coelho & Roy (2020) [13] 中介绍的似然比检验统计量的改进精确卡方渐近分布 (CHI 方法) 得到

$$-2\rho\log\Lambda \sim (1-w_1)\chi^2_f + w_1\chi^2_{f+4}. \tag{2-5}$$

其中参数 ρ, f, w_1 分别为

$$\begin{aligned}(1)\ \rho = 1-\frac{1}{f}\Bigg\{&\sum_{k=1}^{uvm}x_{ak}^{-1}Q(\xi_{ak}) + \sum_{k=1}^{v(u-1)}\sum_{j=1}^{m}x_{bkj}^{-1}Q(\xi_{bkj}) + \sum_{k=1}^{v-1}\sum_{j=1}^{m}x_{ckj}^{-1}Q(\xi_{ckj}) \\ &-\sum_{k=1}^{uv}\sum_{j=1}^{m}y_{ajk}^{-1}Q(\eta_{ajk}) - \sum_{j=1}^{m}y_{bj}^{-1}Q(\eta_{bj}) - \sum_{j=1}^{m}y_{cj}^{-1}Q(\eta_{cj})\Bigg\},\end{aligned}$$

其中

$$x_{ak} = x_{bkj} = x_{ckj} = y_{ajk} = \frac{n}{2}, \quad y_{bj} = \frac{nv(u-1)}{2}, \quad y_{cj} = \frac{n(v-1)}{2}.$$

$$(2)\ f=-2\left\{\sum_{k=1}^{uvm}\xi_{ak}+\sum_{k=1}^{v(u-1)}\sum_{j=1}^{m}\xi_{bkj}+\sum_{k=1}^{v-1}\sum_{j=1}^{m}\xi_{ckj}-\sum_{k=1}^{uv}\sum_{j=1}^{m}\eta_{akj}\right.$$
$$\left.-\sum_{j=1}^{m}\eta_{bj}-\sum_{j=1}^{m}\eta_{cj}-\frac{1}{2}(v(u-1)m-m)-\frac{1}{2}((v-1)m-m)\right\},$$

其中

$$\xi_{ak}=\xi_{bkj}=\xi_{ckj}=\eta_{ajk}=-\frac{j}{2},\quad \eta_{bj}=-\frac{j+v(u-1)-1}{2},\quad \eta_{cj}=-\frac{j+v-2}{2}.$$

$$(3)\ w_1=-\frac{1}{6\rho^2}\left\{\sum_{k=1}^{uvm}x_{ak}^{-2}Q(\xi_{ak}^*)+\sum_{k=1}^{v(u-1)}\sum_{j=1}^{m}x_{bkj}^{-2}Q(\xi_{bkj}^*)+\sum_{k=1}^{v-1}\sum_{j=1}^{m}x_{ckj}^{-2}Q(\xi_{ckj}^*)\right.$$
$$\left.-\sum_{k=1}^{uv}\sum_{j=1}^{m}y_{ajk}^{-2}Q(\eta_{ajk}^*)-\sum_{j=1}^{m}y_{bj}^{-2}Q(\eta_{bj}^*)-\sum_{j=2}^{m}y_{cj}^{-1}Q(\eta_{cj}^*)\right\},$$

其中

$$\xi_{ak}^*=(1-\rho)x_{ak}+\xi_{ak},\quad \xi_{bkj}^*=(1-\rho)x_{bkj}+\xi_{bkj},\quad \xi_{ckj}^*=(1-\rho)x_{ckj}+\xi_{ckj},$$
$$\eta_{ajk}^*=(1-\rho)y_{ajk}+\eta_{ajk},\quad \eta_{bj}^*=(1-\rho)y_{bj}+\eta_{bj},\quad \eta_{cj}^*=(1-\rho)y_{cj}+\eta_{cj}.$$

且 $Q(\cdot)$ 是二次多项式: $Q(x)=x^2-x+\frac{1}{6}$.

在现代社会中, 由于数据容量和数据维数不断地扩大, 基于固定维数的经典卡方近似方法已逐渐失效. 因此针对高维的数据特征, 我们需要寻求新的方法来解决似然比统计量的渐近分布问题.

§2.3 主要结果

本文主要研究的是当维数 N 与样本容量 p 都趋于无穷时, 似然比统计量的渐近分布性质. 其主要思路是从似然比统计量的高阶矩出发, 然后利用矩母函数/特征函数的连续性定理等性质来求解统计量的渐近性质和相关结果. 接下来的四个小节, 将逐次给出本文的主要结果.

§2.3.1 似然比检验统计量矩的表达式

首先针对前文 (2-4) 式给出的似然比统计量 Λ, 这里给出其矩的表达式的结果如下:

定理 2.3.1 设 Λ 由 *(2-4)* 式定义. 那么, 当原假设 H_0 成立时, 对任意的 $h>0$, 有

$$E(\Lambda^h)=[v(u-1)]^{mnv(u-1)h/2}(v-1)^{mn(v-1)h/2}\times\prod_{k=1}^{uv-1}\frac{\Gamma_m\left(\frac{n-(uv-k)-1}{2}+\frac{n}{2}h\right)}{\Gamma_m\left(\frac{n-(uv-k)}{2}\right)}$$
$$\times\frac{\Gamma_m\left(\frac{(n-1)v(u-1)}{2}\right)}{\Gamma_m\left(\frac{(n-1)v(u-1)}{2}+\frac{nv(u-1)}{2}h\right)}\times\frac{\Gamma_m\left(\frac{(n-1)(v-1)}{2}\right)}{\Gamma_m\left(\frac{(n-1)(v-1)}{2}+\frac{n(v-1)}{2}h\right)}, \tag{2-6}$$

其中 $\Gamma_m(x)$ 为 m 元伽马函数 *(其定义可参见 Murihead[31] 中定理 2.1.12).*

根据似然比统计量 Λ 矩的性质, 在接下来的三小节中, 本文将分别从矩母函数的连续性定理, 特征函数的连续性定理等出发, 给出似然比统计量 Λ 的渐近性质及一些相关结果.

§2.3.2 基于高阶伽马函数展开方法的渐近正态性

伽马函数 (Gamma Function) 是统计学中广泛使用的函数之一, 是阶乘函数在实数与复数上扩展的一类函数. 在高维情况下, 尤其是在多元统计当中, 多元伽马函数 Γ_p 广泛地应用于 Wishart 分布和逆 Wishart 分布的概率密度函数以及矩阵变量贝塔分布中, 比如: Muirhead(1982) [31], Anderson (2003) [2] 中有详细的应用. 本节中, 我们采用多元伽马函数的展开形式, 这个方法广泛地使用于 Bai et al. (2009) [3], Jiang et al. (2012) [21], Jiang & Yang (2013) [22], Jiang & Qi (2015) [23] 中.

在本小节中, 基于原检验 H_0, 在样本容量 N 和样本维数 p 趋于无穷时研究了似然比统计量的渐近性质. 根据定理 2.3.1 中似然比统计量 Λ 的矩的结果, 基于矩母函数的一致连续性和高阶伽马函数展开的方法来寻求似然比统计量的渐近性质. 下面给出本小节的定理结果.

定理 2.3.2 假定 $u=u(n)$, $v=v(n)$, $m=m(n)$ 是三列依赖于 $n=N-1$ 的正整数序列, 使得 $uvm<n$ 对任意的 $n>4$ 都成立, 并且满足

$$u=u(n)\to\infty,\quad v=v(n)\to\infty,\quad m=m(n)\to\infty,\quad \frac{uvm}{n}\to y\in(0,1],$$

设 Λ 为 *(2-4)* 式中所定义的似然比统计量. 当 H_0 成立, 在 $n\to\infty$ 时, 有

$$\frac{\log\Lambda-\mu_n}{\sigma_n}\xrightarrow{d}\boldsymbol{N}(0,1), \tag{2-7}$$

其中

$$\begin{aligned}\mu_n = &-(m-(n-1)v(u-1)+\frac{1}{2})v(u-1)\log(1-\frac{m}{(n-1)v(u-1)})\\ &-(m-(n-1)(v-1)+\frac{1}{2})(v-1)\log(1-\frac{m}{(n-1)(v-1)})\\ &+(-n+uvm+\frac{3}{2})\log(\frac{n-uvm-1}{n-m-1})+m(uv-1)\log(\frac{n-m-1}{n-1}),\end{aligned} \tag{2-8}$$

$$\begin{aligned}\sigma_n^2 = &2[v(u-1)]^2\log(1-\frac{m}{(n-1)v(u-1)})-2\log(\frac{n-uvm-1}{n-m-1})\\ &+2(v-1)^2\log(1-\frac{m}{(n-1)(v-1)}).\end{aligned} \tag{2-9}$$

接下来考虑弱化定理 2.3.2 的假设条件. 定理 2.3.2 证明了 $uvm/n \to y \in (0,1]$ 的特殊情况, 接下来仅假设 $uvm < n$ 对于任意 $n > 4$ 成立, 即在 $uvm/n \to y \in [0,1]$ 成立时, 我们可以得到更一般的结果:

定理 2.3.3 *假定 $u = u(n)$, $v = v(n)$, $m = m(n)$ 是三列依赖于 $n = N-1$ 的正整数序列, 使得 $uvm < n$ 对所有的 $n > 4$ 都成立. 设 Λ 是 (2-4) 式所定义的统计量. 假设*

$$u = u(n) \to \infty, \quad v = v(n) \to \infty, \quad m = m(n) \to \infty,$$

那么, 在原假设 H_0 成立, 当 $n \to \infty$ 时, 有

$$\frac{\log\Lambda - \mu_n}{\sigma_n} \xrightarrow{d} \boldsymbol{N}(0,1),$$

其中 μ_n 和 σ_n^2 由定理 2.3.2 中定义 (2-8) 式和 (2-9) 式所示.

注记 2.3.4 *定理 2.3.2 和定理 2.3.3 中关于似然比统计量的期望 μ_n 和 σ_n 是相同的, 定理 2.3.3 相较于定理 2.3.2 在计算方面更具有挑战性. 但是从理论分析的角度来看, 定理 2.3.2 和定理 2.3.3 中关于似然比统计量的渐近性质都具有很好的实践意义, 这两种方法的适用性基本一致.*

§2.3.3 基于服从贝塔分布的随机变量乘积展开的渐近正态性

贝塔分布 (Beta Distribution) 是常见的几种分布之一, 其中多元贝塔分布与与 Wishart 分布密切相关. 在 Muirhead(1982) [31], Anderson (2003) [2] 中介绍的许多假设检验问题中, 基于原假设将似然比统计量的矩的形式转化为贝塔分布随机变量乘积的矩的形式, 进而将

复杂的似然比统计量的形式转化为较为直观的贝塔分布乘积的形式. 在 Dette & Tomecki (2019) [15] 与 Dette & Dörnemann (2020) [14] 中基于这个思路对似然比统计量进行了渐近估计.

在本小节中, 基于原检验 H_0, 在样本容量 N 和样本维数 p 趋于无穷时研究了似然比统计量 Λ 的渐近性质. 根据定理 2.3.1 中统计量 Λ 的矩的结果, 主要采用了似然比统计量矩的函数和 Lindeberg 条件的方法来寻求似然比统计量的渐近性质. 下面给出本小节的定理结果.

定理 2.3.5 假定 $u = u(n)$, $v = v(n)$, $m = m(n)$ 是三列依赖于 $n = N - 1$ 的正整数序列, 使得 $uvm < n$ 对任意的 $n > 4$ 都成立, 并且满足

$$u = u(n) \to \infty, \quad v = v(n) \to \infty, \quad m = m(n) \to \infty, \quad \frac{uvm}{n} \to y \in (0, 1],$$

设 Λ 为 *(2-4)* 式中所定义的似然比统计量. 当 H_0 成立, 在 $n \to \infty$ 时, 有

$$\frac{\log \Lambda - \tilde{\mu}_n}{\sigma_n} \xrightarrow{d} \boldsymbol{N}(0, 1). \tag{2-10}$$

其中

$$\begin{aligned}
\tilde{\mu}_n = & -(m - (n-1)v(u-1) + \frac{1}{2})v(u-1)\log(1 - \frac{m}{(n-1)v(u-1)}) \\
& - (m - (n-1)(v-1) + \frac{1}{2})(v-1)\log(1 - \frac{m}{(n-1)(v-1)}) \\
& + (-n + uvm + \frac{3}{2})\log(\frac{n - uvm - 1}{n - m - 1}) + m(uv-1)\log(\frac{n-m-1}{n-1}) \\
& + v^2(u-1)^2\log[1 - \frac{m}{(n-1)v(u-1)}] + (v-1)^2\log[1 - \frac{m}{(n-1)(v-1)}],
\end{aligned} \tag{2-11}$$

$$\begin{aligned}
\sigma_n^2 = & 2[v(u-1)]^2\log(1 - \frac{m}{(n-1)v(u-1)}) - 2\log(\frac{n - uvm - 1}{n - m - 1}) \\
& + 2(v-1)^2\log(1 - \frac{m}{(n-1)(v-1)}).
\end{aligned} \tag{2-12}$$

注记 2.3.6 第 *2.3.2* 节与本小节在一个类似的框架下考虑问题. 从理论实践的意义来说, 这两种方法是从不同的角度来分析基于双可交换协方差矩阵结构的似然比统计量的渐近形式. 同时可以很容易地看出

$$\lim_{n \to \infty} (\tilde{\mu}_n - \mu_n) = 0,$$

表明这两种方法求出的渐近分布的期望和方差在极限状态是相等的.

§2.3.4 基于 Edgeworth 展开方法的近似分布和一致误差界

Edgeworth 展开方法 (Edgeworth Expansion) 关键思想在于写出分布的特征函数, 其概率密度函数 f 可以根据具有已知和适当的分布的特征函数来近似, 并通过傅里叶逆变换来求解 f, 其优点在于可以控制误差. Edgeworth 方法的思路最早是由 Edgeworth (1896) [17] 等研究者在解决独立随机变量和的问题时提出的. 近些年由于在高维领域具有适用性, 所以受到了广大研究者的关注和研究. 不过, 由于 Edgeworth 展开的多项式可能过于复杂而难以计算, 所以需要进一步的处理, 在 Withers (1983) [54], Withers (1984) [55] 中, 针对这一问题, 研究者进行了优化. 如今, 由于 Edgeworth 展开方法可以通过对特征函数的逆变换来逼近真实分布, 所以具有很高的实用价值, 一些研究成果也可以参考 Tonda & Fujikoshi (2004) [49], Wakaki (2006) [51], Wakaki (2007) [52].

在本小节中, 基于原检验 H_0, 在样本容量 N 和样本维数 p 趋于无穷时研究了似然比统计量 Λ 的渐近性质. 主要基于泰勒展开, 特征函数的逆变换和傅里叶逆变换的方法来寻求似然比统计量 Λ 的渐近性质和一致误差界.

在本节的定理之前, 我们首先介绍一些本节需要用到的表达式:

$$\begin{aligned}\hat{\mu}_n = &-\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi(\frac{n-(uv-k)m-j}{2})+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)})\\&+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi(\frac{n-1}{2}-\frac{j-1}{2(v-1)}),\end{aligned} \tag{2-13}$$

$$\begin{aligned}\hat{\sigma}_n^2 = &\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi^{(1)}(\frac{n-(uv-k)m-j}{2})+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi^{(1)}(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)})\\&+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi^{(1)}(\frac{n-1}{2}-\frac{j-1}{2(v-1)}),\end{aligned} \tag{2-14}$$

其中 ψ 是对数伽马函数的导数, 且满足

$$\psi^s(a)=(\frac{d}{dx})^{s+1}\log\Gamma(a)=\begin{cases}\mathbb{C}+\sum_{k=0}^{\infty}(\frac{1}{1+k}-\frac{1}{k+a}) & s=0,\\ \sum_{k=0}^{\infty}\frac{(-1)^{s+1}s!}{(k+a)^{s+1}} & s=1,2,\cdots,\end{cases}$$

其中 $\mathbb{C}$ 为欧拉常数.

$$\kappa^{(s)} = (-1)^s \left\{ \sum_{j=1}^{m} \sum_{k=1}^{uv-1} \psi^{(s-1)}[\frac{n-(uv-k)m-j}{2}] + \sum_{j=1}^{m} \sum_{k=1}^{v(u-1)} \psi^{(s-1)}[\frac{n-1}{2} - \frac{j-1}{2v(u-1)} + \frac{k-1}{v(u-1)}] + \sum_{j=1}^{m} \sum_{k=1}^{v-1} \psi^{(s-1)}[\frac{n-1}{2} - \frac{j-1}{2(v-1)}] \right\},$$

以及

$$\gamma_{a,s} = \sum_{s_1+\cdots+s_a=s} \prod_{l=1}^{a} \frac{\kappa^{(s_l+3)}}{(s_l+3)!\sigma_n^{s_l+3}}.$$

由此, 可得到下面的结果:

定理 2.3.7 假定 $u = u(n)$, $v = v(n)$, $m = m(n)$ 是三列依赖于 $n = N - 1$ 的正整数序列, 使得 $u > 1$, $v > 1$, $m > 1$, $uvm < n$ 对所有的 $n > 4$ 都成立. 令 $w = (n - (uv + \frac{1}{2}) - \frac{3}{2})/2$, 并且定义

$$Z = \frac{T - \hat{\mu}_n}{\hat{\sigma}_n},$$

其中 $T = -\frac{2}{n}\log\Lambda$. 那么, 当 $n \to \infty$ 时, 有

$$P(Z \le x) = \Phi_l(x) + O\left(\frac{1}{w^{s+1}}\right), \tag{2-15}$$

其中

$$\Phi_l(x) = \Phi(x) - \phi(x)\left[\sum_{a=1}^{l} \frac{1}{a!} \sum_{s=0}^{l-a} \gamma_{a,s} h_{3a+s-1}(x)\right], \tag{2-16}$$

并且 $h_s(x)$ 为 s 阶 *Hermite* 多项式, 其定义为

$$(\frac{d}{dx})^s \exp(-\frac{x^2}{2}) = (-1)^r h_s(x)\exp(-\frac{x^2}{2}), \quad s = 1, 2, \cdots. \tag{2-17}$$

进一步地, 这里给出定理 2.3.7 中统计量 Z 的渐近分布的一致误差界.

定理 2.3.8 在定理 2.3.7 的假设下, 对所有的 $0 < w < \hat{\sigma}_n/2$, 有

$$\sup_x |P(Z \le x) - \Phi_l(x)| < \frac{1}{2\pi}[U_1(r) + U_2(r) + U_3(r)],$$

其中 $\Phi_l(x)$ 由 *(2-16)* 式定义, 且

$$U_1[r]=\frac{1}{w^{l+1}}\left\{\sum_{a=1}^{l}\frac{1}{a!}R_{a,l-a+1}[r](\frac{1}{2})^{-(l+2a+1)/2}\Gamma(\frac{l+2a+1}{2})\right.$$
$$\left.+\frac{1}{(l+1)!}(B[r])^{l+1}(\frac{C_r}{2})^{-(3l+3)/2}\Gamma(\frac{3l+3}{2})\right\},$$

$$U_2[r]=\frac{8}{m[(u^3v^3-3)m+uv-3]}(1+\alpha)^{-\frac{1}{8}m[(u^3v^3-3)m+uv-3]}\frac{1+\alpha}{\alpha},$$

$$U_3[r]=\exp(-\frac{w^2r^2}{2}(1-c))\left\{1+\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}(\frac{c}{2})^{-(3a+s)/2}\Gamma(\frac{3a+s}{2})\gamma_{a,s}\right\}.$$

其中

$$B[r]=\sum_{s=0}^{\infty}b_sr^s,\ C_r=1-2rB[r],\ \alpha=(\frac{wr}{n})^2,\ c\ \text{为常数且}\ 0<c<1.$$

§2.4 本章小结

在本章中, 依次给出了多元实正态分布总体下的双可交换协方差矩阵的似然比统计量的一系列性质.

首先, 基于原检验 H_0, 在第一节中求出似然比统计量的高阶矩的具体表达形式, 然后以此作为余下三节的出发点. 在第二节中, 根据 Jiang & Yang (2013) [22] 和 Jiang & Qi (2015) [23] 中提出的针对高维数据的似然比统计量渐近正态性的方法, 运用高阶伽马函数展开和矩母函数的一致有界性, 解决了在 $uvm\to\infty$, $n\to\infty$, 且 $uvm<n$ 时, 基于双交换协方差矩阵结构的似然比统计量的渐近正态性. 在第三节中, 通过将似然比统计量转化为服从独立贝塔分布随机变量的乘积的形式来求解统计量的渐近性质, 其主要思路来自于 Dette & Dõrnemann (2020) [15] 中针对无穷分组的统计量的渐近正态估计. 在第四节中, 通过 Edgeworth 展开的方法, 同时使用特征函数的逆转函数对似然比统计量进行渐近估计和误差界的一致估计, 这个方法广泛的使用于 Tonda & Fujikoshi (2004) [49], Wakaki (2006) [51], Wakaki (2007) [52], Fujikoshi et al. (2008) [18], Akita et al. (2010) [1] 等文献当中.

通过对三种方法的对比, 我们可以看出: (1) 对于前两种方法, 我们基本可以视作为同一个结果的两个不同求解方法, 其计算过程较为简单与直观, 以实用角度来看基本一致, 具有很好的实用价值. (2) 第三种方法, 通过 Edgeworth 展开, 我们可以通过控制参数 l 的大小来控制分布的精确程度, 但是 l 并不是越大越好. 同时第三种方法相较于前两种方

法, 我们可以计算渐近分布与真实分布之间的误差界, 但是相对而言, 计算量和计算难度也要高于前两种方法.

综上所述, 基于原检验 H_0 和似然比统计量 Λ, 本文采用了三种方法对其求解渐近分布. 在接下来的章节中, 我们将会通过数值模拟和实例分析来验证模型的有效性.

第三章　数值研究

在这一章, 我们关注的重点是所提出的似然比统计量在数值模拟和实例分析上的具体表现情况.

§3.1　数值模拟

在本节中, 我们将使用 Monte Carle 方法 [39] 通过一个有限样本来对比传统卡方近似和本文提出的三种方法在处理双可交换协方差矩阵结构假设检验问题时的表现. 我们将用 10000 次独立迭代运行模拟来估计; 同时为了简化计算的复杂度, 我们采用有限的固定的样本容量 N 和变动的样本维数 $p(=uvm)$. 通过图像与表格, 可以看出当样本维数 p 逐渐增大并接近样本容量 N 时, 本文提出的渐近分布方法的拟合效果要比卡方近似的拟合效果好.

首先, 我们将绘制卡方分布密度曲线和标准正态分布密度曲线来跟频率直方图进行直观地比较. 我们选择样本容量 $N=100$, 且样本维数 $p=10,\ 30,\ 50,\ 70,\ 90$.

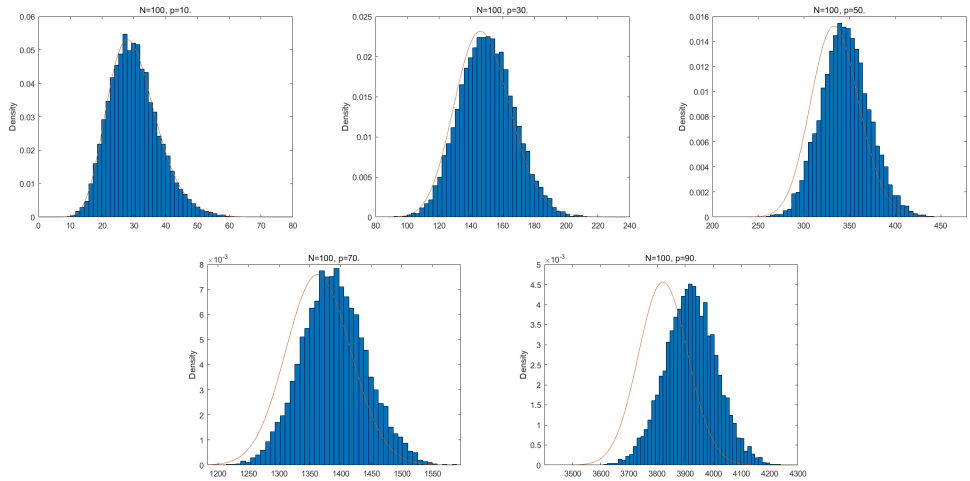

图 3-1　CHI 中 $-2\rho\log\Lambda$ 频率分布直方图与卡方分布密度曲线对比

图 3–1 显示, 在样本容量固定为 N 时, 在样本维数 p 较小时卡方近似方法的频率直方图与卡方密度函数曲线的拟合度很好, 但是在随着维数 p 的逐渐增大频率直方图与密度函数曲线的拟合度越来越差. 这表明传统的卡方近似方法在处理高维数据时是失效的.

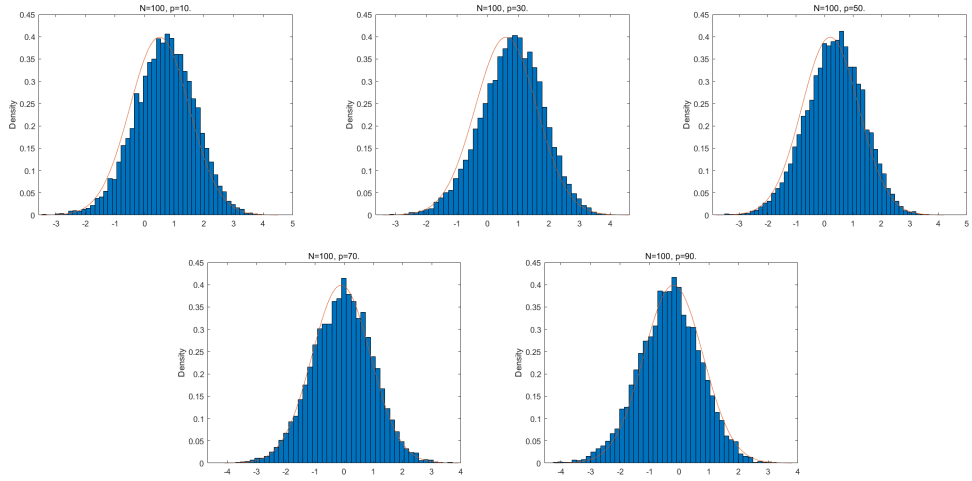

图 3-2 HGM 中 $(\log\Lambda-\mu_n)/\sigma_n$ 频率分布直方图与标准正态分布概率密度曲线对比

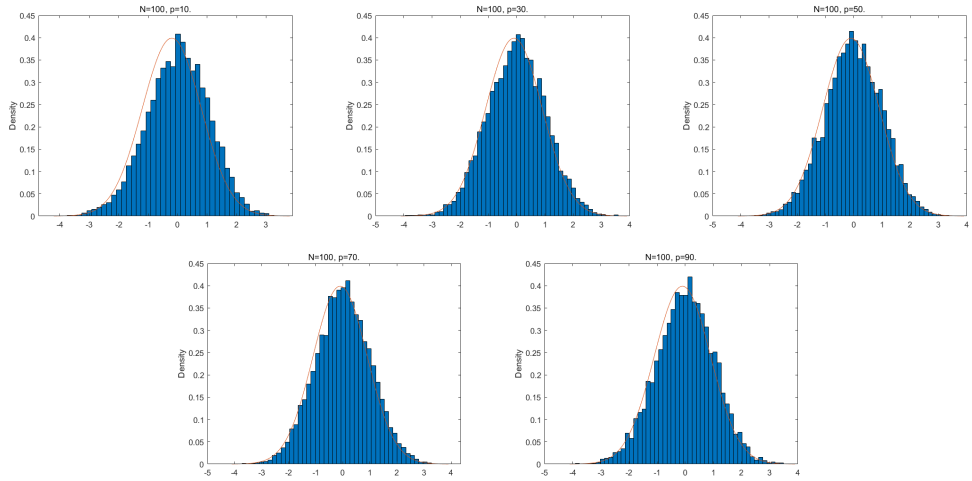

图 3-3 HBM 中 $(\log\Lambda-\tilde{\mu}_n)/\tilde{\sigma}_n$ 频率分布直方图与标准正态分布概率密度曲线对比

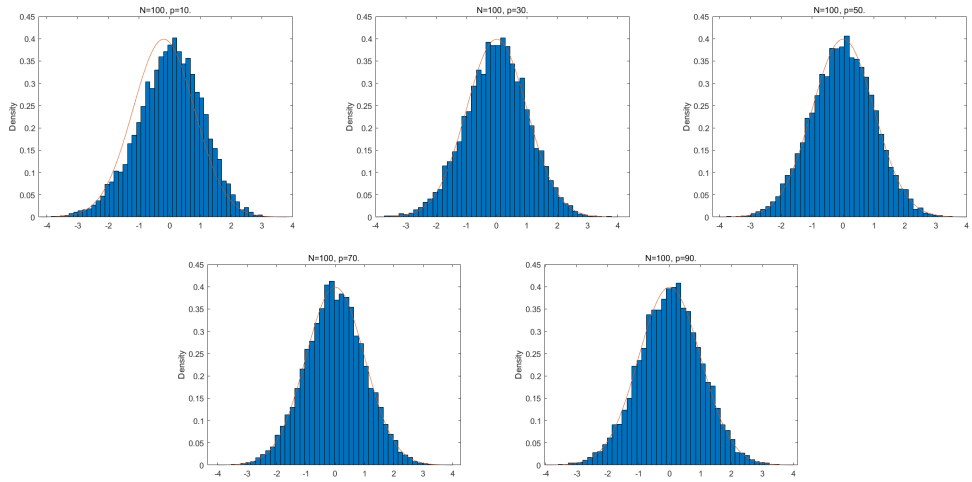

图 3-4 HEM 中 Z 频率分布直方图与标准正态分布概率密度曲线对比

图 3–2, 3–3, 3–4 显示, 在样本容量固定为 N 时, 无论样本维数 p 是高是低, 本文提出的三种高维近似方法的频率直方图与标准正态分布密度函数曲线的拟合度很好. 这也表明这些方法方法在处理高维数据时是有效的.

接着, 我们使用本文中介绍的四种方法来检验双可交换协方差矩阵结构: 在 (2-5) 式中提到的传统卡方近似 (CHI 方法), 在定理 2.3.2 中的高阶伽马函数渐近展开方法 (HGM 方法), 在定理 2.3.5 中的转为为贝塔分布随机变量乘积的展开方法 (HBM 方法), 在定理 2.3.7 中的高阶 Edgeworth 展开方法 (HEM 方法). 对于双可交换协方差矩阵结构, 我们选择 $X_1, \cdots, X_N$ 是来自样本容量 $N = 200$, 且分布为 $N(\mu_1, \Sigma_1)$ 的随机样本, 其中:

$$\mu_1 = 0, \quad \Sigma_1 = I_{uvm}.$$

为了更全面的比较三种方法, 我们给定显著性水平 $\alpha = 0.05$. 在 u, v, m 不同的情况下, 根据 10000 次独立实验得出四种方法的经验 Size 和 Power 值. 如表 3–1 所示,

表 3–1: 在给定显著性水平 $\alpha = 0.05$ 时,似然比统计量的 Size 和 Power

	Size				Power			
	CHI	HGM	HBM	HEM	CHI	HGM	HBM	HEM
u=2, v=3, m=5	0.0496	0.0618	0.0598	0.0548	0.7340	0.4513	0.5156	0.5627
u=2, v=5, m=3	0.0492	0.0581	0.0524	0.0474	0.7399	0.5134	0.5256	0.5782
u=5, v=2, m=3	0.0501	0.0609	0.0514	0.0489	0.7455	0.5241	0.5256	0.4682
u=2, v=3, m=10	0.0687	0.0507	0.0456	0.0507	0.7851	0.5234	0.4423	0.6445
u=2, v=10, m=3	0.0724	0.0550	0.0522	0.0497	0.7809	0.7681	0.7731	0.7111
u=10, v=2, m=3	0.0758	0.0552	0.0524	0.0523	0.7476	0.7584	0.7637	0.7412
u=2, v=3, m=15	0.1283	0.0512	0.0521	0.0487	0.8735	0.7740	0.7145	0.7521
u=2, v=15, m=3	0.1277	0.0552	0.0563	0.0493	0.8387	0.7112	0.7271	0.6251
u=15, v=2, m=3	0.1234	0.0552	0.0489	0.0504	0.8897	0.7256	0.7251	0.7241
u=2, v=3, m=20	0.5811	0.0513	0.0503	0.0522	0.9022	0.7411	0.7626	0.7178
u=2, v=20, m=3	0.6131	0.0536	0.0524	0.0512	0.9256	0.7134	0.7415	0.7167
u=20, v=2, m=3	0.6565	0.0457	0.0501	0.0527	0.9121	0.7513	0.8134	0.7424
u=2, v=3, m=5	0.9964	0.0536	0.0524	0.0513	0.9431	0.8514	0.8567	0.8123
u=2, v=25, m=3	0.9816	0.0509	0.0514	0.0501	0.9730	0.8654	0.8634	0.8143
u=25, v=2, m=3	0.9991	0.0524	0.0501	0.0489	0.9714	0.8532	0.8765	0.8567
u=2, v=3, m=30	1.000	0.0517	0.0499	0.0523	0.9453	0.9147	0.9264	0.9434
u=2, v=30, m=3	1.000	0.0528	0.0539	0.0513	0.9700	0.9434	0.9821	0.9234
u=30, v=2, m=3	1.000	0.0521	0.0555	0.0505	0.9552	0.9264	0.9449	0.9225

从表 3–1 中可以看出, 当样本维数 uvm 很小时, 卡方近似方法的表现更好一点. 一个共同特征是当样本维数 uvm 很小时 (如 $u=2$, $v=3$, $m=5$, $uvm=30$), 基于卡方近似方法的 Size 值与给定的显著性水平 $\alpha=0.05$ 十分接近, 这表明卡方近似方法在低维情况是有效的, 且其表现要优于本文提出的 HGM 方法, HBM 方法和 HEM 方法. 但随着样本维数 uvm 逐渐地增大, 卡方近似方法的 Size 值开始迅速增大, 并与给定显著性水平 α 差异较大, 这说明卡方近似方法将总是拒绝原假设. 在样本维数 uvm 逐渐增大时, HGM 方法, HBM 方法和 HEM 方法的 Size 值大小是相对稳定, 其值稳定在给定显著性水平 $\alpha=0.05$ 附近. 当样本维数 uvm 逐渐增大时, 卡方近似方法的 Power 值要比其他三种方法的 Power 值要大, 这主要是因为在样本维数 uvm 接近样本容量 N 的时候, HGM 方法, HBM 方法和 HEM 方法的收敛速率变差了. 因此, 在处理高维双可交换协方差矩阵结构的假设检验问题时, 本文提出的三种方法都是有效的, 且效果优于卡方近似方法.

最后, 由于在上一个随机模拟中选取的分布过于特殊, 因此接下来我们选择样本容量为 $N=1000$, 使用显著性水平 $\alpha=0.05$, 根据 10000 次独立实验得出四种方法的经验 Size 和 Power 值. 且 $X_1,\cdots,X_N$ 是分布为 $N(\mu_2,\Sigma_2)$ 的一个样本, 其中:

$$\mu_2=0,\qquad \Sigma_2=I_v\otimes(I_u\otimes U_0+(J_u-I_u)\otimes U_1)+(J_v-I_v)\otimes(J_u\otimes W),$$

且

$$U_0=\begin{pmatrix}1&0&\cdots&0\\0&2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&m\end{pmatrix}_{m\times m},\ U_1=\begin{pmatrix}2&1&\cdots&1\\1&2&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&2\end{pmatrix}_{m\times m},\ W=\begin{pmatrix}1&2&\cdots&2\\2&1&\cdots&2\\\vdots&\vdots&\ddots&\vdots\\2&2&\cdots&1\end{pmatrix}_{m\times m}.$$

因此与表 3–1 相似, 我们可以根据显著性水平 α 同理得出经验 Size–Power 表.

相较于表 3–1 中选取的特殊样本分布, 表 3–2 中选取的样本分布情况较为复杂和一般. 但是结果与表 3–1 类似, 在样本维数 uvm 远远小于样本容量 N 时, 卡方近似方法的 Size 值表现比 HGM 方法, HBM 方法和 HEM 方法好; 但卡方近似方法的 Power 值却一直比其他三种方法的 Power 值大, 更逼近 1.0000, 这主要是因为 HGM 方法, HBM 方法和 HEM 方法的收敛速率在样本维数接近样本容量时变差了. 在 $uvm>100$ 时, HGM 方法, HBM 方法和 HEM 方法的 Size 值和 Power 值就表现得比卡方近似方法好, Size 值稳定在给定的显著性水平 $\alpha=0.05$ 附近, 且三种方法的 Size 值基本保持在同一水平, 在高维情况下相较于卡方近似方法有很大的提升; 但是通过对比 Power 值可以发现, 第一

种方法 HGM 总体表现得比其他两种方法都好. 由于样本维数 uvm 是由 u , v , m 相乘得出的, 所以通过对比三个变量在给定 uvm 值时对 Size–Power 值的影响, 可以看出, 不同的 u, v, m 并不会对 Size–Power 值产生较大影响, 其值基本处于同一的水平线上, 这表明三个参数仅仅在整体乘积方面对渐近分布的效果产生影响. 因此, 在处理高维双可交换协方差矩阵结构的假设检验问题时, 本文提出的三种方法都是效的, 其单个的 u, v, m 并不会对效果产生重大的影响.

表 3–2: 在给定显著性水平 $\alpha = 0.05$ 时,似然比统计量的 Size 和 Power

	Size				Power			
	CHI	HGM	HBM	HEM	CHI	HGM	HBM	HEM
u=2, v=3, m=5	0.0523	0.0574	0.0563	0.0515	0.6342	0.6521	0.5432	0.5121
u=2, v=5, m=3	0.0514	0.0543	0.0586	0.0471	0.6234	0.5241	0.5264	0.5212
u=5, v=2, m=3	0.0553	0.0512	0.0499	0.0533	0.7023	0.5534	0.5563	0.5624
u=3, v=3, m=10	0.0489	0.0525	0.0586	0.0536	0.7531	0.7852	0.6234	0.6234
u=3, v=10, m=3	0.0501	0.0553	0.0573	0.0537	0.7309	0.7532	0.7464	0.7131
u=10, v=3, m=3	0.0521	0.0582	0.0521	0.0524	0.7621	0.8653	0.7234	0.7234
u=3, v=3, m=50	0.7941	0.0536	0.0546	0.0523	0.8324	0.5412	0.6639	0.7147
u=3, v=50, m=3	0.7747	0.0513	0.0512	0.0512	0.8532	0.4214	0.5324	0.7413
u=50, v=3, m=3	0.7872	0.0573	0.0523	0.0535	0.7432	0.6211	0.7436	0.7452
u=2, v=10, m=30	0.9534	0.0493	0.0531	0.0555	0.9124	0.8142	0.7834	0.7923
u=10, v=2, m=30	0.9765	0.0534	0.0525	0.0526	0.9643	0.7817	0.7743	0.8642
u=30, v=10, m=2	0.9632	0.0587	0.0489	0.0526	0.9413	0.7923	0.8252	0.8243
u=5, v=12, m=16	1.0000	0.0554	0.0512	0.0531	0.9872	0.8553	0.8964	0.8741
u=5, v=16, m=12	1.0000	0.0517	0.0532	0.0536	0.9432	0.9122	0.8717	0.8416
u=16, v=12, m=5	1.0000	0.0564	0.0511	0.0534	0.9400	0.8927	0.9224	0.9413

进一步, 对本文提出的三种方法进行比较. 从理论的角度看, HGM 方法和 HBM 方法得到的表达式较为简洁, 其期望和方差在极限处也是相等的, 通过模拟得出的 Size 值和 Power 值也是十分接近的. HEM 方法得出的表达式可以通过控制参数 l 来控制渐近分布的精度, 但是表达式比 HGM 方法和 HBM 方法更复杂. 通过对比表 3–1 和表 3–2 中的 Size 值可以发现, 在参数设定相同时, 基于 HEM 方法得到的 Size 值要比基于 HGM 方法和 HBM 方法得到的 Size 值略微小一点. 这说明采用 HGM 方法和 HBM 方法进行假设检验时比采用 HGM 方法犯第一类错误的概率要大些. 通过对比表 3–1 和表 3–2 中的 Power 值可以发现, 在参数设定相同时, 三种方法的 Power 值都随着样本维数 uvm 的

增大而增大, 但是采用 HGM 方法和 HBM 方法时得到的 Power 值要略微小于采用 HEM 方法时得到的 Power 值, 换句话说, 通过采用 HEM 方法得到的较大 Power 值是来自于牺牲第一类错误的准确性或测试集的大小.

综上图表所述, 本文提出的 HGM 方法, HBM 方法和 HEM 方法对于高维似然比统计量的拟合相较于卡方近似方法有很大的提升, 可以有效的解决高维双可交换协方差矩阵结构的假设检验问题. 且三种方法优势互补, 都具有很好的实践意义.

§3.2 实例分析

接下来, 我们将用两个实例来说明本文提出的似然比检验的效果.

§3.2.1 骨质疏松数据

为了解释说明我们提出的正态近似方法, 我们在一个真实的数据集上测试原假设. 原始数据由使用 GE-Lunar Prodigy 型双 X 射线骨密度仪获得的骨密度值组成. 选择 12 例患者 ($n = 12$) 进行骨密度测定研究. 测量得到的股骨径和股骨转子区 ($m = 2$), 左和右股骨 ($u = 2$). 由于股骨颈主要是皮质骨, 而股骨转子区本质上是松质骨, 所以这两种测量结果可以看作是两个不同的随机变量. 这四个测量位置大约两年后重新测量一次 ($v = 2$). 由此可知数据是具有三级多元的形式, 符合我们在第二章模型介绍中的假定, 同时在 Roy & Leiva (2011) [37] 中已经证明了此数据满足双可交换协方差矩阵结构的假定. 且针对数据集获得的样本协方差矩阵 A 如 Coelho & Roy (2020) [13] 中所叙述的那样:

$$A = \left[\begin{array}{cccc|cccc} \begin{bmatrix} 0.00623 & 0.00442 \\ 0.00442 & 0.00682 \end{bmatrix} & & 0.00432 & 0.00476 & & 0.00639 & 0.00489 & 0.00442 & 0.00400 \\ & & 0.00588 & 0.00854 & & 0.00479 & 0.00739 & 0.00575 & 0.00775 \\ 0.00432 & 0.00588 & \begin{bmatrix} 0.00789 & 0.00753 \\ 0.00753 & 0.01136 \end{bmatrix} & & 0.00485 & 0.00667 & 0.00767 & 0.00690 \\ 0.00476 & 0.00854 & & & 0.00517 & 0.00923 & 0.00707 & 0.01029 \\ \hline 0.00639 & 0.00479 & 0.00485 & 0.00517 & \begin{bmatrix} 0.00706 & 0.00553 \\ 0.00553 & 0.00865 \end{bmatrix} & & 0.00522 & 0.00461 \\ 0.00489 & 0.00739 & 0.00667 & 0.0093 & & & 0.00713 & 0.00920 \\ 0.00442 & 0.00575 & 0.00767 & 0.00707 & 0.00522 & 0.00713 & \begin{bmatrix} 0.00875 & 0.00753 \\ 0.00753 & 0.01090 \end{bmatrix} & \\ 0.00400 & 0.00775 & 0.00690 & 0.01029 & 0.00461 & 0.00920 & & \end{array}\right]$$

从矩阵 A 中我们可以看出, 股骨颈和股骨转子区的钙含量的方差–协方差矩阵 (U_0) 在第一年和第二年非常相似. 而且, 这两年左右股骨的之间的钙含量的协方差矩阵 (U_1) 也非常相似. 最后, 这两年间股骨两部分的钙含量的协方差矩阵 (W) 也相似. 因此, 对于总体协方差矩阵具有双可交换协方差矩阵结构的假设无法被拒绝.

因此, 由 (2-4) 式可知, 似然比统计量 Λ 为:

$$\Lambda = \left(2^4 \frac{|A|}{|A_1||A_2 + A_3||A_3|}\right)^{n/2},$$

其中 A_1, A_2, A_3, A_4 为 A 中的 2×2 的对角块矩阵. 所以 $\Lambda = 4.46637 \times 10^{-12}$.

在这个例子中, 我们首先使用传统的卡方分布来估计似然比统计量, 则有 $-2\log\Lambda \overset{d}{\sim} \chi^2_{[muv(muv+1)/2]-[3m(m+1)/2]} = \chi^2_{27}$, 因此我们可以得到其 p 值约为 0.00246, 结果表明卡方近似对于小样本可能是完全不充分的. 事实上, 即使对于相当大的样本, 卡方近似也可能导致完全不充分的 p 值. 接着我们使用定理 2.3.2 和定理 2.3.5 中提出的 HGM 方法和 HBM 方法来估计似然比统计量, 有其 p 值为 0.3401, 结果表明通过使用 HGM 方法和 HBM 方法我们接受原假设, 认为数据的协方差矩阵满足双可交换的结构. 这些结果清楚地表明, 传统的卡方近似(与 HGM 方法和 HBM 方法相比)导致对零假设的拒绝太多, 或者在一般情况下 p 值过低, 显然不适合实际目的.

§3.2.2　矿物含量数据

第二个例子来自于 Leiva & Roy (2011) [26], 研究人员使用此数据检验数据集是否具有双可交换协方差矩阵结构. 研究人员通过光子吸收法测量了骨骼 (桡骨、肱骨和尺骨) 中的矿物质含量, 用以研究膳食补充剂是否会减缓 24 名老年妇女的骨质流失. 测量记录了优势侧和非优势侧的三个部位的骨骼 (Johnson & Wichern (2007) [24], p.43), 很明显 $u = 2, m = 3$. 同时 Johnson & Wichern (2007) [24] 中 353 页给出了这 24 名女性在参与实验项目一年后同位置的骨骼矿物质含量. 因此, 将这两个数据集并排合并成一个新的数据集. 可知, 新数据集具有三层多元结构, 其中 $v = 2, u = 2, m = 3$. 同理于第一个例子, 由矩阵 A 可知 (详见 Coelho & Roy (2020) [13] Supplementary material 中附录 E.2), 我们可以考虑检验总体协方差矩阵是否具有双可交换协方差结构的假设, 且 $\Lambda = 6.18767 \times 10^{-35}$.

利用定理 2.3.6 中 Edgeworth 展开的方法. 令 $l = 2$, 拟合的渐近概率密度由 (2-15) 式给出, 我们得到渐进的 p 值为 0.000042. 因此, 我们应该拒绝协方差结构是双可交换协方差矩阵结构的零假设. 如果我们使用传统的卡方近似来拟合似然比统计量的分布, 我们会得到 $-2\log\Lambda \overset{d}{\sim} \chi^2_{[muv(muv+1)/2]-[3m(m+1)/2]} = \chi^2_{60}$, 这将得到一个 p 值为 1.08083×10^{-10}. 虽然对于同一个 α 水平, 这个 p 值会导致相同的决策, 即拒绝原假设, 但是通过对比可知, 卡方近似产生的 p 值与 Edgeworth 方法产生的 p 值相差甚远.

第四章　主要结果的证明

在这一章, 本文将依次给出定理 2.3.2, 定理 2.3.3, 定理 2.3.5, 定理 2.3.7, 定理 2.3.8 的技术性证明.

§4.1　预备引理

引理 4.1.1 令 $n > p = p_n$, $\gamma_n = [-\log(1-\frac{p}{n})]^{\frac{1}{2}}$, 假定在 $n \to \infty$ 时, 有 $s = s_n = O(\gamma_n^{-1})$ 和 $t = t_n = O(\gamma_n^{-1})$. 那么, 当 $n \to \infty$ 时, 有

$$\log \frac{\Gamma_p(\frac{n}{2}+t)}{\Gamma_p(\frac{n}{2}+s)} = p(t-s)(\log n - 1 - \log 2) + \gamma_n^2\big[(t^2-s^2) - (p-n+\frac{1}{2})(t-s)\big] + o(1).$$

这一结果是由 Jiang & Yang (2013) [22] 中的引理 5.4 给出的.

引理 4.1.2 令 $\{p = p_n \in \mathbb{N}; n \geq 1\}$, $\{m = m_n \in \mathbb{N}; n \geq 1\}$, $\{t_n \in \mathbb{R}; n \geq 1\}$, 且随着 $n \to \infty$ 时, 满足 (i) $p_n \to \infty$, $p_n = o(n)$; (ii) 存在 $\epsilon \in (0,1)$ 使得 $\epsilon \leq \frac{m_n}{n} \leq \epsilon^{-1}$ 对足够大的 n 都成立; (iii) $t = t_n = O(\frac{n}{p})$. 则, 当 $n \to \infty$ 时, 有

$$\log \frac{\Gamma_p(\frac{m-1}{2}+t)}{\Gamma_p(\frac{m-1}{2})} = \alpha_n t + \beta_n t^2 + \gamma_n(t) + o(1),$$

其中

$$\begin{aligned}
\alpha_n &= -[2p + (m-p-\frac{3}{2})\log(1-\frac{p}{m-1})];\\
\beta_n &= -[\frac{p}{m-1} + \log(1-\frac{p}{m-1})];\\
\gamma_n(t) &= p\big[(\frac{m-1}{2}+t)\log(\frac{m-1}{2}+t) - \frac{m-1}{2}\log\frac{m-1}{2}\big].
\end{aligned}$$

这一结果是由 Jiang & Qi (2015) [23] 中的引理 5.1 给出的.

引理 4.1.3 令 $T_n (n \in \mathbb{N})$ 为一组有限的集合, $\{(X_n(i))_{i\in T_n} | n \in \mathbb{N}\}$ 表示一组随机变量数组, $\{(g_n(i))_{i\in T_n} | n \in \mathbb{N}\}$ 表示一组权重系数. 当满足:

(A.1) 对于 $\forall n \in \mathbb{N}$, 随机变量 $(X_n(i))_{i\in T_n}$ 为独立的.

(A.2) 对于 $\forall i \in T_n,\ n \in \mathbb{N}$, 随机变量 $(X_n(i))_{i\in T_n}$ 为中心化的, 即 $E[X_n(i)] = 0$.

(A.3) 对于 $\forall n \in \mathbb{N}$, $\exists$ 常数 C, $E[X_n^4(i)] \le CE[X_n^2(i)]^2$.

(A.4) 在 $n \to \infty$ 时, $\sup_{i\in T_n} g_n^2(i)Var(X_n(i)) \to 0$.

(A.5) 存在一个常数 $\sigma^2 > 0$, 使得在 $n \to \infty$ 时, 有

$$\sum_{i\in T_n} g_n^2(i)Var(X_n(i)) \to \sigma^2.$$

则随机变量

$$Z := \sum_{i\in T_n} g_n(i)Var(X_n(i)),$$

收敛到一个期望为 0, 方差为 σ^2 的正态分布.

这一结果是由 Dette & Dõrnemann (2020) [15] 中定理 A.1 给出的.

§4.2 证明定理 2.3.2 和定理 2.3.3

定理 2.3.2 的证明:

因为对于任意一个随机变量 $\boldsymbol{X} \sim \boldsymbol{N}(0,1)$ 时, 关于 $\boldsymbol{X}$ 的矩母函数为 $E(e^{sx}) = e^{\frac{s^2}{2}}$. 根据 Resnick (2019) [34] 中定理 9.5.2 给出的 Levy 连续性定理可知, 若要证明 $\frac{\frac{2}{n}\log\Lambda - \mu_n}{\sigma_n} \xrightarrow{d} \boldsymbol{N}(0,1)$, 仅需证明当 $n \to \infty$ 时, 存在 $\delta > 0$, 使得

$$E\exp\left(\frac{\frac{2}{n}\log\Lambda - \mu_n}{\sigma_n}s\right) \to e^{\frac{s^2}{2}}, \tag{4-1}$$

对所有 $|s| < \delta$ 成立即可.

因为在 $n \to \infty$ 时, 对于任意的 $y \in (0,1)$, 则有 $\frac{m}{(n-1)v(u-1)} \to y_1 \in (0,1)$, $\frac{m}{(n-1)(v-1)} \to y_2 \in (0,1)$, $\frac{n-uvm-1}{n-m-1} \to y_3 \in (0,1)$, 有

$$\begin{aligned}\sigma_n^2 &= 2[v(u-1)]^2\log(1-\frac{m}{(n-1)v(u-1)})\\ &\quad + 2(v-1)^2\log(1-\frac{m}{(n-1)(v-1)}) - 2\log(\frac{n-uvm-1}{n-m-1})\\ &\to 2[v(u-1)]^2\log(1-y_1) + 2(v-1)^2\log(1-y_2) - 2\log(y_3) > 0,\end{aligned}$$

成立, 以及当 $y = 1$ 时, $\sigma_n^2 \to \infty$.

因此, 定义

$$\delta_0 := \inf\{\sigma_n; n \geq 4\} > 0.$$

令 $h = h_n = s/\sigma_n$, 则当$|s| < \frac{\delta_0}{2}$ 固定时, 有 $\{h_n : n > 4\}$ 有界以及 $|h_n| < \frac{1}{2}$ 对所有的 $n > 4$ 都成立. 由此, 即可利用 (2-6) 式中统计量的高阶矩的结果得

$$\begin{aligned} E[\exp(\frac{2}{n}h\log\Lambda)] = [v(u-1)]^{mv(u-1)h}(v-1)^{m(v-1)h}\prod_{k=1}^{uv-1}\frac{\Gamma_m(\frac{n-(uv-k)-1}{2}+h)}{\Gamma_m(\frac{n-(uv-k)}{2})} \\ \times\frac{\Gamma_m(\frac{(n-1)v(u-1)}{2})}{\Gamma_m(\frac{(n-1)v(u-1)}{2}+v(u-1)h)}\frac{\Gamma_m(\frac{(n-1)(v-1)}{2})}{\Gamma_m(\frac{(n-1)(v-1)}{2}+(v-1)h)}. \end{aligned} \tag{4-2}$$

取定 $A_{(n-1)v(u-1)} = [-\log(1-\frac{m}{(n-1)v(u-1)})]^{\frac{1}{2}}$, 则当 $n \to \infty$ 时, 有

$$\frac{s^2}{\sigma^2}[-\log(1-\frac{m}{(n-1)v(u-1)})] = \begin{cases} \frac{s^2}{2}\cdot\frac{-\log(1-y_1)}{2[v(u-1)]^2\log(1-y_1)+2(v-1)^2\log(1-y_2)-2\log(y_3)} & y\in(0,1), \\ \frac{s^2}{2} & y=1, \end{cases}$$

这即表明, 当 $n \to \infty$ 时, 有 $t = O(A^{-1}_{(n-1)v(u-1)})$.

利用引理 4.1.1 和给定的条件 $uvm/n \to y \in (0,1]$, 可得当 $n \to \infty$ 时, 有

$$\begin{aligned} \log\frac{\Gamma_m(\frac{(n-1)v(u-1)}{2})}{\Gamma_m(\frac{(n-1)v(u-1)}{2}+v(u-1)h)} &= -mv(u-1)h(\log(n-1)v(u-1)-1-\log 2) \\ &+ A^2_{(n-1)v(u-1)}[-(v(u-1)h)^2+(m-(n-1)v(u-1)+\frac{1}{2})v(u-1)h]+o(1) \\ &= -mv(u-1)h\log(n-1)-mv(u-1)h\log v(u-1) \\ &+ mv(u-1)h(1+\log 2)+(v(u-1)h)^2\log(1-\frac{m}{(n-1)v(u-1)}) \\ &- (m-(n-1)v(u-1)+\frac{1}{2})v(u-1)h)\log(1-\frac{m}{(n-1)v(u-1)})+o(1). \end{aligned} \tag{4-3}$$

同样地, 取定 $B_{(n-1)(v-1)} = [-\log(1-\frac{m}{(n-1)(v-1)})]^{\frac{1}{2}}$ 和 $C_{n-(uv-k)m-1} = [-\log(1-\frac{m}{n-(uv-k)m-1})]^{\frac{1}{2}}$, 当 $n \to \infty$ 时, 也有 $t = O(B^{-1}_{(n-1)(v-1)})$ 和 $t = O(C^{-1}_{n-(uv-k)m-1})$. 那么, 再次利用引理 4.1.1, 当 $n \to \infty$ 时, 可得

$$\begin{aligned} \log\frac{\Gamma_m(\frac{(n-1)(v-1)}{2})}{\Gamma_m(\frac{(n-1)(v-1)}{2}+(v-1)h)} &= -m(v-1)h\log(v-1)-m(v-1)h\log(v-1) \\ &+ m(v-1)h(1+\log 2)+((v-1)h)^2\log(1-\frac{m}{(n-1)(v-1)}) \\ &- (m-(n-1)(v-1)+\frac{1}{2})(v-1)h\log(1-\frac{m}{(n-1)(v-1)})+o(1), \end{aligned} \tag{4-4}$$

$$\sum_{k=1}^{uv-1}\log\frac{\Gamma_m(\frac{n-(uv-k)-1}{2}+h)}{\Gamma_m(\frac{n-(uv-k)}{2})}=\sum_{k=1}^{uv-1}m\log[n-(uv-k)m-1]-m(uv-1)(1+\log 2)$$
$$-\sum_{k=1}^{uv-1}h^2\log(1-\frac{m}{n-(uv-k)m-1})$$
$$+\sum_{k=1}^{uv-1}\log(1-\frac{m}{n-(uv-k)m-1})(m-(n-(uv-k)m-1)+\frac{1}{2})h+o(1). \tag{4-5}$$

因此, 将 (4-3)–(4-5) 式代入 (4-2) 式, 可得当 $n\to\infty$ 时, 有

$$\begin{aligned}\log Ee^{\frac{2}{n}h\log\Lambda}&=\frac{1}{2}h^2\left\{2[v(u-1)]^2\log(1-\frac{m}{(n-1)v(u-1)})+2(v-1)^2\log(1-\frac{m}{(n-1)(v-1)})\right.\\&\left.-2\log(\frac{n-uvm-1}{n-m-1})\right\}+h\left\{-(m-(n-1)v(u-1)+\frac{1}{2})v(u-1)\right.\\&\times\log(1-\frac{m}{(n-1)v(u-1)})-(m-(n-1)(v-1)+\frac{1}{2})(v-1)\\&\times\log(1-\frac{m}{(n-1)(v-1)})+(-n+uvm+\frac{3}{2})\log(\frac{n-uvm-1}{n-m-1})\\&\left.+m(uv-1)\log(\frac{n-m-1}{n-1})\right\}+o(1)\\&=\mu_n h+\sigma_n^2\frac{h^2}{2}+o(1).\end{aligned}$$

其中

$$\begin{aligned}\mu_n&=-(m-(n-1)v(u-1)+\frac{1}{2})v(u-1)\log(1-\frac{m}{(n-1)v(u-1)})\\&-(m-(n-1)(v-1)+\frac{1}{2})(v-1)\log(1-\frac{m}{(n-1)(v-1)})\\&+(-n+uvm+\frac{3}{2})\log(\frac{n-uvm-1}{n-m-1})+m(uv-1)\log(\frac{n-m-1}{n-1}),\\\sigma_n&=2[v(u-1)]^2\log(1-\frac{m}{(n-1)v(u-1)})\\&+2(v-1)^2\log(1-\frac{m}{(n-1)(v-1)})-2\log(\frac{n-uvm-1}{n-m-1}).\end{aligned}$$

注意到 $h=h_n=s/\sigma_n$, 可得当 $n\to\infty$ 时, 有

$$\log Ee^{\frac{2}{n}h\log\Lambda}=\frac{s^2}{2}+\frac{\mu_n}{\sigma_n}s+o(1), \tag{4-6}$$

这即是说 (4-6) 式对所有的 $|s|<\frac{\delta_0}{2}$ 成立.

证毕. ■

定理 2.3.3 的证明:

由子序列定理可知, 当且仅当对任意的子列 X_{n_j} 存在进一步地子列 $X_{n_{j_k}}$ 使得 $X_{n_{j_k}} \xrightarrow{d} X$. 这即是说, 欲证明在 $n \to \infty$ 时,

$$H_n := \frac{\log \Lambda - \mu_n}{\sigma_n} \xrightarrow{d} \boldsymbol{N}(0,1), \tag{4-7}$$

等价于, 对于任意子列 n_k,存在一个子列 n_{k_j}, 使得当 $n \to \infty$ 时, $H_{n_{k_j}} \xrightarrow{d} \boldsymbol{N}(0,1)$. 由假定条件可知, 对于任意 n, 有 $uvm/n \in [0,1]$, 即对于任意一个子列 n_k, 存在一个子列 n_{k_j}, 使得 $uvm_{n_{k_j}}/n_{n_{k_j}} \in [0,1]$. 在不失一般性下, 我们仅需要在满足 $\lim_{n\to\infty} uvm_n/n \in [0,1]$ 时, 使得 (4-7) 式成立.

同理于定理 2.3.2, 根据 Resnick (2019) [34] 中定理 9.5.2 给出的 Levy 连续性定理可知, 则只需证明当 $n \to \infty$ 时, 有

$$E \exp\left(\frac{\log \Lambda - \mu_n}{\sigma_n} s\right) \to e^{\frac{s^2}{2}}, \tag{4-8}$$

对所有满足 $|s| \le 1$ 的 s 都成立, 或等价地,

$$\log E(\frac{2}{n}\Lambda)^t = \mu_n t + \frac{\sigma_n^2}{2} t^2 + o(1),$$

对所有满足 $|s| \le 1$ 的 t 都成立, 其中 $t = t_n = \frac{s}{\sigma_n}$.

由 (2-6) 式可知:

$$\begin{aligned} E \exp[(\frac{2}{n} t \log \Lambda)] = {} & [v(u-1)]^{mv(u-1)h} (v-1)^{m(v-1)h} \prod_{k=1}^{uv-1} \frac{\Gamma_m(\frac{n-(uv-k)-1}{2} + h)}{\Gamma_m(\frac{n-(uv-k)}{2})} \\ & \times \frac{\Gamma_m(\frac{(n-1)v(u-1)}{2})}{\Gamma_m(\frac{(n-1)v(u-1)}{2} + v(u-1)h)} \frac{\Gamma_m(\frac{(n-1)(v-1)}{2})}{\Gamma_m(\frac{(n-1)(v-1)}{2} + (v-1)h)}. \end{aligned}$$

由给定的假设 $\frac{m}{n} \to 0$, 易知 $\sigma_n^2 \sim \frac{m^2}{n^2}$, 这即表明 $\frac{tm}{n}$ 有界. 进一步地, 在 $n \to \infty$ 时, 可得 $t = O(\frac{n}{m})$. 由此, 利用引理 4.1.2 可知, 当 $n \to \infty$ 时, 有

$$\begin{aligned} & \log \frac{\Gamma_m(\frac{(n-1)v(u-1)}{2})}{\Gamma_m(\frac{(n-1)v(u-1)}{2} + v(u-1)h)} \\ & = [2m + ((n-1)v(u-1) - m - \frac{1}{2}) \log(1 - \frac{m}{(n-1)v(u-1)})] v(u-1)h \\ & + [\frac{m}{(n-1)v(u-1)} + \log(1 - \frac{m}{(n-1)v(u-1)})][v(u-1)h]^2 \\ & - m[(\frac{(n-1)v(u-1)}{2} + v(u-1)h) \log(\frac{(n-1)v(u-1)}{2} + v(u-1)h) \\ & - \frac{(n-1)v(u-1)}{2} \log(\frac{(n-1)v(u-1)}{2})] + o(1). \end{aligned} \tag{4-9}$$

由泰勒展开可知, 当 $f(x)=x\log x$ 在 $x=x_0$ 处展开时, 有 $f(x)=x_0\log_0+(1+\log(x_0))(x-x_0)+\frac{1}{2x_0}(x-x_0)^2$, 所以针对上式的第三项, 可以将其改写为:

$$\begin{aligned}(\frac{(n-1)v(u-1)}{2}&+v(u-1)h)\log(\frac{(n-1)v(u-1)}{2}+v(u-1)h)\\&=\frac{(n-1)v(u-1)}{2}\log(\frac{(n-1)v(u-1)}{2})\\&+v(u-1)h(1+\log(\frac{(n-1)v(u-1)}{2}))+O(\frac{1}{(n-1)v(u-1)}).\end{aligned}\tag{4-10}$$

结合 (4-9) 式和 (4-10) 式, 有

$$\begin{aligned}&\log\frac{\Gamma_m(\frac{(n-1)v(u-1)}{2})}{\Gamma_m(\frac{(n-1)v(u-1)}{2}+v(u-1)h)}\\&=[2m+((n-1)v(u-1)-m-\frac{1}{2})\log(1-\frac{m}{(n-1)v(u-1)})]v(u-1)h\\&+[v(u-1)h]^2\log(1-\frac{m}{(n-1)v(u-1)})-mv(u-1)h(1+\log(\frac{(n-1)v(u-1)}{2}))\\&+\frac{mv(u-1)}{n-1}h^2+o(1),\end{aligned}\tag{4-11}$$

同理于上, 我们有:

$$\begin{aligned}&\log\frac{\Gamma_m(\frac{(n-1)(v-1)}{2})}{\Gamma_m(\frac{(n-1)(v-1)}{2}+(v-1)h)}\\&=[2m+((n-1)(v-1)-m-\frac{1}{2})\log(1-\frac{m}{(n-1)(v-1)})](v-1)h\\&+[(v-1)h]^2\log(1-\frac{m}{(n-1)(v-1)})]-m(v-1)h(1+\log(\frac{(n-1)(v-1)}{2}))\\&+\frac{m(v-1)}{n-1}h^2+o(1),\end{aligned}\tag{4-12}$$

$$\begin{aligned}&\sum_{k=1}^{uv-1}\log\frac{\Gamma_m(\frac{n-(uv-k)-1}{2}+h)}{\Gamma_m(\frac{n-(uv-k)}{2})}\\&=-\sum_{k=1}^{uv-1}[2m+(n-(uv-k)m-m-\frac{3}{2})\log(1-\frac{m}{n-(uv-k)m-1})]h\\&-\sum_{k=1}^{uv-1}h^2\log(1-\frac{m}{n-(uv-k)m-1})]+\sum_{k=1}^{uv-1}mh(1+\log(\frac{n-(uv-k)m-1}{2}))\\&-\sum_{k=1}^{uv-1}\frac{m}{n-(uv-k)m-1}h^2+o(1),\end{aligned}\tag{4-13}$$

结合 (4-11)–(4-13) 式, 且

$$\frac{mv(u-1)}{n-1}+\frac{m(v-1)}{n-1}-\sum_{k=1}^{uv-1}\frac{m}{n-(uv-k)m-1}=\sum_{k=1}^{uv-1}\frac{-(uv-k)m}{(n-1)(n-(uv-k)m-1)}=o(1).$$

即可得到

$$\log Ee^{\frac{2}{n}h\log\Lambda}=\frac{s^2}{2}+\frac{\mu_n}{\sigma_n}s+o(1),$$

其中在 $n\to\infty$ 时, μ_n 和 σ_n 如 (2-8) 式和 (2-9) 式所述, 可得到 (4-8) 式对所有满足 $|s|\le 1$ 的 s 都成立.

证毕. ■

§4.3 证明定理 2.3.5

为了证明该定理, 我们使用引理 4.1.3 并证明满足该结果的条件(A.1)–(A.5). 由 Muirhead (1982) [31], Anderson (2003) [2] 可知, 在原假设 H_0 成立时, 似然比统计量 Λ 可以转化为独立的 Beta 分布的随机变量的乘积. 因此, 由 (2-4) 式可知, Λ_a, Λ_b, Λ_c 均可以转换为服从 Beta 分布随机变量乘积的形式, 所以有:

$$\Lambda_a\overset{d}{\sim}\prod_{j=1}^{m}\prod_{k=1}^{uv-1}(X_{jk}),\qquad \text{其中}\quad X_{jk}\sim Beta(\frac{n-(uv-k)m-j}{2},\frac{(uv-k)m}{2});$$

$$\Lambda_b\overset{d}{\sim}\prod_{j=1}^{m}\prod_{k=1}^{v(u-1)}(X_{jk}^{*}),\quad \text{其中}\quad X_{jk}^{*}\sim Beta(\frac{n-j}{2},\frac{j-1}{2}+\frac{2k-j-1}{2v(u-1)});$$

$$\Lambda_c\overset{d}{\sim}\prod_{j=1}^{m}\prod_{k=1}^{v-1}(X_{jk}^{**}),\qquad \text{其中}\quad X_{jk}^{**}\sim Beta(\frac{n-j}{2},\frac{j-1}{2}+\frac{2k-j-1}{2(v-1)}).$$

因为对于任意一个随机变量 $X\sim Beta(a,b)$, 我们有:

$$E[\log(X)]=\psi_0(a)-\psi_0(a+b),$$
$$Var[\log(X)]=\psi_1(a)-\psi_1(a+b).$$

其中 $\psi_k(x)=(\frac{d}{dx})^{k+1}\log\Gamma(x)$ $(k\ge 0)$.

令 $S_n := \log\Lambda = \log\Lambda_a\Lambda_b\Lambda_c$, 所以有

$$\begin{aligned}
E(S_n) &= E(\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\log\Lambda_a + \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\log\Lambda_b + \sum_{j=1}^{m}\sum_{k=1}^{v-1}\log\Lambda_c)\\
&= \sum_{j=1}^{m}\sum_{k=1}^{uv-1}\left[\psi_0(\frac{n-(uv-k)m-j}{2}) - \psi_0(\frac{n-j}{2})\right]\\
&+ \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\left[\psi_0(\frac{n-j}{2}) - \psi_0(\frac{n-j}{2}+\frac{2k-j-1}{2v(u-1)})\right]\\
&+ \sum_{j=1}^{m}\sum_{k=1}^{v-1}\left[\psi_0(\frac{n-j}{2}) - \psi_0(\frac{n-j}{2}+\frac{2k-j-1}{2(v-1)})\right]\\
&= A_1 - A_2 - A_3,
\end{aligned}$$

其中

$$\begin{aligned}
A_1 &= \sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi_0(\frac{n-(uv-k)m-j}{2});\\
A_2 &= \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi_0(\frac{n-j}{2}+\frac{2k-j-1}{2v(u-1)});\\
A_3 &= \sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi_0(\frac{n-j}{2}+\frac{2k-j-1}{2(v-1)}).
\end{aligned}$$

因为 $\psi_0(z) = \log(z) - \frac{1}{z} + O(z^{-2})$, $\sum_{k=1}^{n}\frac{1}{k} = \log n + \mathbb{C} + \frac{1}{2n} + O(n^{-2})$, $\mathbb{C}$为欧拉常数.

所以

$$\begin{aligned}
A_1 &= \sum_{i=n-uvm}^{n-m-1}\psi_0(\frac{i}{2}) = \sum_{i=n-uvm}^{n-m-1}\left\{\log\frac{i}{2} - \frac{1}{i} + O(i^{-2})\right\}\\
&= \log\frac{(n-m-1)!}{(n-uvm-1)!} - (uvm-m)\log 2 - \log\frac{(n-m-1)}{(n-uvm-1)} + o(1),
\end{aligned}$$

同理可得

$$\begin{aligned}
A_2 &= \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\{\log[(n-1)v(u-1)+2k-j-1] - \log 2v(u-1)\\
&\quad - \frac{v(u-1)}{(n-1)v(u-1)+2k-j-1} + O(n^{-2})\Big\}\\
&= -mv(u-1)\log[2v(u-1)] + \sum_{k=1}^{v(u-1)}\log\frac{[(n-1)v(u-1)+2k-2]!}{[(n-1)v(u-1)+2k-2-m]!} + o(1),
\end{aligned}$$

$$A_3 = -m(v-1)\log[2(v-1)]\sum_{k=1}^{v-1}\log\frac{[(n-1)(v-1)+2k-2]!}{[(n-1)(v-1)+2k-2-m]!} + o(1).$$

由 Stirling 公式: $\log(n!) = n\log n - n + \frac{1}{2}\log(2\pi n) + \frac{1}{12n} + O(n^{-3})$, 可知:

$$\begin{aligned} A_1 &= (n-m-\frac{3}{2})\log(n-m-1) - (n-m-\frac{3}{2})\log(n-uvm-1) \\ &\quad + (m-uvm) + \frac{m-uvm}{12(n-m-1)(n-uvm-1)} + (m-uvm)\log 2 + o(1), \end{aligned} \tag{4-14}$$

$$\begin{aligned} A_2 &= v(u-1)[(n-1)v(u-1) - \frac{1}{2}]\log[(n-1)v(u-1)] - mv(u-1) \\ &\quad - v^2(u-1)^2\log[1 - \frac{m}{(n-1)v(u-1)}] - mv(u-1)\log[2v(u-1)] \\ &\quad - v(u-1)[(n-1)v(u-1) - m - \frac{1}{2}]\log[(n-1)v(u-1) - m] + o(1), \end{aligned} \tag{4-15}$$

$$\begin{aligned} A_3 &= (v-1)[(n-1)(v-1) - \frac{1}{2}]\log[(n-1)(v-1)] - m(v-1) \\ &\quad - (v-1)^2\log[1 - \frac{m}{(n-1)(v-1)}] - m(v-1)\log[2(v-1)] \\ &\quad - (v-1)[(n-1)(v-1) - m - \frac{1}{2}]\log[(n-1)(v-1) - m] + o(1). \end{aligned} \tag{4-16}$$

结合 (4-14)–(4-16) 式可得,

$$\begin{aligned} E(S_n) &= -(m-(n-1)v(u-1) + \frac{1}{2})v(u-1)\log(1 - \frac{m}{(n-1)v(u-1)}) \\ &\quad - (m-(n-1)(v-1) + \frac{1}{2})(v-1)\log(1 - \frac{m}{(n-1)(v-1)}) \\ &\quad + (-n + uvm + \frac{3}{2})\log(\frac{n-uvm-1}{n-m-1}) + m(uv-1)\log(\frac{n-m-1}{n-1}) \\ &\quad + v^2(u-1)^2\log[1 - \frac{m}{(n-1)v(u-1)}] + (v-1)^2\log[1 - \frac{m}{(n-1)(v-1)}] + o(1). \end{aligned}$$

同理可求统计量的方差

$$\begin{aligned} Var(S_n) &= Var(\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\log\Lambda_a + \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\log\Lambda_b + \sum_{j=1}^{m}\sum_{k=1}^{v-1}\log\Lambda_c) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{uv-1}\left[\psi_1(\frac{n-(uv-k)m-j}{2}) - \psi_1(\frac{n-j}{2})\right] \\ &\quad + \sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\left[\psi_1(\frac{n-j}{2}) - \psi_1(\frac{n-j}{2} + \frac{2k-j-1}{2v(u-1)})\right] \\ &\quad + \sum_{j=1}^{m}\sum_{k=1}^{v-1}\left[\psi_1(\frac{n-j}{2}) - \psi_1(\frac{n-j}{2} + \frac{2k-j-1}{2(v-1)})\right] \\ &= B_1 - B_2 - B_3. \end{aligned}$$

其中

$$B_1=\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi_1(\frac{n-(uv-k)m-j}{2});$$
$$B_2=\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi_1(\frac{n-j}{2}+\frac{2k-j-1}{2v(u-1)});$$
$$B_3=\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi_1(\frac{n-j}{2}+\frac{2k-j-1}{2(v-1)}).$$

因为 $\psi_1(z)=\frac{1}{z}+O(z^{-2})$, $\sum_{k=1}^{n}\frac{1}{k}=\log n+\mathbb{C}+\frac{1}{2n}+O(n^{-2})$, $\mathbb{C}$为欧拉常数. 所以

$$\begin{aligned}B_1&=\sum_{i=n-uvm}^{n-m-1}\psi_1(\frac{i}{2})=\sum_{i=n-uvm}^{n-m-1}\frac{2}{i}+o(1)=2\left\{\sum_{i=1}^{n-m-1}\frac{1}{i}-\sum_{i=1}^{n-uvm-1}\frac{1}{i}\right\}+o(1)\\&=2\log\left(\frac{n-m-1}{n-uvm-1}\right)+O(n^{-1})=2\log\left(\frac{1-m/n-1/n}{1-uvm/n-1/n}\right)+o(1),\end{aligned}$$

同理可得,

$$\begin{aligned}B_2&=\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\frac{2v(u-1)}{(n-1)v(u-1)+2k-j-1}+O\left((\frac{2v(u-1)}{(n-1)v(u-1)+2k-j-1})^{-2}\right)\\&=2v(u-1)\sum_{k=1}^{m}\left\{\log[(n-1)v(u-1)+2k-2]+\frac{1}{2[(n-1)v(u-1)+2k-2]}\right.\\&\quad\left.-\log[(n-1)v(u-1)+2k-m-2]-\frac{1}{2[(n-1)v(u-1)+2k-m-2]}\right\}+o(1)\\&=2v^2(u-1)^2\log(\frac{(n-1)v(u-1)}{(n-1)v(u-1)-m})+o(1),\end{aligned}$$

$$B_3=2(v-1)^2\log(\frac{(n-1)(v-1)}{(n-1)(v-1)-m})+o(1).$$

故有

$$\begin{aligned}Var(S_n)&=2[v(u-1)]^2\log(1-\frac{m}{(n-1)v(u-1)})\\&\quad+2(v-1)^2\log(1-\frac{m}{(n-1)(v-1)})-2\log(\frac{n-uvm-1}{n-m-1})+o(1).\end{aligned}\tag{4-17}$$

综上所述, 且利用引理 4.1.3 可知:

(1) (A.1), (A.2) 是显然成立的.

(2) 利用 Dette & Tomecki (2019) [15] 中引理 A.7 和定理 A.8 可知, 这里借用 Lindeberg

中心极限定理中的 Lindeberg 条件可知, 对于任意的 $\xi > 0$,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{Var(\log\Lambda)}\sum_{i=a}^{c}E\left[(\log\Lambda_i)^2 1_{\left(|\log\Lambda_i|\geq\xi\sqrt{Var(\log\Lambda)}\right)}\right]\\
&\leq\lim_{n\to\infty}\frac{1}{Var(\log\Lambda)}\sum_{i=a}^{c}E\left[(\log\Lambda_i)^2\cdot\frac{|\log\Lambda_i|^2}{\xi^2 Var(\log\Lambda)}\right]\\
&=\lim_{n\to\infty}\frac{1}{\xi^2}\frac{1}{(Var(\log\Lambda))^2}\sum_{i=a}^{c}E\left[(\log\Lambda_i)^4\right]=0,
\end{aligned}$$

所以

$$\frac{\sum_{i=a}^{c}E(\log\Lambda_i)^4}{(Var(\log\Lambda))^2}<\frac{\sum_{i=a}^{c}E(\log\Lambda_i)^2}{Var(\log\Lambda)}.$$

所以存在一些常数 C, 使得 $E(\log\Lambda)^4 < CE(\log\Lambda)^2$ 不等式成立, 即 (A.3) 成立.

(3) 又因为 ψ_1 为减函数且值恒大于零, 所以

$$\begin{aligned}
\sup_{j,k}Var(\log X_{j,k})&=\sup_{j,k}\left\{\psi_1(\frac{n-(uv-k)m-j}{2})-\psi_1(\frac{n-j}{2})\right\}\\
&\leq\psi_1(\frac{n-uvm}{2})=\frac{2}{n-uvm}+O(\frac{1}{(n-uvm)^2})\\
&=o(1),
\end{aligned}$$

同理可得 $\sup_{j,k}Var(\log X^*_{j,k})=\sup_{j,k}Var(\log X^{**}_{j,k})=o(1)$(注意 $uvm/n<1$). 所以对于任意 $n\to\infty$ 有:

$$\sup_{j,k}g_n^2(j,k)Var(\log X_{j,k})=\sup_{j,k}g_n^2(j,k)Var(\log X^*_{j,k})=\sup_{j,k}g_n^2(j,k)Var(\log X^{**}_{j,k})=0,$$

其中 $g_n^2(j,k)=1$, 即证明了 (A.4) 成立.

(4) 由 (4-17) 式可知, 由于统计量的方差是存在, 所以存在一个常数 $\sigma^2=g_n^2(i)[Var(X_{j,k})+Var(\log X^*_{j,k})+Var(\log X^{**}_{j,k})]$ 使得 (A.5) 成立, 其中 $g_n^2(j,k)=1$.

综上, 条件 (A.1)–(A.5) 成立, 所以由引理 4.1.3 可知, 定理 2.3.5 成立.

证毕. ■

§4.4　证明定理 2.3.7 和定理 2.3.8

定理 2.3.7 的证明:

令 $T=-\frac{2}{n}\log\Lambda$, 对于所有的 $t\in\mathbb{R}$, 由 (2-6) 式可知, 统计量 T 的特征函数为

$$\begin{aligned}\varphi_T(t)&=E[\exp(itT)]=E(\Lambda^{-\frac{2}{n}it})\\&=\prod_{j=1}^{m}\prod_{k=1}^{uv-1}\frac{\Gamma\left(\frac{n-(uv-k)m-j}{2}-it\right)}{\Gamma\left(\frac{n-(uv-k)m-j}{2}\right)}\\&\times\prod_{j=1}^{m}\prod_{k=1}^{v(u-1)}\frac{\Gamma\left(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)}\right)}{\Gamma\left(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)}-it\right)}\\&\times\prod_{j=1}^{m}\prod_{k=1}^{v-1}\frac{\Gamma\left(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}\right)}{\Gamma\left(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}-it\right)}.\end{aligned}\tag{4-18}$$

对上式 (4.18) 两边同时取对数, 可得

$$\begin{aligned}\log\varphi_T(t)&=\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\left\{\log\Gamma(\frac{n-(uv-k)m-j}{2}-it)-\log\Gamma(\frac{n-(uv-k)m-j}{2})\right\}\\&+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\left\{\log\Gamma(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1}-\log\Gamma(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1}-it)\right\}\\&+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\left\{\log\Gamma(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}-\log\Gamma(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}-it)\right\}.\end{aligned}\tag{4-19}$$

利用泰勒展开公式, 可得对任意实数 a 和 b, 有

$$\log\Gamma(a+b)=\log\Gamma(a)+\sum_{s=1}^{\infty}\frac{b^s}{s!}\psi^{(s-1)}(a),\tag{4-20}$$

其中 $\psi^s(a)$ 是对数伽马函数, 其定义式如下:

$$\psi^s(a)=(\frac{d}{dx})^{s+1}\log\Gamma(a)=\begin{cases}\mathbb{C}+\sum_{k=0}^{\infty}(\frac{1}{1+k}-\frac{1}{k+a}) & s=0,\\ \sum_{k=0}^{\infty}\frac{(-1)^{s+1}s!}{(k+a)^{s+1}} & s=1,2,\cdots,\end{cases}\tag{4-21}$$

其中 $\mathbb{C}$ 为欧拉常数.

则 $\log\varphi_T(t)$ 可以变形为:

$$\begin{aligned}\log\varphi_T(t) &= \sum_{j=1}^{m}\sum_{k=1}^{uv-1}\sum_{s=1}^{\infty}\frac{1}{s!}(-it)^s\psi^{s-1}(\frac{n-(uv-k)m-j}{2})\\ &-\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\sum_{s=1}^{\infty}\frac{1}{s!}(-it)^s\psi^{s-1}(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1}\\ &-\sum_{j=1}^{m}\sum_{k=1}^{v-1}\sum_{s=1}^{\infty}\frac{1}{s!}(-it)^s\psi^{s-1}(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1})\\ &= it\hat{\mu}_n+\frac{(it)^2}{2}\hat{\sigma}_n^2+\sum_{s=3}^{\infty}\frac{1}{s!}(it)^s\kappa^{(s)},\end{aligned} \tag{4-22}$$

其中

$$\begin{aligned}\hat{\mu}_n &= -\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi(\frac{n-(uv-k)m-j}{2})+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)})\\ &+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi(\frac{n-1}{2}-\frac{j-1}{2(v-1)}),\end{aligned} \tag{4-23}$$

$$\begin{aligned}\hat{\sigma}_n^2 &= \sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi^{(1)}(\frac{n-(uv-k)m-j}{2})+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi^{(1)}(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)})\\ &+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi^{(1)}(\frac{n-1}{2}-\frac{j-1}{2(v-1)}),\end{aligned} \tag{4-24}$$

以及对任意的 $s\geq 3$, 有

$$\begin{aligned}\kappa^{(s)} = (-1)^s\Bigg\{&\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\psi^{(s-1)}(\frac{n-(uv-k)m-j}{2})\\ &+\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\psi^{(s-1)}(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)})\\ &+\sum_{j=1}^{m}\sum_{k=1}^{v-1}\psi^{(s-1)}(\frac{n-1}{2}-\frac{j-1}{2(v-1)})\Bigg\}.\end{aligned} \tag{4-25}$$

令 $Z=(T-\hat{\mu}_n)/\hat{\sigma}_n$, 可知 Z 的特征函数可以表达为

$$\begin{aligned}\varphi_Z(t)&=E\exp(itZ)=\exp(-\frac{\mu_n}{\sigma_n}it)\varphi_T(\frac{t}{\sigma_n})\\&=\exp\left\{\frac{it\mu_n}{\sigma_n}+\frac{(it)^2}{2}+\sum_{s=3}^{\infty}\frac{1}{s!}(it)^s\frac{\kappa^{(s)}}{\sigma_n^s}-\frac{it\mu_n}{\sigma_n}\right\}\\&=e^{-\frac{t^2}{2}}\left\{1+\sum_{a=1}^{\infty}\frac{1}{a!}\left[\sum_{s=3}^{\infty}\frac{(it)^s\kappa^{(s)}}{s!\sigma_n^s}\right]^a\right\}\\&=e^{-\frac{t^2}{2}}\left\{1+\sum_{a=1}^{\infty}\frac{(it)^{3a}}{a!}\left[\sum_{s=0}^{\infty}\frac{(it)^s\kappa^{(s+3)}}{(s+3)!\sigma_n^{s+3}}\right]^a\right\}.\end{aligned}$$

因为

$$\begin{aligned}\left[\sum_{s=0}^{\infty}\frac{(it)^s\kappa^{(s+3)}}{(s+3)!\sigma_n^{s+3}}\right]^a&=\sum_{s_1=1}^{\infty}\cdots\sum_{s_a=1}^{\infty}\prod_{l=1}^{a}\frac{(it)^{s_l}}{(s_l+3)!}\frac{\kappa^{(s_l+3)}}{\sigma_n^{(s_l+3)!}}\\&=\sum_{s=0}^{\infty}\sum_{s_1+\cdots s_a=s}\prod_{l=1}^{a}\frac{(it)^{s_l}}{(s_l+3)!}\frac{\kappa^{(s_l+3)}}{\sigma_n^{s_l+3}}.\end{aligned}$$

所以

$$\begin{aligned}\varphi_Z(t)&=e^{-\frac{t^2}{2}}\left\{1+\sum_{a=1}^{\infty}\frac{(it)^{3a}}{a!}\sum_{s=0}^{\infty}\sum_{s_1+\cdots s_a=s}\prod_{l=1}^{a}\frac{(it)^{s_l}}{(s_l+3)!}\frac{\kappa^{(s_l+3)}}{\sigma_n^{s_l+3}}\right\}\\&=\exp(-\frac{t^2}{2})\left\{1+\sum_{a=1}^{\infty}\sum_{s=0}^{\infty}(it)^{3a+s}\gamma_{a,s}\right\},\end{aligned}\tag{4-26}$$

其中

$$\gamma_{a,s}=\sum_{s_1+\cdots+s_a=s}\prod_{l=1}^{a}\frac{\kappa^{(s_l+3)}}{(s_l+3)!\sigma_n^{s_l+3}}.\tag{4-27}$$

从而, 利用特征函数的逆变换公式, 可得 Z 的分布函数为

$$P(Z\le x)=\Phi(x)+\sum_{a=1}^{\infty}R_a(x),\tag{4-28}$$

其中

$$\begin{aligned}&\Phi(x)\text{ 为标准正态分布的分布函数;}\\&R_a(x)\text{ 满足 }\int_{-\infty}^{\infty}e^{itx}dR_a(x)=\frac{1}{a!}\sum_{s=0}^{\infty}(it)^{3a+s}e^{-\frac{t^2}{2}}\gamma_{a,s}.\end{aligned}\tag{4-29}$$

因此下一步需要对 $R_a(x)$ 进行计算简化, 由分部积分法可知:

$$\begin{aligned}\exp(-\frac{t^2}{2}) &= \int_{-\infty}^{\infty} e^{itx} d\Phi(x) = \int_{-\infty}^{\infty} \Phi^{(1)}(x) d(\frac{e^{itx}}{it}) \\ &= (-it)^{-1} \int_{-\infty}^{\infty} e^{itx} d\Phi^{(1)}(x) \\ &\vdots \\ &= (-it)^{-(3a+s)} \int_{-\infty}^{\infty} e^{itx} d\Phi^{(3a+s)}(x),\end{aligned} \tag{4-30}$$

其中 $\Phi^{(k)}(x) = (\frac{d}{dx})^k \Phi(x)$. 可得

$$(it)^{3a+s} e^{-\frac{t^2}{2}} = \int_{-\infty}^{\infty} e^{itx} d[(-\frac{d}{dx})^{3a+s} \Phi(x)],$$

以及

$$\int_{-\infty}^{\infty} e^{itx} d[\frac{1}{a!} \sum_{s=0}^{\infty} \gamma_{a,s} (-\frac{d}{dx})^{3a+s} \Phi(x)] = \frac{1}{a!} \sum_{s=0}^{\infty} (it)^{3a+s} e^{-\frac{t^2}{2}} \gamma_{a,s}. \tag{4-31}$$

通过 (4-29) 式和 (4-31) 式可知

$$R_a(x) = \frac{1}{a!} \sum_{a=0}^{\infty} \gamma_{a,s} (-\frac{d}{dx})^{3a+s} \Phi(x).$$

又因为

$$(-\frac{d}{dx})^s \Phi(x) = -h_{s-1}(x) \phi(x),$$

对于所有的 $s \geq 1$ 都成立, 其中 $h_s(x)$ 为 s 阶 Hermite 多项式, 且满足 $(\frac{d}{dx})^s \exp(-\frac{x^2}{2}) = (-1)^s h_s(x) \exp(-\frac{x^2}{2})$.

所以

$$R_a(x) = -\frac{1}{a!} \sum_{s=0}^{\infty} \gamma_{a,s} h_{3a+s-1}(x).$$

这意味着可改写 (4-28) 式为

$$P(Z \leq x) = \Phi(x) - \phi(x) \left[\sum_{a=1}^{\infty} \frac{1}{a!} \sum_{s=0}^{\infty} \gamma_{a,s} h_{3a+l-1}(x) \right]. \tag{4-32}$$

定义

$$\varphi_Z^{(l)}(t) = \exp(-\frac{t^2}{2}) \left\{ 1 + \sum_{a=1}^{l} \frac{1}{a!} \sum_{s=0}^{l-a} (it)^{3a+s-1} \gamma_{a,s} \right\},$$

和

$$\Phi_l(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_Z^{(l)}(t) dt.$$

则类比于 (4-28) 式和 (4-32) 式可得:

$$\Phi_l(x)=\Phi(x)-\phi(x)\left[\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}\gamma_{a,s}h_{3a+s-1}(x)\right]. \tag{4-33}$$

所以, 通过 (4-28) 式和 (4-33) 式可得

$$\begin{aligned}P(Z\le x)-\Phi_l(x)&=-\phi(x)\left\{\sum_{a=1}^{\infty}\frac{1}{a!}\sum_{s=0}^{\infty}\gamma_{a,s}h_{3a+s-1}(x)-\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}\gamma_{a,s}h_{3a+s-1}(x)\right\}\\&=-\phi(x)\left\{\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{\infty}\gamma_{a,s}h_{3a+s-1}(x)+\sum_{a=l+1}^{\infty}\frac{1}{a!}\sum_{s=0}^{\infty}\gamma_{a,s}h_{3a+s-1}(x)\right.\\&\quad\left.-\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}\gamma_{a,s}h_{3a+s-1}(x)\right\}\\&=-\phi(x)\left\{\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=l-a+1}^{\infty}\gamma_{a,s}+\sum_{a=l+1}^{\infty}\sum_{s=0}^{\infty}\gamma_{a,s}\right\}h_{3a+s-1}(x).\end{aligned} \tag{4-34}$$

接下来考虑最后一行等式右边的有界性, 因为 $\phi(x)$ 和 $h_{3a+s-1}(x)$ 都是关于 x 的有界量, 所以此时考虑 $\gamma_{a,s}$ 的有界性. 由 $\gamma_{a,s}=\sum_{s_1+\cdots+s_a=s}\prod_{l=1}^{a}\frac{\kappa^{(s_l+3)}}{(s_l+3)!\sigma_n^{s_l+3}}$ 可知, 需要首先确定 $\kappa^{(s_l+3)}$ 的有界性.

由 (4-25) 式可得, 当 $s=2$ 时,有

$$\begin{aligned}\kappa^{(2)}&<\sum_{k=1}^{uv-1}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{1}{(\frac{n-(uv-k)m-y}{2}+x)^2}dxdy\\&\quad-\sum_{k=1}^{v(u-1)}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{1}{(\frac{n-1}{2}-\frac{y-1}{2v(u-1)}+\frac{k-1}{v(u-1)}+x)^2}dxdy\\&\quad-\sum_{k=1}^{v-1}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{1}{(\frac{n-1}{2}-\frac{y-1}{2(v-1)}+\frac{k-1}{v-1}+x)^2}dxdy\\&<2\log\frac{n-m-\frac{3}{2}}{n-uvm-\frac{3}{2}},\end{aligned}$$

当 $s=3$ 时, 有

$$\begin{aligned}\kappa^{(2)} &< \int_{\frac{1}{2}}^{uv-\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{2}{(\frac{n-(uv-z)m-y}{2}+x)^3}dxdydz\\ &-\int_{\frac{1}{2}}^{v(u-1)+\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{2}{(\frac{n-1}{2}-\frac{y-1}{2v(u-1)}+\frac{z-1}{v(u-1)}+x)^3}dxdydz\\ &-\int_{\frac{1}{2}}^{v-\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{2}{(\frac{n-1}{2}-\frac{y-1}{2(v-1)}+\frac{z-1}{v-1}+x)^3}dxdydz\\ &<\frac{4}{m}\log\frac{(n-\frac{3}{2}m-\frac{3}{2})(n-(uv-\frac{1}{2})m-\frac{3}{2})}{(n-\frac{1}{2}m-\frac{3}{2})(n-(uv+\frac{1}{2})m-\frac{3}{2})}\\ &+2v^2(u-1)^2\log\frac{(n-2-\frac{m+\frac{1}{2}}{v(u-1)})(n-\frac{1}{2v(u-1)})}{(n-\frac{m+\frac{1}{2}}{v(u-1)})(n-2-\frac{1}{2v(u-1)})}\\ &+2(v-1)^2\log\frac{(n-2-\frac{m+\frac{1}{2}}{v-1})(n-\frac{1}{2(v-1)})}{(n-\frac{m+\frac{1}{2}}{v-1})(n-2-\frac{1}{2(v-1)})},\end{aligned}$$

当 $s\geq 4$ 时, 有

$$\begin{aligned}\kappa^{(s)} &< \int_{\frac{1}{2}}^{uv-\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{(s-1)!}{(\frac{n-(uv-z)m-y}{2}+x)^{s-1}}dxdydz\\ &-\int_{\frac{1}{2}}^{v(u-1)+\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{(s-1)!}{(\frac{n-1}{2}-\frac{y-1}{2v(u-1)}+\frac{z-1}{v(u-1)}+x)^{s}}dxdydz\\ &-\int_{\frac{1}{2}}^{v-\frac{1}{2}}\int_{\frac{1}{2}}^{m+\frac{1}{2}}\int_{-\frac{1}{2}}^{\infty}\frac{(s-1)!}{(\frac{n-1}{2}-\frac{y-1}{2(v-1)}+\frac{z-1}{v-1}+x)^{s}}dxdydz\\ &<\frac{2^{s-1}(s-4)!}{m}\left\{\left(\frac{1}{n-(uv+\frac{1}{2})m-\frac{3}{2}}\right)^{s-3}-\left(\frac{1}{n-\frac{3}{2}m-\frac{3}{2}}\right)^{s-3}\right.\\ &\quad\left.+\left(\frac{1}{n-\frac{1}{2}m-\frac{3}{2}}\right)^{s-3}-\left(\frac{1}{n-(uv-\frac{1}{2})m-\frac{3}{2}}\right)^{s-3}\right\}\\ &+2^{s-2}(s-4)!v^2(u-1)^2\left\{\left(\frac{1}{n-\frac{m+\frac{1}{2}}{v(u-1)}}\right)^{s-3}-\left(\frac{1}{n-\frac{m+\frac{1}{2}}{v(u-1)}-2}\right)^{s-3}\right.\\ &\quad\left.+\left(\frac{1}{n-\frac{1}{2v(u-1)}-2}\right)^{s-3}-\left(\frac{1}{n-\frac{1}{2v(u-1)}}\right)^{s-3}\right\}\\ &+2^{s-2}(s-4)!(v-1)^2\left\{\left(\frac{1}{n-\frac{m+\frac{1}{2}}{v-1}}\right)^{s-3}-\left(\frac{1}{n-\frac{m+\frac{1}{2}}{v-1}-2}\right)^{s-3}\right.\\ &\quad\left.+\left(\frac{1}{n-\frac{1}{2(v-1)}-2}\right)^{s-3}-\left(\frac{1}{n-\frac{1}{2(v-1)}}\right)^{s-3}\right\},\end{aligned}$$

综上所述可知, 令 $w,\ b_0,\ b_s\ (s=1,2,\cdots)$ 的定义如下:

$$
\begin{aligned}
w &= \frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{2},\\
b_0 &= \frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{3m}\log\frac{(n-\frac{3}{2}m-\frac{3}{2})(n-(uv-\frac{1}{2})m-\frac{3}{2})}{(n-\frac{1}{2}m-\frac{1}{2})(n-(uv+\frac{1}{2})m-\frac{3}{2})}\\
&\quad+\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{6}v^2(u-1)^2\log\frac{(n-2-\frac{m+\frac{1}{2}}{v(u-1)})(n-\frac{1}{2v(u-1)})}{(n-\frac{m+\frac{1}{2}}{2v(u-1)})(n-2-\frac{1}{2v(u-1)})}\\
&\quad+\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{6}(v-1)^2\log\frac{(n-2-\frac{m+\frac{1}{2}}{v-1})(n-\frac{1}{2(v-1)})}{(n-\frac{m+\frac{1}{2}}{v-1})(n-2-\frac{1}{2(v-1)})},\\
b_s &= \frac{2(n-(uv+\frac{1}{2})m-\frac{3}{2})}{(s+3)(s+2)(s+1)sm}\left\{1-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{3}{2}m-\frac{3}{2}}\right)^s\right.\\
&\quad\left.+\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{1}{2}m-\frac{3}{2}}\right)^s-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-(uv-\frac{1}{2})m-\frac{3}{2}}\right)^s\right\}\\
&\quad+\frac{v^2(u-1)^2}{(s+3)(s+2)(s+1)s}\left\{\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{m+\frac{1}{2}}{v(u-1)}}\right)^s-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{m+\frac{1}{2}}{v(u-1)}}\right)^s\right.\\
&\quad\left.+\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{1}{2v(u-1)}}\right)^s-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{1}{2v(u-1)}}\right)^s\right\}\\
&\quad+\frac{(v-1)^2}{(s+3)(s+2)(s+1)s}\left\{\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{m+\frac{1}{2}}{v-1}}\right)^s-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{m+\frac{1}{2}}{v-1}}\right)^s\right.\\
&\quad\left.+\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{1}{2(v-1)}}\right)^s-\left(\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{1}{2(v-1)}}\right)^s\right\},
\end{aligned}
$$

所以在 n 和 s 都足够大时, b_s 是有界的.

基于此, 在定理 2.3.7 中条件所示, 显然有 $w=\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{2}\to\infty$. 进一步地, 对于所有的 $s\geq 3$, 有

$$0<\kappa^{(s)}=s!w^{-(s-2)}b_{s-3}\ (s=3,4,\cdots). \tag{4-35}$$

所以, 根据 (4-35) 式和 (4-27) 式, 有

$$\gamma_{a,s}=O\left(\frac{1}{w^{a+s}}\right). \tag{4-36}$$

将 (4-36) 式和 (4-34) 式的后半部分放在一起, 即证明了

$$\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=l-a+1}^{\infty}\gamma_{a,s}+\sum_{a=l+1}^{\infty}\frac{1}{a!}\sum_{s=0}^{\infty}\gamma_{a,s}=O\left(\frac{1}{w^{l+1}}\right). \tag{4-37}$$

进一步地, 将 (4-37) 式代入 (4-34) 式, 即得对于任意的实数 x, 有

$$P(Z \leq x) - \Phi_l(x) = O\left(\frac{1}{w^{l+1}}\right).$$

证毕. ■

定理 2.3.8 的证明:

利用傅里叶逆变换, 可以得到 Edgeworth 展开的一致界:

$$\sup_x |P(T \leq x) - \Phi_l(x)| \leq \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{|t|}|\varphi_Z(t) - \varphi_Z^{(l)}(t)|dt \leq \frac{1}{2\pi}(I_1[r] + I_2[r] + I_3[r]), \quad (4\text{-}38)$$

其中

$$\begin{aligned} I_1[r] &= \int_{|t|\leq wr}\frac{1}{|t|}|\varphi_Z(t) - \varphi_Z^{(l)}(t)|dt; \\ I_2[r] &= \int_{|t|>wr}\frac{1}{|t|}|\varphi_Z^{(l)}(t)|dt; \\ I_3[r] &= \int_{|t|>wr}\frac{1}{|t|}|\varphi_Z(t)|dt. \end{aligned}$$

其中 $0 < r < 1$.

接下来, 本文将要对 $I_1[r]$, $I_2[r]$ 和 $I_3[r]$ 的上界分别进行估算.
首先针对 $I_1[r]$ 项. 令

$$L_1[r] = \begin{cases} \frac{11r^2-15r+6}{32r^2} + \frac{(1-r)^3}{6r^3}\log(1-r) & 0 < |r| < 1, \\ 0 & r = 0, \end{cases} \quad (4\text{-}39)$$

则 $L_1[r]$ 可以被展开为:

$$L_1[r] = \sum_{s=1}^{\infty}\frac{1}{s(s+1)(s+2)(s+3)}r^s.$$

故令 $B[r] :\triangleq \sum_{s=0}^{\infty} b_s r^s$ 可以展开为:

$$\begin{aligned}
B[r] = b_0 &+ \frac{2(n-(uv+\frac{1}{2})m-\frac{3}{2})}{m}\left\{L_1[r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{3}{2}m-\frac{3}{2}}r]\right.\\
&\left.+L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{1}{2}m-\frac{3}{2}}r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-(uv-\frac{1}{2})m-\frac{3}{2}}r]\right\}\\
&+ v^2(u-1)^2\left\{L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{m+\frac{1}{2}}{v(u-1)}}r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{m+\frac{1}{2}}{v(u-1)}}r]\right.\\
&\left.+L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{1}{2v(u-1)}}r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{1}{2v(u-1)}}r]\right\}\\
&+ (v-1)^2\left\{L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{m+\frac{1}{2}}{v-1}}r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{m+\frac{1}{2}}{v-1}}r]\right.\\
&\left.+L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-\frac{2}{2(v-1)}}r] - L_1[\frac{n-(uv+\frac{1}{2})m-\frac{3}{2}}{n-2-\frac{1}{2(v-1)}}r]\right\}.
\end{aligned}$$

令 $R_{a,s} = r^{-1}\left\{(B[r])^s - \sum_{a=1}^{s-1}(\sum_{s_1+\cdots+s_a=s})r^a\right\}$, 则对于 $|t| \le wr$ 有:

$$\begin{aligned}
\frac{1}{|t|}|\varphi_Z(t) - \varphi_Z^{(l)}(t)| &\le exp(-\frac{t^2}{2})\left\{\sum_{a=1}^{l}\frac{1}{a!}|t|^{3a-1}\sum_{s=l-a+1}^{\infty}(\sum_{s_1+\cdots+s_a=s} b_{s_1}\cdots b_{s_a})w^{-(s+a)}|t|^s\right.\\
&\left.+\sum_{a=l+1}^{\infty}\frac{1}{a!}|t|^{3a_1}(\sum_{s=0}^{\infty}b_s w^{-(s+1)}|t|^s)^a\right\}\\
&\le w^{-(l+1)}exp(-\frac{t^2}{2})\left\{\sum_{a=1}^{l}\frac{1}{a!}|t|^{l+2a}R_{a,s}[r]\right.\\
&\left.+\frac{1}{(l-1)!}|t|^{3l+2}(B[r])^{l+1}\exp(t^2 rBr)\right\},
\end{aligned}$$

对上式两侧同时求积分, 可得:

$$\begin{aligned}
I_1[r] &\le \frac{1}{w^{l+1}}\left\{\sum_{a=1}^{l}\frac{1}{a!}R_{a,l-a+1}[r](\frac{1}{2})^{-(l+2a+1)/2}\Gamma(\frac{l+2a+1}{2})\right.\\
&\left.+\frac{1}{(l+1)!}(B[r])^{l+1}(\frac{C_r}{2})^{-(3l+3)/2}\Gamma(\frac{3l+3}{2})\right\} \qquad (4\text{-}40)\\
&\triangleq U_1(r),
\end{aligned}$$

其中 $C_r = 1 - 2rB[r]$.

接着针对 $I_2[r]$ 项, 由

$$\begin{aligned} I_2[r] &= \int_{|t|>wr} \frac{1}{|t|}|E[\exp(itZ)]|dt \\ &= \int_{|t|>wr} \frac{1}{|t|}|E[\exp(-i\frac{2}{n}\tilde{t}\log\Lambda)]|dt, \end{aligned} \tag{4-41}$$

其中 $\tilde{t} = t/\sigma_n$.

因为对于任意实数 $x, y(x>0)$ 有,

$$\left|\frac{\Gamma(x+iy)}{\Gamma(x)}\right|^2 = \prod_{k=0}^{\infty}\left[1+\left(\frac{y}{x+k}\right)^2\right]^{-1}.$$

所以

$$\begin{aligned} \log|\varphi_T(t)| &= -\frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\sum_{a=0}^{\infty}\log(1+\frac{t^2}{(\frac{n-(uv-k)m-j}{2}+a)^2}) \\ &+ \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{v(u-1)}\sum_{a=0}^{\infty}\log(1+\frac{t^2}{(\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)}+a)^2}) \\ &+ \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{v-1}\sum_{a=0}^{\infty}\log(1+\frac{t^2}{(\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}+a)^2}) \\ &< -\frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{uv-1}\int_{\frac{n-(uv-k)m-j}{2}}^{\frac{n-1}{2}-\frac{j-1}{2v(u-1)}+\frac{k-1}{v(u-1)}}\log(1+\frac{t^2}{x^2})dx \\ &- \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{v-1}\int_{\frac{n-(uv-k)m-j}{2}}^{\frac{n-1}{2}-\frac{j-1}{2(v-1)}+\frac{k-1}{v-1}}\log(1+\frac{t^2}{x^2})dx \\ &< -\frac{1}{8}m[(u^2v^2-3)m+uv-3]\log(1+\frac{t^2}{n^2}). \end{aligned} \tag{4-42}$$

因此

$$\begin{aligned} I_2[r] &< \int_{wr}^{\infty}\frac{2}{t}(1+\frac{t^2}{n^2})^{-\frac{1}{8}m[(u^3v^3-3)m+uv-3]}dt \\ &< \frac{8}{m[(u^3v^3-3)m+uv-3]}(1+\alpha)^{-\frac{1}{8}m[(u^3v^3-3)m+uv-3]}\frac{1+\alpha}{\alpha} \\ &\triangleq U_2[r], \end{aligned} \tag{4-43}$$

其中 $\alpha = (\frac{wr}{n})^2$.

最后针对 $I_3[r]$ 项. 因为

$$\begin{aligned} I_3[r] &= 2\int_{wr}^{\infty}\exp(-\frac{t^2}{2})t^{-1}dt + 2\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}\gamma_{a,s}\int_{wr}^{\infty}\exp(-\frac{t^2}{2})|t|^{3a+s-1}dt \\ &= 2\int_{wr}^{\infty}\exp(-\frac{t^2}{2}(1+c-c))t^{-1}dt + 2\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}\gamma_{a,s}\int_{wr}^{\infty}\exp(-\frac{t^2}{2}(1+c-c))|t|^{3a+s-1}dt \\ &< \exp(-\frac{w^2r^2}{2}(1-c)) + \exp(-\frac{w^2r^2}{2}(1-c))\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}(\frac{c}{2})^{-(3a+s)/2}\Gamma(\frac{3a+s}{2})\gamma_{a,s}. \end{aligned}$$

其中 $0<c<1$, 故可知

$$\begin{aligned} I_3[r] &< \exp(-\frac{w^2r^2}{2}(1-c))\left\{1+\sum_{a=1}^{l}\frac{1}{a!}\sum_{s=0}^{l-a}(\frac{c}{2})^{-(3a+s)/2}\Gamma(\frac{3a+s}{2})\gamma_{a,s}\right\} \\ &\triangleq U_3[r]. \end{aligned} \tag{4-44}$$

结合 (4-38) 式, (4-40) 式, (4-43) 式和 (4-44) 式, 即得定理的结果.

证毕. ■

第五章　总结与展望

在本文中, 我们在双可交换协方差矩阵结构下考虑高维似然比统计量的渐近正态性问题. 本文以似然比统计量的矩函数为出发点, 分别聚焦高阶伽马函数的展开, 高阶贝塔分布随机变量乘积展开和高阶 Edgeworth 展开三个方法, 求解了基于原假设 H_0 下的似然比统计量 Λ 的渐近分布及其相关结果. 同时通过数值模拟和实例分析, 与传统的卡方近似方法进行了比较, 可以看出本文中所提出的三种近似方法在处理高维情况下的双可交换协方差矩阵结构的似然比统计量的假设检验问题具有很好的效果; 且在给定显著性水平 α 下的经验 Size–Power 图表中显示, 本文提出的渐近分布的效果要好于传统的卡方近似分布.

但是, 由于本文侧重于给定的高维情况下的似然比统计量的渐近正态性问题, 也限于作者的研究水平和能力, 对此类模型依然存在许多问题尚未深入研究. 下面根据本文的方法和内容, 主要列举两个需要今后进一步研究的问题.

一、在本文中, 我们考虑的是基于原假设 H_0 下的似然比统计量的渐近分布性质, 在 Chen & Jiang (2018) [8] 中给出了基于备择假设下似然比统计量的渐近分布结果, 这对本文的内容也是有很好的启示, 这是一项非常有意义的工作, 值得进一步研究.

二、在本文中, 我们考虑的是基于高维数据的似然比统计量的渐近分布性质, 即 $N > uvm$. 但是随着科学技术的发展, 超高维数据大量涌入了许多科研领域, 这类数据具有样本维数 p 远远大于样本容量 N 的特征. 在 Li & Chen (2012) [28]、Cai et al. (2013) [7] 和 Zheng et al. (2015) [56] 等文献中利用秩差中心极限来处理具有这种数据特征的似然比统计量的渐近分布问题, 这对本文的内容也是一个启发. 因此, 需要提出新的理论和方法来解决这类问题.

参考文献

[1] Akita, T., Jin, J., & Wakaki, H. (2010). High-dimensional Edgeworth expansion of a test statistic on independence and its error bound. Journal of Multivariate Analysis, 101(8), 1806-1813.

[2] Anderson, T. (2003). An introduction to multivariate statistical analysis. New York: John Wiley.

[3] Bai, Z., Jiang, D., Yao, J. F., & Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. The Annals of Statistics, 37(6B), 3822-3840.

[4] Bai, Z., & Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. Statistica Sinica, 311-329.

[5] Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. Journal of the Royal Statistical Society: Series B (Methodological), 16(2), 296-298.

[6] Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. Biometrika, 36(3/4), 317-346.

[7] Cai, T., Liu, W., & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. Journal of the American Statistical Association, 108(501), 265-277.

[8] Chen, H., & Jiang, T. (2018). A study of two high-dimensional likelihood ratio tests under alternative hypotheses. Random Matrices: Theory and Applications, 7(01), 1750016.

[9] Cardeño, L., & Nagar, D.K. (2001). Testing block sphericity of a covariance matrix. Divulgaciones Matemáticas, 9(1), 25-34.

[10] Chen, S. X., Zhang, L. X., & Zhong, P. S. (2010). Tests for high-dimensional covariance matrices. Journal of the American Statistical Association, 105(490), 810-819.

[11] Coelho, C. A., & Marques, F. J. (2013). The multi-sample block-scalar sphericity test: exact and near-exact distributions for its likelihood ratio test statistic. Communications in Statistics-Theory and Methods, 42(7), 1153-1175.

[12] Coelho, C. A., & Roy, A. (2017). Testing the hypothesis of a block compound symmetric covariance matrix for elliptically contoured distributions. Test, 26, 308-330.

[13] Coelho, C. A., & Roy, A. (2020). Testing the hypothesis of a doubly exchangeable covariance matrix. Metrika, 83(1), 45-68.

[14] Dette, H., & Dõrnemann, N. (2020). Likelihood ratio tests for many groups in high dimensions. Journal of Multivariate Analysis, 178, 104605.

[15] Dette, H., & Tomecki, D. (2019). Determinants of block Hankel matrices for random matrixvalued measures. Stochastic Processes and their Applications, 129(12), 5200-5235.

[16] Eaton, M. L. (1983). Multivariate statistics: a vector space approach. New York: John Wiley.

[17] Edgeworth, F. Y. (1896). The asymmetrical probability-curve. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 41(249), 90-99.

[18] Fujikoshi, Y., Himeno, T., & Wakaki, H. (2008). Asymptotic results in canonical discriminant analysis when the dimension is large compared to the sample size. Journal of Statistical Planning and Inference, 138(11), 3457-3466.

[19] Fujikoshi, Y., Ulyanov, V. V., & Shimizu, R. (2011). Multivariate statistics: high-dimensional and large-sample approximations. John Wiley & Sons.

[20] Gleser, L. J. (1966). A note on the sphericity test. The Annals of Mathematical Statistics, 37, 464-467.

[21] Jiang, D., Jiang, T., & Yang, F. (2012). Likelihood ratio tests for covariance matrices of highdimensional normal distributions. Journal of Statistical Planning and Inference, 142(8), 2241-2256.

[22] Jiang, T., & Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for highdimensional normal distributions. The Annals of Statistics, 41(4), 2029-2074.

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

[24] Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis. Prentice Hall: Upper Saddle River, NJ.

[25] Ledoit, O., & Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. The Annals of Statistics, 30(4), 1081-1102.

[26] Leiva, R., & Roy, A. (2011). Linear discrimination for multi-level multivariate data with separable means and jointly equicorrelated covariance structure. Journal of Statistical Planning and Inference, 141(5), 1910-1924.

[27] Leiva, R., & Roy, A. (2012). Linear discrimination for three-level multivariate data with a separable additive mean vector and a doubly exchangeable covariance structure. Computational Statistics & Data Analysis, 56(6), 1644-1661.

[28] Li, J., & Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. The Annals of Statistics, 40(2), 908-940.

[29] Marques, F. J., & Coelho, C. A. (2012). Near-exact distributions for the likelihood ratio test statistic of the multisample block-matrix sphericity test. Applied Mathematics and Computation, 219(6), 2861-2874.

[30] Moschopoulos, P. G. (1992). The hypothesis of multisample block sphericity. Sankhyā: The Indian Journal of Statistics, 54, 260-270.

[31] Muirhead, R. J. (2009). Aspects of multivariate statistical theory. John Wiley & Sons.

[32] Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika, 175-240.

[33] Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika, 263-294.

[34] Resnick, S. I. (2019). A probability path. Springer Science & Business Media.

[35] Roy, A., & Leiva, R. (2008). Likelihood ratio tests for triply multivariate data with structured correlation on spatial repeated measurements. Statistics & Probability Letters, 78(13), 1971-1980.

[36] Roy, A., & Fonseca, M. (2012). Linear models with doubly exchangeable distributed errors. Communications in Statistics-Theory and Methods, 41(13-14), 2545-2569.

[37] Roy, A., & Leiva, R. (2011). Estimating and testing a structured covariance matrix for three-level multivariate data. Communications in Statistics-Theory and Methods, 40(11), 1945-1963.

[38] Roy, A., Leiva, R., Žežula, I., & Klein, D. (2015). Testing the equality of mean vectors for paired doubly multivariate observations in blocked compound symmetric covariance matrix setup. Journal of Multivariate Analysis, 137, 50-60.

[39] Rubinstein, R. Y., & Kroese, D. P. (2016). Simulation and the Monte Carlo method. John Wiley & Sons.

[40] Schott, J. R. (2001). Some tests for the equality of covariance matrices. Journal of Statistical Planning and Inference, 94(1), 25-36.

[41] Schott, J. R. (2005). Testing for complete independence in high dimensions. Biometrika, 92(4),951-956.

[42] Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Computational Statistics & Data Analysis, 51(12), 6535-6542.

[43] Shults, J., & Morrow, A. L. (2002). The use of quasi-least squares to adjust for two levels of correlation. Biometrics 58:521 – 530.

[44] Shults, J., Whitt, M. C., & Kumanyika, S. (2004). Analysis of data with multiple sources of correlation in the framework of generalized estimating equations. Statistics in Medicine, 23(20), 3209-3226.

[45] Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. Journal of the Japan Statistical Society, 35(2), 251-272.

[46] Srivastava, M. S., & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. Journal of Multivariate Analysis, 99(3), 386-402.

[47] Szatrowski, T. H. (1977). Estimation and testing for block compound symmetry and other patterned covariance matrices with linear and nonlinear structure. Stanford, CA: Stanford University Statistics Department.

[48] Szatrowski, T. H. (1982). Testing and estimation in the block compound symmetry problem. Journal of Educational Statistics, 7(1), 3-18.

[49] Tonda, T., & Fujikoshi, Y. (2004). Asymptotic expansion of the null distribution of LR statistic for multivariate linear hypothesis when the dimension is large. Communications in Statistics-Theory and Methods, 33(5), 1205-1220.

[50] Votaw, D. (1948). Testing compound symmetry in a normal multivariate distribution. The Annals of Mathematical Statistics, 19(4), 447-473.

[51] Wakaki, H. (2006). Edgeworth expansion of Wilks' lambda statistic. Journal of Multivariat Analysis, 97(9), 1958-1964.

[52] Wakaki, H. (2007). Error bounds for high-dimensional Edgeworth expansions for some tests on covariance matrices. Hiroshima Statistical Research Group Technical Report, 07-04.

[53] Wang, Q., & Yao, J. (2013). On the sphericity test with large-dimensional observations. Electronic Journal of Statistics, 7(1), 2164-2192.

[54] Withers, C. S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. The Annals of Statistics, 11(2), 577-587.

[55] Withers, C. S. (1984). Asymptotic expansions for distributions and quantiles with power series cumulants. Journal of the Royal Statistical Society: Series B (Methodological), 46(3), 389-396.

[56] Zheng, S., Bai, Z., & Yao, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. The Annals of Statistics, 43(2), 546-591.

致　谢

时光荏苒, 岁月如梭. 三年的硕士生活已接近尾声, 三年的时光充满了酸甜苦辣, 但更多是收获与成长.

首先, 我要感谢我的硕士导师解俊山老师. 解老师为人和蔼, 学识渊博. 在三年的研究生学习过程中, 耐心细致地指导我, 给我提了许多宝贵的意见和建议, 给了我许多引导与启发. 从最开始指点我入门, 到中期引导我探寻数字的美妙, 再到最后撰写硕士毕业论文, 解老师都从各个方面给我提出了建议, 使得我在困惑之中有了新的方向与目标. 本篇论文即是在解老师的悉心指导下完成的. 在此, 再次向解老师表达我深切的谢意和最美好的祝福.

其次, 我要感谢一起进行讨论班的王玲娣老师, 同门师兄弟们. 在讨论班上, 王老师指点我学习论文, 为我拨开迷雾指点迷津, 给予了我很大的帮助. 同门的师兄弟各有所长, 用自身的经历为我提出建议, 给与我帮助, 使得我避免踏入误区之中. 同班同学及朋友之间相互帮助, 相互鼓励, 让我有了不断前行的动力.

此外, 我要感谢我的父母与朋友, 感谢他们一直以来对我的支持与理解, 是你们的默默支持使我读过了最后的一段繁忙的时光. 此外, 感谢所有关心与帮助过我的人, 祝大家幸福与美满!

最后, 衷心感谢各位评审专家在百忙之中抽出宝贵时间对本论文进行审阅、评论.

袁百城

2021 年 6 月