中国科学: 数学 2020 年 第 50 卷 第 5 期: 729～750
SCIENTIA SINICA Mathematica
论 文

基于 L^2 范数的高维数据双因素方差分析方法

献给方开泰教授 80 华诞

周布[1], 郭佳[2], 张金廷[3*]

1. 浙江工商大学统计与数学学院, 杭州 310018;
2. 浙江工业大学管理学院, 杭州 310023;
3. 华侨大学统计学院现代应用统计与大数据研究中心, 厦门 361021
E-mail: bu.zhou@u.nus.edu, jia.guo@u.nus.edu, stazjt@nus.edu.sg

收稿日期: 2019-02-24; 接受日期: 2019-04-27; 网络出版日期: 2020-05-07; * 通信作者
浙江省一流学科 A 类 (浙江工商大学统计学) 资助项目

摘要 针对观测样本数小于观测变量个数的高维双因素方差分析问题, 本文提出一种基于 L^2 范数的检验方法. 本文将一般线性假设的 Wald 类型统计量中的样本协方差阵的逆移除, 得到新的检验统计量, 使用著名的 Welch-Satterthwaite 卡方近似方法来逼近新检验统计量的零分布, 讨论这种卡方近似方法与文献中常用的正态近似方法的关系, 并证明新的基于 L^2 范数的检验方法在几类变换下具有不变性. 新检验方法的近似和渐近功效也在本文中得到研究. 模拟实验和实际数据应用表明新的方法对高维数据表现出色.

关键词 高维数据 L^2 范数 卡方型混合 Welch-Satterthwaite 卡方近似 双因素方差分析 一般线性假设检验

MSC (2010) 主题分类 62H15, 62F05

1 引言

双因素多元方差分析 (multivariate analysis of variance, MANOVA) 广泛地应用于实验科学, 如生物学、心理学和物理学等. 一些具体的双因素多元方差分析应用实例参见文献 [1–3]. 双因素多元方差分析模型主要用于在实验中比较两个因素的效应, 其中每个因素都可以有多个不同的水平. 实验中, 在每个因素的每个水平下都有多个独立重复观测. 可以把每一个给定因素的给定水平看作一个组, 相应实验条件下的独立重复观测则可看作该组的样本. 如果每个观测只包含一个指标或变量, 则问题为双因素单元方差分析, 在文献 [4] 中有详细介绍. 观测包含多个指标时, 样本为多元的, 指标或观测变量的个数就是数据的维数. 当数据维数远小于所有组总样本数时, 假若各组协方差矩阵是相同的, 解决双因素多元方差分析问题可以使用 Wilks 似然比 (Wilks' likelihood ratio, WLR) 检验、Lawley-Hotelling 迹 (Lawley-Hotelling trace, LHT) 检验、Bartlett-Nanda-Pillai (BNP) 迹检验和 Roy 的最大根检验等方

英文引用格式: Zhou B, Guo J, Zhang J T. An L^2-norm based test for high-dimensional two-way MANOVA (in Chinese). Sci Sin Math, 2020, 50: 729–750, doi: 10.1360/N012019-00060

法 (参见文献 [5, 第 8 章] 和 [6, 第 6.4.2 小节]). 而当同协方差阵假设不成立时, 文献 [7–10] 及其引用的相关文献对该问题也进行了一些研究.

高维数据给很多经典的统计问题带来新的挑战. 高维数据, 顾名思义就是变量个数一般很大, 接近甚至可能远远大于数据观测样本数. 在这种情况下, 样本协方差矩阵是奇异的, 导致许多经典的统计方法都不能使用. 高维数据也常出现在对较少的个体做出很多种不同类型测量的情形下. 一个具有启发性的例子就是 Grenoble 大学体育运动实验室 (Laboratoire Sport et Performance Motrice, EA 597, UFRAPS, Grenoble University, France) 的 David Amarantini 和 Luc Martin 通过实验得到的矫正器数据 (orthosis data). 该实验的目的是研究肌肉如何应对外部的扰动. 在实验中, 志愿者穿着一个可调节刚度的弹簧矫形器, 以模拟外部扰动. 每次实验以 200 Hz 的采样频率连续记录 10 s 的相关数据, 然后通过身体部分运动学计算出膝盖处的合成力矩变化曲线. 这个数据集包含了 7 位年轻男志愿者在 4 种不同实验条件下的数据—对照条件 (无矫形器)、矫形器状况 (仅矫形器) 和两个弹簧条件 (弹簧 1 和 2). 对于每个受试者和每个实验条件, 分别记录 10 次重复实验的结果; 而对于每次重复实验, 分别在 256 个等距时间点上计算膝盖上的力矩. 关于该实验和数据更详细的介绍可参见文献 [11, 第 1.2.7 小节]. 图 1 展示了前两个条件 (对照条件和矫正条件) 下前三个受试者的原始数据. 我们关注的是不同的实验对象和实验条件是否会对实验结果膝盖处的平均力矩有影响. 注意到, 此数据的维度为 256, 远大于各组样本数 10.

因为关注两个不同因素的效应, 且上述数据的维度远大于各组样本数, 该问题可以看作一个高维双因素多元方差分析问题. 在高维双因素多元方差分析问题中, 由于样本协方差阵通常是奇异的或接近奇异的, 前面提到的经典检验方法不再适用. 据了解, 目前仅有较少的工作研究高维数据的双因素多元方差分析问题, 尽管该问题在实际中也已经很常见. 较为相关的工作多见于高维单因素方差问题中, 只限于研究高维数据实验中单个因子不同水平的效应, 如 k 个不同水平的效应. 当 $k = 2$ 时, 该类问题就是高维数据的两总体检验问题. Dempster[12, 13] 较早关注了该问题并提出了一个非精确检验. 文献 [14] 推导出了 Dempster 检验的渐近正态性并提出了一个新的非精确渐近正态检验, 文献 [15] 提出了一个尺度不变的检验, 文献 [16] 研究了基于 U 统计量的检验方法. 在一些正则条件下, 这些文献的作者皆证明其提出的检验统计量有渐近正态分布. 对于一般的 $k \geqslant 3$, 该类问题就是高维数据的单因素方差分析问题. 对此, 文献 [17] 研究了一个基于 Dempster 迹的检验方法, 文献 [18] 通过扩展文献 [14] 的工作提出了一个非精确检验, 文献 [19] 通过扩展文献 [15] 的工作提出了一个尺度不变的检验, 而文献 [20] 提出了一个基于 L^2 范数的检验方法. 对于 k 个总体的协方差矩阵不相同时的异方差高维单因素方差分析问题, 也已经有相关工作进行了研究. 其中文献 [21] 提出了一个基于 Bennett[22] 变换的检验, 文献 [23–25] 各自研究了渐近正态的检验方法.

作为对文献 [20] 的推广和扩展, 本文提出并研究一个基于 L^2 范数的高维数据双因素方差分析方法. 对于正态数据, 我们所提出的基于 L^2 范数的检验统计量的零分布是一个卡方型混合分布, 这表明检验统计量零分布的形状通常是偏斜的. 因此, 文献中常见的正态近似可能无法很好地用于近似该检验统计量的零分布. 为了解决这个问题, 我们提出使用 Welch-Satterthwaite 卡方近似方法来逼近统计量的零分布. 我们将讨论何时零分布会渐近趋向正态分布, 并比较正态近似和卡方近似两种方法.

本文其余部分内容安排如下. 检验方法的原理在第 2 节中介绍. 其中, 第 2.1 小节简要回顾双因素多元方差分析的一些基本结果, 第 2.2 小节构造出基于 L^2 范数的检验统计量, 第 2.3 小节研究检验的零分布, 第 2.4 小节给出统计量零分布的近似方法. 第 3 节给出统计量的一些不变性质. 检验的近似与渐近功效函数在第 4 节中推导得出. 模拟实验和有关实际数据的应用分别在第 5 和 6 节中展示. 第 7 节总结全文并提出几点值得进一步研究的问题. 主要定理的技术证明可参见附录.

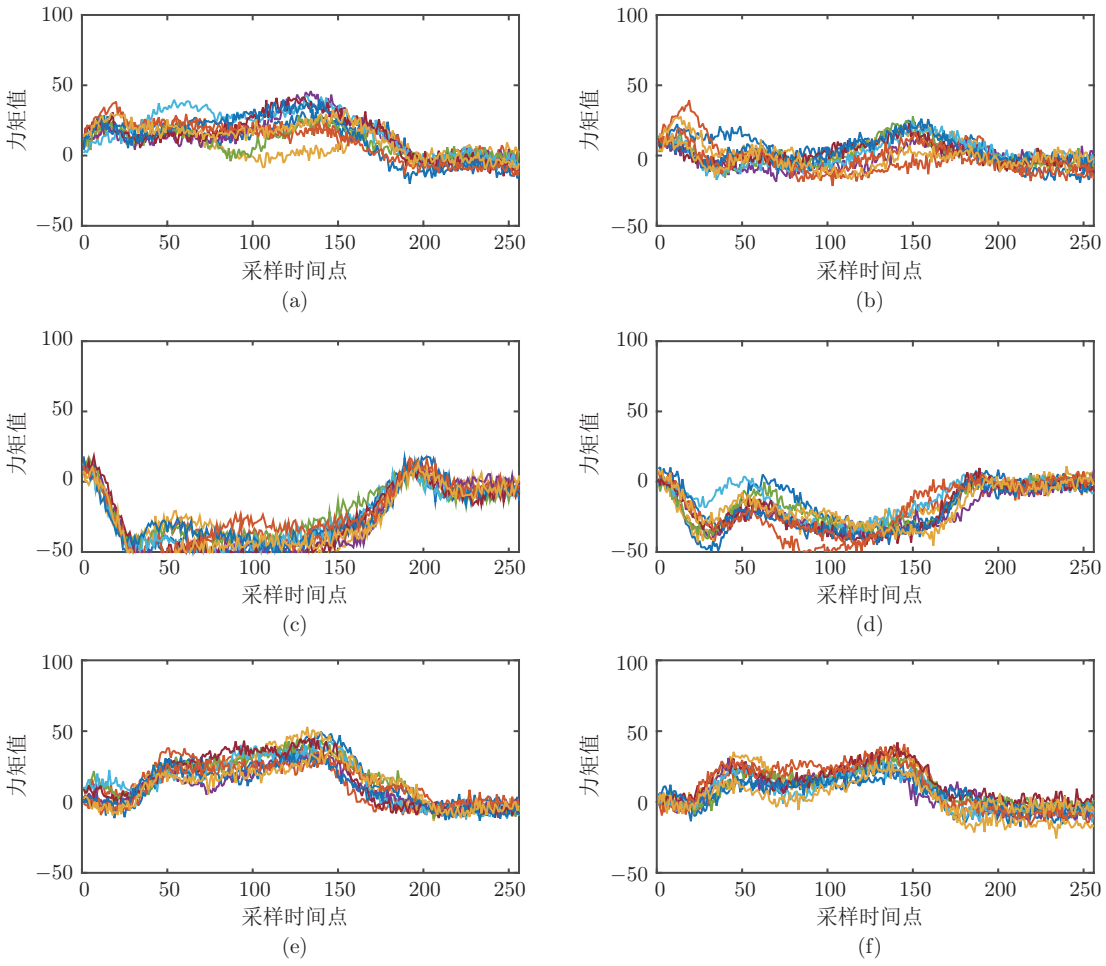

图 1　(网络版彩图) 矫形器数据中前两个条件下前 3 个受试者的膝盖处合成力矩原始曲线. 同一子图中 10 条不同颜色的曲线分别代表 10 次不同的重复观测. 其中不同行的子图对应不同的受试者的样本观测, 不同列的子图对应不同的实验条件 (对照和矫正)

2　主要方法

2.1　多元双因素方差分析模型中的主效应与交互效应

考虑一个有两个因素的双向实验, 其中两个因素 A 和 B 分别有 a 和 b 个水平. 假设在第 (i,j) 组, $i=1,\ldots,a$, $j=1,\ldots,b$, 有以下 p 维独立同分布 (i.i.d.) 的随机样本:

$$\boldsymbol{y}_{ijk}=\boldsymbol{\mu}_{ij}+\boldsymbol{\epsilon}_{ijk},\quad \mathrm{E}(\boldsymbol{\epsilon}_{ijk})=\boldsymbol{0},\quad \mathrm{Cov}(\boldsymbol{\epsilon}_{ijk})=\boldsymbol{\Sigma},\quad k=1,\ldots,n_{ij}, \tag{2.1}$$

其中 $\boldsymbol{\mu}_{ij}:p\times 1$ 和 $\boldsymbol{\Sigma}:p\times p$ 分别是第 (i,j) 组的均值向量和各组公共的协方差矩阵. 假设不同组的样本相互独立. (2.1) 可以被进一步写成众所周知的双向 MANOVA 模型

$$\boldsymbol{y}_{ijk}=\boldsymbol{\mu}_0+\boldsymbol{\alpha}_i+\boldsymbol{\beta}_j+\boldsymbol{\gamma}_{ij}+\boldsymbol{\epsilon}_{ijk},\quad k=1,\ldots,n_{ij},\quad i=1,\ldots,a,\quad j=1,\ldots,b, \tag{2.2}$$

其中 $\boldsymbol{\mu}_0$ 是总均值向量, $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\beta}_j$ 分别为因素 A 和 B 的第 i 和 j 个主效应, $\boldsymbol{\gamma}_{ij}$ 是因素 A 和 B 的第 (i,j) 个交互效应.

对于模型 (2.2), 我们感兴趣的是测试两个因素的主效应及其交互效应是否显著. 三个零假设可分别表示如下:

$$\begin{aligned} &H_{0A}: \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_a, \\ &H_{0B}: \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_b, \\ &H_{0AB}: \boldsymbol{\gamma}_{11} = \cdots = \boldsymbol{\gamma}_{1b} = \cdots = \boldsymbol{\gamma}_{a1} = \cdots = \boldsymbol{\gamma}_{ab}. \end{aligned} \tag{2.3}$$

进一步定义

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_a^\top]^\top, \quad \boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_b^\top]^\top, \quad \boldsymbol{\gamma} = [\boldsymbol{\gamma}_{11}^\top, \ldots, \boldsymbol{\gamma}_{1b}^\top, \ldots, \boldsymbol{\gamma}_{a1}^\top, \ldots, \boldsymbol{\gamma}_{ab}^\top]^\top,$$

则三个零假设 (2.3) 还可以用约束矩阵等价地写为

$$\begin{aligned} &H_{0A}: (\boldsymbol{H}_a \otimes \boldsymbol{I}_p)\boldsymbol{\alpha} = \boldsymbol{0}, \quad \boldsymbol{H}_a = (\boldsymbol{I}_{a-1}, -\boldsymbol{1}_{a-1}), \\ &H_{0B}: (\boldsymbol{H}_b \otimes \boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{0}, \quad \boldsymbol{H}_b = (\boldsymbol{I}_{b-1}, -\boldsymbol{1}_{b-1}), \\ &H_{0AB}: (\boldsymbol{H}_{ab} \otimes \boldsymbol{I}_p)\boldsymbol{\gamma} = \boldsymbol{0}, \quad \boldsymbol{H}_{ab} = (\boldsymbol{I}_{a-1}, -\boldsymbol{1}_{a-1}) \otimes (\boldsymbol{I}_{b-1}, -\boldsymbol{1}_{b-1}), \end{aligned} \tag{2.4}$$

其中 $\boldsymbol{I}_r$ 和 $\boldsymbol{1}_r$ 分别表示 r 维的单位矩阵和 $r \times 1$ 维的元素皆为 1 的向量, $\otimes$ 表示矩阵的 Kronecker 乘积.

本小节的目标是将主效应和交互效应都表示为双因素 MANOVA 模型 (2.2) 中可以被估计的各组均值向量的线性组合. 然而模型 (2.2) 是不可识别的, 因为 (2.2) 中的参数 $\boldsymbol{\mu}_0$、$\boldsymbol{\alpha}_i$、$\boldsymbol{\beta}_j$ 和 $\boldsymbol{\gamma}_{ij}$ 没有唯一的定义. 为了使模型可识别, 应该施加一些限制条件. 给定一系列的权重系数 $w_{ij} > 0$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 与文献 [7] 一样, 我们施加以下约束:

$$\sum_{i=1}^{a} w_{i\cdot}\boldsymbol{\alpha}_i = \boldsymbol{0}, \quad \sum_{j=1}^{b} w_{\cdot j}\boldsymbol{\beta}_j = \boldsymbol{0}, \tag{2.5}$$

$$\sum_{j=1}^{b} w_{ij}\boldsymbol{\gamma}_{ij} = \boldsymbol{0}, \quad i = 1, \ldots, a-1, \tag{2.6}$$

$$\sum_{i=1}^{a} w_{ij}\boldsymbol{\gamma}_{ij} = \boldsymbol{0}, \quad j = 1, \ldots, b-1, \tag{2.7}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b} w_{ij}\boldsymbol{\gamma}_{ij} = \boldsymbol{0}, \tag{2.8}$$

其中 $w_{i\cdot} = \sum_{j=1}^{b} w_{ij}$, $w_{\cdot j} = \sum_{i=1}^{a} w_{ij}$.

有多种方法可用于设定权重 w_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 具体可参见文献 [7,26]. 在本文中, 与文献 [7] 一样, 我们只考虑等权重和样本数适应权重两种简单方法. 对于这两种方法, 权重都可以被分解为 $w_{ij} = u_i v_j$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 且满足 $u_i > 0$, $\sum_{i=1}^{a} u_i = 1$; $v_j > 0$, $\sum_{j=1}^{b} v_j = 1$. 对于等权重方法, $u_i = 1/a$, $v_j = 1/b$, $i = 1, \ldots, a$, $j = 1, \ldots, b$. 而对于样本数适应权重方法, $u_i = \sum_{j=1}^{b} n_{ij}/n$, $i = 1, \ldots, a$; $v_j = \sum_{i=1}^{a} n_{ij}/n$, $j = 1, \ldots, b$, 其中样本总数 $n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$. 对于这两种加权方法, 在约束 (2.5)–(2.8) 下, 通过一些代数运算可以得到

$$\begin{aligned} &\boldsymbol{\alpha} = (\boldsymbol{A}_a \otimes \boldsymbol{I}_p)\boldsymbol{\mu}, \quad \boldsymbol{A}_a = (\boldsymbol{I}_a - \boldsymbol{1}_a\boldsymbol{u}^\top) \otimes \boldsymbol{v}^\top, \\ &\boldsymbol{\beta} = (\boldsymbol{A}_b \otimes \boldsymbol{I}_p)\boldsymbol{\mu}, \quad \boldsymbol{A}_b = \boldsymbol{u}^\top \otimes (\boldsymbol{I}_b - \boldsymbol{1}_b\boldsymbol{v}^\top), \\ &\boldsymbol{\gamma} = (\boldsymbol{A}_{ab} \otimes \boldsymbol{I}_p)\boldsymbol{\mu}, \quad \boldsymbol{A}_{ab} = (\boldsymbol{I}_a - \boldsymbol{1}_a\boldsymbol{u}^\top) \otimes (\boldsymbol{I}_b - \boldsymbol{1}_b\boldsymbol{v}^\top), \end{aligned} \tag{2.9}$$

其中 $\boldsymbol{\mu}=[\boldsymbol{\mu}_{11}^\top,\ldots,\boldsymbol{\mu}_{1b}^\top,\ldots,\boldsymbol{\mu}_{a1}^\top,\ldots,\boldsymbol{\mu}_{ab}^\top]^\top$, 矩阵 $\boldsymbol{A}_a$、$\boldsymbol{A}_b$ 和 $\boldsymbol{A}_{ab}$ 不满秩, 其秩分别是 $a-1$、$b-1$ 和 $(a-1)(b-1)$.

一旦将主效应和交互效应都表示为双因素 MANOVA 模型 (2.2) 中可以被估计的组均值的线性组合 (2.9), 三个零假设 (2.4) 对应的检验问题就可以被等价地表示成一般线性假设检验 (general linear hypothesis testing, GLHT) 问题 (2.12) 的形式. 一般线性假设检验问题可以看作广义线性模型检验问题的一种特殊形式. 对于传统的非高维背景下的广义线性模型, 在张尧庭和方开泰的经典专著《多元统计分析引论》[27] 的第七章中有详细的介绍. 而作为一个相关的问题, 一般线性假设检验问题在本文下一节中有具体的表述. 其中三个零假设 (2.4) 对应的一般线性假设的系数矩阵分别为

$$\boldsymbol{C}_a=(\boldsymbol{H}_a\boldsymbol{A}_a)\otimes\boldsymbol{I}_p,\quad \boldsymbol{C}_b=(\boldsymbol{H}_b\boldsymbol{A}_b)\otimes\boldsymbol{I}_p,\quad \boldsymbol{C}_{ab}=(\boldsymbol{H}_{ab}\boldsymbol{A}_{ab})\otimes\boldsymbol{I}_p. \tag{2.10}$$

将 (2.4) 和 (2.9) 代入 (2.10), 并注意到 $\boldsymbol{H}_a\boldsymbol{1}_a=\boldsymbol{0}$, $\boldsymbol{H}_b\boldsymbol{1}_b=\boldsymbol{0}$, 有

$$\boldsymbol{C}_a=(\boldsymbol{H}_a\otimes\boldsymbol{v}^\top)\otimes\boldsymbol{I}_p,\quad \boldsymbol{C}_b=(\boldsymbol{u}^\top\otimes\boldsymbol{H}_b)\otimes\boldsymbol{I}_p,\quad \boldsymbol{C}_{ab}=(\boldsymbol{H}_a\otimes\boldsymbol{H}_b)\otimes\boldsymbol{I}_p. \tag{2.11}$$

从 (2.11) 可知, $\boldsymbol{C}_a$ 不依赖于 $\boldsymbol{u}$, $\boldsymbol{C}_b$ 不依赖于 $\boldsymbol{v}$, 以及 $\boldsymbol{C}_{ab}$ 不依赖于 $\boldsymbol{u}$ 和 $\boldsymbol{v}$. 注意到矩阵 $\boldsymbol{H}_a\otimes\boldsymbol{v}^\top$、$\boldsymbol{u}^\top\otimes\boldsymbol{H}_b$ 和 $\boldsymbol{H}_a\otimes\boldsymbol{H}_b$ 皆行满秩, 其秩分别是 $(a-1)$、$(b-1)$ 和 $(a-1)(b-1)$.

2.2 基于 L^2 范数的一般线性假设检验统计量

使用前一小节中定义的向量 $\boldsymbol{\mu}$, 在双因素 MANOVA 模型 (2.2) 背景下考虑 GLHT 问题

$$H_0:\boldsymbol{C}\boldsymbol{\mu}=\boldsymbol{0}\quad\text{vs.}\quad H_1:\boldsymbol{C}\boldsymbol{\mu}\neq\boldsymbol{0}, \tag{2.12}$$

其中 $\boldsymbol{C}=\boldsymbol{C}_0\otimes\boldsymbol{I}_p$, $\boldsymbol{C}_0:q\times(ab)$ 是一个已知的行满秩矩阵. 第 2.1 小节中的三个检验问题 (2.3) 所对应的 $\boldsymbol{C}$ 矩阵已在 (2.11) 中给出. 注意到假设 (2.12) 非常通用, 我们也可以通过选择特定的矩阵 $\boldsymbol{C}$ 来构建事后多重比较和两两比较检验. 例如, 把 (2.10) 中的 $\boldsymbol{H}_a$ 换为一个 $(a-1)\times a$ 的对比矩阵, 其第一行的前两个元素分别为 1 和 -1, 其余所有元素为 0, 我们就可以测试第一个因素的前两个水平的差异.

为了构建 GLHT 问题 (2.12) 的检验统计量, 将模型 (2.1) 的组均值向量和组协方差矩阵通常的无偏估计分别记为

$$\hat{\boldsymbol{\mu}}_{ij}=n_{ij}^{-1}\sum_{k=1}^{n_{ij}}\boldsymbol{y}_{ijk},\quad \hat{\boldsymbol{\Sigma}}_{ij}=(n_{ij}-1)^{-1}\sum_{k=1}^{n_{ij}}(\boldsymbol{y}_{ijk}-\hat{\boldsymbol{\mu}}_{ij})(\boldsymbol{y}_{ijk}-\hat{\boldsymbol{\mu}}_{ij})^\top,\quad i=1,\ldots,a,\quad j=1,\ldots,b. \tag{2.13}$$

令 $\hat{\boldsymbol{\mu}}=[\hat{\boldsymbol{\mu}}_{11}^\top,\ldots,\hat{\boldsymbol{\mu}}_{1b}^\top,\ldots,\hat{\boldsymbol{\mu}}_{a1}^\top,\ldots,\hat{\boldsymbol{\mu}}_{ab}^\top]^\top$ 为 $\boldsymbol{\mu}$ 的估计, 并记 $\boldsymbol{D}_0=\mathrm{diag}(n_{11}^{-1},\ldots,n_{1b}^{-1},\ldots,n_{a1}^{-1},\ldots,n_{ab}^{-1})$, 则有

$$\mathrm{E}(\boldsymbol{C}\hat{\boldsymbol{\mu}})=\boldsymbol{C}\boldsymbol{\mu},\quad \mathrm{Cov}(\boldsymbol{C}\hat{\boldsymbol{\mu}})=\boldsymbol{C}\boldsymbol{\Sigma}_n\boldsymbol{C}^\top, \tag{2.14}$$

其中 $\boldsymbol{\Sigma}_n=\boldsymbol{D}_0\otimes\boldsymbol{\Sigma}$. 令 $\hat{\boldsymbol{\Sigma}}_n=\boldsymbol{D}_0\otimes\hat{\boldsymbol{\Sigma}}$, 其中 $\hat{\boldsymbol{\Sigma}}=(n-ab)^{-1}\sum_{i=1}^a\sum_{j=1}^b(n_{ij}-1)\hat{\boldsymbol{\Sigma}}_{ij}$ 为合并样本协方差阵. 当 $\hat{\boldsymbol{\Sigma}}_n$ 可逆时, 问题 (2.12) 的 Wald 类型检验统计量为

$$T_w=(\boldsymbol{C}\hat{\boldsymbol{\mu}})^\top(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_n\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\mu}}). \tag{2.15}$$

但对于引言中介绍的高维数据, 维数 p 可以远远大于各组样本数 n_{ij}, 甚至总样本数 n, 所以, $\hat{\boldsymbol{\Sigma}}_n$ 可能是奇异的, 这时统计量 (2.15) 没有定义. 考虑到这一点, 把 (2.15) 中的 $\hat{\boldsymbol{\Sigma}}_n$ 用 $\boldsymbol{D}=\boldsymbol{D}_0\otimes\boldsymbol{I}_p$ 代替, 那么对 GLHT 问题 (2.12) 就可以构造出一个新的检验统计量

$$T_n=(\boldsymbol{C}\hat{\boldsymbol{\mu}})^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}(\boldsymbol{C}\hat{\boldsymbol{\mu}})=\|(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1/2}\boldsymbol{C}\hat{\boldsymbol{\mu}}\|^2, \tag{2.16}$$

其中 $\|\boldsymbol{x}\|^2=\boldsymbol{x}^{\top}\boldsymbol{x}$ 表示一个向量 $\boldsymbol{x}$ 的 L^2 范数的平方, $\boldsymbol{M}^{1/2}$ 表示矩阵 $\boldsymbol{M}$ 的矩阵平方根, $\boldsymbol{M}^{-1/2}$ 表示矩阵 $\boldsymbol{M}$ 的矩阵平方根的逆矩阵. 在统计量 T_n 中, 用 L^2 范数来度量 $\boldsymbol{C}\boldsymbol{\mu}$ 的无偏估计 $\boldsymbol{C}\hat{\boldsymbol{\mu}}$ 与零向量 $\boldsymbol{0}$ 的差异. 显然在 H_0 成立时, T_n 的值应该较小, 反之偏大.

记

$$\boldsymbol{x}_{ij}=\hat{\boldsymbol{\mu}}_{ij}-\boldsymbol{\mu}_{ij},\quad i=1,\ldots,a,\quad j=1,\ldots,b. \tag{2.17}$$

那么 $\boldsymbol{x}_{ij}$ $(i=1,\ldots,a, j=1,\ldots,b)$ 相互独立且 $\mathrm{E}(\boldsymbol{x}_{ij})=\boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{x}_{ij})=\boldsymbol{\Sigma}/n_{ij}$. 令 $\boldsymbol{x}=[\boldsymbol{x}_{11}^{\top},\ldots,\boldsymbol{x}_{1b}^{\top},\ldots,\boldsymbol{x}_{a1}^{\top},\ldots,\boldsymbol{x}_{ab}^{\top}]^{\top}$, 则有

$$T_n=(\boldsymbol{C}\boldsymbol{x})^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}(\boldsymbol{C}\boldsymbol{x})+2(\boldsymbol{C}\boldsymbol{x})^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}(\boldsymbol{C}\boldsymbol{\mu})+(\boldsymbol{C}\boldsymbol{\mu})^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}(\boldsymbol{C}\boldsymbol{\mu}). \tag{2.18}$$

可以看出, 在上面的表达式中, 右侧的第一项与组均值向量 $\boldsymbol{\mu}_{ij}$ $(i=1,\ldots,a, j=1,\ldots,b)$ 无关, 第二项的均值为零, 第三项为各组均值向量的一个函数. 此外, 当 (2.12) 中的零假设 H_0 为真时, 最后两项为零, 否则为非零. 将 (2.18) 右侧的第一项记为

$$T_{n0}=(\boldsymbol{C}\boldsymbol{x})^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}(\boldsymbol{C}\boldsymbol{x}), \tag{2.19}$$

它与零假设下的 T_n 具有相同的分布. 也就是说, 在 H_0 下, 有 $T_{n0}\overset{d}{=}T_n$, 其中 $X\overset{d}{=}Y$ 表示随机变量 X 和 Y 具有相同的分布.

为了检验 (2.12), 需要推导出 T_{n0} 的分布. 令 $\boldsymbol{x}_{n,ij}=\sqrt{n_{ij}}\boldsymbol{x}_{ij}$, $i=1,\ldots,a$, $j=1,\ldots,b$, 并记 $\boldsymbol{x}_n=[\boldsymbol{x}_{n,11}^{\top},\ldots,\boldsymbol{x}_{n,1a}^{\top},\ldots,\boldsymbol{x}_{n,a1}^{\top},\ldots,\boldsymbol{x}_{n,ab}^{\top}]^{\top}$, 有

$$\mathrm{E}(\boldsymbol{x}_n)=\boldsymbol{0},\quad \mathrm{Cov}(\boldsymbol{x}_n)=\boldsymbol{I}_{ab}\otimes\boldsymbol{\Sigma}. \tag{2.20}$$

令 $\boldsymbol{A}=\boldsymbol{D}^{1/2}\boldsymbol{C}^{\top}(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^{\top})^{-1}\boldsymbol{C}\boldsymbol{D}^{1/2}=\boldsymbol{A}_0\otimes\boldsymbol{I}_p$, 其中 $\boldsymbol{A}_0=\boldsymbol{D}_0^{1/2}\boldsymbol{C}_0^{\top}(\boldsymbol{C}_0\boldsymbol{D}_0\boldsymbol{C}_0^{\top})^{-1}\boldsymbol{C}_0\boldsymbol{D}_0^{1/2}$, 很容易验证 $\boldsymbol{A}_0$ 是一个幂等阵, 所以, $\boldsymbol{A}_0^2=\boldsymbol{A}_0$ 且 $\mathrm{tr}(\boldsymbol{A}_0)=q$. 有

$$T_{n0}=\boldsymbol{x}_n^{\top}\boldsymbol{A}\boldsymbol{x}_n=\sum_{i,i_1=1}^{a}\sum_{j,j_1=1}^{b}a_{0,(i,j)(i_1,j_1)}\boldsymbol{x}_{n,ij}^{\top}\boldsymbol{x}_{n,i_1j_1}, \tag{2.21}$$

其中 $a_{0,(i,j)(i_1,j_1)}$ 表示 $\boldsymbol{A}_0$ 的第 $[((i-1)b+j),((i_1-1)b+j_1)]$ 个元素. 先计算 T_n 在零假设下的前两阶矩. 记

$$\kappa_{ij}=\mathrm{E}\,\|\boldsymbol{y}_{ij1}-\boldsymbol{\mu}_{ij}\|^4-\mathrm{tr}^2(\boldsymbol{\Sigma})-2\,\mathrm{tr}(\boldsymbol{\Sigma}^2),\quad i=1,\ldots,a,\quad j=1,\ldots,b, \tag{2.22}$$

κ_{ij} 为衡量随机向量 $\boldsymbol{y}_{ij1}$ 的非正态性的重要参数 [28]. 当样本 (2.1) 服从正态分布时, 有 $\kappa_{ij}=0$, $i=1,\ldots,a$, $j=1,\ldots,b$.

以下定理给出了 T_{n0} 的前两阶矩.

定理 2.1 T_{n0} 的前两阶矩分别为

$$\mathrm{E}(T_{n0}) = q\,\mathrm{tr}(\boldsymbol{\Sigma}), \quad \mathrm{Var}(T_{n0}) = 2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + \sum_{i=1}^{a}\sum_{j=1}^{b} a_{0,(i,j)}^2 \frac{\kappa_{ij}}{n_{ij}}, \tag{2.23}$$

其中 $a_{0,(i,j)}$ 是 $\boldsymbol{A}_0$ 的第 $(i-1)b+j$ 个对角元素.

从 (2.23) 可看出, 数据的平均化极大减弱了数据的非正态性对 T_n 的影响. 在本文中, 令 χ_d^2 表示自由度为 d 的卡方分布. 对正态数据, 表达式 (2.23) 可以简化为下面推论中给出的形式.

推论 2.1 假设 $\boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $k = 1, \ldots, n_{ij}$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 则对于任何固定的 n_{ij} $(i = 1, \ldots, a,\ j = 1, \ldots, b)$ 和 p, 有

$$T_{n0} \overset{d}{=} \sum_{r=1}^{p} \lambda_r A_r, \quad \text{其中 } A_r\ (r = 1, \ldots, p) \overset{\text{i.i.d.}}{\sim} \chi_q^2, \tag{2.24}$$

λ_r $(r = 1, \ldots, p)$ 为 $\boldsymbol{\Sigma}$ 的按降序排列的特征值. 此外, $\mathrm{E}(T_{n0}) = q\,\mathrm{tr}(\boldsymbol{\Sigma})$, $\mathrm{Var}(T_{n0}) = 2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)$, $\mathrm{E}[T_{n0} - \mathrm{E}(T_{n0})]^3 = 8q\,\mathrm{tr}(\boldsymbol{\Sigma}^3)$, 且 T_{n0} 的偏度可以表示为

$$\frac{\mathrm{E}[T_{n0} - \mathrm{E}(T_{n0})]^3}{\mathrm{Var}^{3/2}(T_{n0})} = \left(\frac{8}{d^*}\right)^{1/2}, \quad \text{其中 } d^* = q\frac{\mathrm{tr}^3(\boldsymbol{\Sigma}^2)}{\mathrm{tr}^2(\boldsymbol{\Sigma}^3)}. \tag{2.25}$$

(2.24) 是在固定了样本数 n_{ij} 和维数 p 下得到的. 注意到 (2.24) 与样本数 n_{ij} 无关, 所以只要样本服从多元正态分布, 即使在 $n_{ij} \to \infty$ 时, (2.24) 的结论仍然成立. 另一方面, 如果 $\mathrm{tr}(\boldsymbol{\Sigma}) < \infty$, 当 $p \to \infty$ 时, 类似于文献 [20, 注 1] 的结论, 也有 (2.24) 在极限意义下成立. 更进一步, 在一些正则条件下, 即使样本不服从正态分布, 还可以证明, 当样本数和维数同时趋于无穷时, (2.24) 也在某种极限意义下成立 [29].

2.3 T_{n0} 的近似与渐近分布

当样本服从正态分布时, 从 (2.24) 可知 T_{n0} 的分布是具有未知系数 λ_r $(r = 1, \ldots, p)$ 的卡方型混合分布 [30]. 所以 T_{n0} 是非负的, 且其分布通常是有偏的. 这启发了我们应用著名的 Welch-Satterthwaite 卡方近似方法, 也称为 Box 卡方近似方法 [31], 来近似 T_{n0} 的分布. 这种近似方法通常是非常准确的, 并且被广泛地用于解决单变量数据的 Behrens-Fisher 问题 [32, 33]. 文献 [34] 也将该近似方法推广并应用到了多变量正态数据的双因素 MANOVA 问题中. 对于样本可能不是正态分布的一般情况, 我们发现这种基于两阶矩匹配逼近 T_{n0} 分布的近似方法仍然是非常准确和有效的. 一是因为 T_{n0} 始终是非负的并且经常是有偏的, 二是因为统计量中对样本的平均显著降低了数据非正态性对 T_{n0} 分布的影响.

Welch-Satterthwaite 卡方近似的关键思想是通过匹配 T_{n0} 和一个形如 $R = \beta\chi_d^2$ 的随机变量的前两阶矩来逼近 T_{n0} 的分布. 因此, Welch-Satterthwaite 卡方近似也被称为两阶矩匹配的卡方近似. 在 $\mathrm{E}(T_{n0}) = \mathrm{E}(R)$ 和 $\mathrm{Var}(T_{n0}) = \mathrm{Var}(R)$ 的条件下, 通过简单的运算可以得到参数

$$\beta = \frac{\mathrm{tr}(\boldsymbol{\Sigma}^2) + \sum_{i=1}^{a}\sum_{j=1}^{b} a_{0,(i,j)}^2 \frac{\kappa_{ij}}{2qn_{ij}}}{\mathrm{tr}(\boldsymbol{\Sigma})}, \quad d = \frac{q\,\mathrm{tr}^2(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma}^2) + \sum_{i=1}^{a}\sum_{j=1}^{b} a_{0,(i,j)}^2 \frac{\kappa_{ij}}{2qn_{ij}}}. \tag{2.26}$$

然后有

$$T_{n0} \overset{d}{\approx} R = \beta\chi_d^2, \tag{2.27}$$

其中 $\overset{d}{\approx}$ 表示近似 (在前两阶矩相等的意义下) 同分布.

一个能说明卡方近似 (2.27) 合理性的事实是, 在一些正则条件下, T_{n0} 和 R 具有相同的正态极限分布. 为了在高维数据背景下研究一般情况下 T_n 的渐近零分布, 需要一些额外的条件. 假设样本 (2.1) 来自文献 [14,16] 中使用的因子模型

$$\boldsymbol{y}_{ijk} = \boldsymbol{\Gamma}\boldsymbol{z}_{ijk} + \boldsymbol{\mu}_{ij}, \quad k = 1, \ldots, n_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b, \tag{2.28}$$

其中矩阵 $\boldsymbol{\Gamma} : p \times m$ 满足 $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{\Sigma}$, $\boldsymbol{z}_{ijk}$ 是一个 $m \times 1$ 维的随机向量并满足分别对 $i = 1, \ldots, a$, $j = 1, \ldots, b$, $\boldsymbol{z}_{ijk}$ $(k = 1, \ldots, n_{ij})$ 独立同分布且 $\mathrm{E}(\boldsymbol{z}_{ijk}) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{z}_{ijk}) = \boldsymbol{I}_m$, 并且当 $(i, j) \neq (i_1, j_1)$ 时, $\boldsymbol{z}_{ij}$ 与 $\boldsymbol{z}_{i_1 j_1}$ 独立. 可以证明在下列条件下, 当 $n, p \to \infty$ 时, T_{n0} 和 $R = \beta\chi^2_d$ 都服从渐近正态分布.

假设 A (A1) 当 n_{ij} 和 $n \to \infty$ 时, $n_{ij}/n \to \tau_{ij} \in (0, 1)$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 其中 $n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$.

(A2) $\mathrm{E}[z^4_{ijk(\ell)}] < \infty$, 其中 $z_{ijk(\ell)}$ 是 $\boldsymbol{z}_{ijk}$ 的第 ℓ 个分量, $\ell = 1, \ldots, m$.

(A3) 对任何满足 $\sum_{\ell=1}^{r} \alpha_\ell \leqslant 8$ 的正整数 r 和 $\ell_1 \neq \cdots \neq \ell_r$, 有

$$\mathrm{E}[z^{\alpha_1}_{ijk(\ell_1)} \cdots z^{\alpha_r}_{ijk(\ell_r)}] = \mathrm{E}[z^{\alpha_1}_{ijk(\ell_1)}] \cdots \mathrm{E}[z^{\alpha_r}_{ijk(\ell_r)}].$$

(A4) 随着 $p \to \infty$, $\boldsymbol{\Sigma}$ 的最大特征值 $\lambda_{\max} = o[\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]$.

假设 (A1) 保证了各组样本量大小相当. 假设 (A2) 和 (A3) 是文献 [14] 所用到的假设的扩展, 它们用来减少样本的非正态性给统计量的高阶矩计算和表达带来的困难. 而假设 (A4) 保证了高维数据各个分量之间是弱相关的, 这是高维数据分析中一种常见的稀疏性假设.

定理 2.2 在假设 A 下, 随着 $n, p \to \infty$, 有 $d \to \infty$. 此外,

$$\frac{T_{n0} - q\,\mathrm{tr}(\boldsymbol{\Sigma})}{[2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \quad \text{且} \quad \frac{\beta\chi^2_d - q\,\mathrm{tr}(\boldsymbol{\Sigma})}{[2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{2.29}$$

根据定理 2.2, 在假设 A 下可以同时得到 T_{n0} 和 $R = \beta\chi^2_d$ 的渐近正态性, 这说明两阶矩匹配卡方近似方法在假设 A 下是渐近精确的. 定理 2.2 也表明当 $n, p \to \infty$ 时, 若 $d = O(1)$, 假设 A 可能会被违反, 否则当 $n, p \to \infty$ 时, d 会趋向于 ∞. 从附录中的定理 2.2 的证明可以看出, 当 d 有界时, 假设 (A4) 不能成立.

为方便起见, 可以将 d 称为两阶矩匹配卡方近似的近似自由度. 当样本正态性成立时, 可以得到以下更强的结果: 当 $p \to \infty$ 时, $d^* \to \infty$ 是 T_{n0} 渐近正态的充分必要条件, 其中 d^* 在 (2.25) 中定义. 像往常一样, 本文用 $\xrightarrow{\mathcal{L}}$ 和 $\xrightarrow{P}$ 分别表示依分布和依概率收敛.

推论 2.2 假设 $\boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $k = 1, \ldots, n_{ij}$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 那么随着 $p \to \infty$, T_{n0} 服从渐近正态分布当且仅当 $d^* \to \infty$. 另外, 有 $0 \leqslant d^* \leqslant d$. 更进一步, 随着 $d^* \to \infty$, 有 (2.29) 成立.

推论 2.2 表明, 若样本正态性成立, 如果 d 有界, d^* 也有界, 这时 T_{n0} 的极限分布不是正态分布. 因此, 当样本的正态性成立时, 可以使用 d 来确定 (2.29) 中的正态近似是否适合近似 T_{n0}. 例如, 当 $d < 10$ 时, 总有 $d^* < 10$, 这时正态近似是不够精确的, 应该尝试使用两阶矩匹配的卡方近似方法. 另一方面, 当 d^* 很大时, 由推论 2.2 可知两阶矩匹配的卡方近似方法会非常接近正态近似方法. 所以一般来讲, 两阶矩匹配的卡方近似方法比正态近似方法有更广泛的适用范围. 事实上, 文献 [29] 在高维两总体的背景下给出了两阶矩匹配的卡方近似的一致误差界, 并证明了两阶矩匹配的卡方近似方法通常比正态近似能更精确地逼近卡方型混合分布. 特别需要指出的是, 文献 [29] 发现基于两阶矩匹配的

卡方近似方法对逼近 L^2 范数类型统计量的零分布有很好的自适应性. 该方法不仅适用于变量之间弱相关的高维数据, 而且对变量之间强相关的高维数据也非常有效. 根据文献 [29] 得到的结果, 我们知道基于两阶矩匹配的卡方近似方法是不需要假设 (A4) 的, 而相应的正态近似方法则依赖于假设 (A4) 或者类似的高维稀疏性假设. 关于两阶矩匹配卡方近似方法和正态近似方法在本文所关注的高维双因素方差分析问题中的具体表现, 将在后面的模拟实验部分进行比较.

2.4 未知参数的估计

为了使用前面提出的两阶矩匹配卡方近似方法或者 (2.29) 中的正态近似方法, 需要利用观测到的样本给出未知参数 β 和 d 的相合估计, 进而得到近似临界值 $\beta\chi_d^2(\alpha)$ 的一个估计. 为此, 只需给出参数 $\mathrm{tr}(\boldsymbol{\Sigma})$、$\mathrm{tr}^2(\boldsymbol{\Sigma})$、$\mathrm{tr}(\boldsymbol{\Sigma}^2)$ 和 κ_{ij} $(i=1,\ldots,a,\ j=1,\ldots,b)$ 的相合估计.

在一些正则条件下, 文献 [28] 证明对 $i=1,\ldots,a,\ j=1,\ldots,b$, 基于第 (i,j) 组的样本, $\mathrm{tr}(\boldsymbol{\Sigma})$、$\mathrm{tr}^2(\boldsymbol{\Sigma})$、$\mathrm{tr}(\boldsymbol{\Sigma}^2)$ 和 κ_{ij} 的无偏且比率相合的估计量分别为 $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij})$ 和

$$\begin{aligned}
\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}_{ij} &= \frac{n_{ij}-1}{n_{ij}(n_{ij}-2)(n_{ij}-3)}[2\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij}^2)+(n_{ij}^2-3n_{ij}+1)\,\mathrm{tr}^2(\hat{\boldsymbol{\Sigma}}_{ij})-n_{ij}Q_{ij}],\\
\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}_{ij} &= \frac{n_{ij}-1}{n_{ij}(n_{ij}-2)(n_{ij}-3)}[(n_{ij}-1)(n_{ij}-2)\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij}^2)+\mathrm{tr}^2(\hat{\boldsymbol{\Sigma}}_{ij})-n_{ij}Q_{ij}],\\
\hat{\kappa}_{ij} &= \frac{-1}{(n_{ij}-2)(n_{ij}-3)}[2(n_{ij}-1)^2\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij}^2)+(n_{ij}-1)^2\,\mathrm{tr}^2(\hat{\boldsymbol{\Sigma}}_{ij})-n_{ij}(n_{ij}+1)Q_{ij}],
\end{aligned}\tag{2.30}$$

其中 $Q_{ij}=\sum_{k=1}^{n_{ij}}\|\boldsymbol{y}_{ijk}-\bar{\boldsymbol{y}}_{ij}\|^4/(n_{ij}-1)$, $i=1,\ldots,a$, $j=1,\ldots,b$. 所以, 基于全部的 ab 组样本 (2.1), $\mathrm{tr}(\boldsymbol{\Sigma})$、$\mathrm{tr}^2(\boldsymbol{\Sigma})$ 和 $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ 的无偏且比率相合的估计量分别为

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}=(n-ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij}),\tag{2.31}$$

$$\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}=(n-ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}_{ij},\tag{2.32}$$

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}=(n-ab)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}_{ij}.\tag{2.33}$$

把前面给出的这些比率相合的估计量代入 (2.26), 自然得到 β 和 d 的估计量

$$\hat{\beta}=\frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}+\sum_{i=1}^{a}\sum_{j=1}^{b}a_{0,(i,j)}^2\frac{\hat{\kappa}_{ij}}{2qn_{ij}}}{\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}},\quad \hat{d}=\frac{\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}}{\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}+\sum_{i=1}^{a}\sum_{j=1}^{b}a_{0,(i,j)}^2\frac{\hat{\kappa}_{ij}}{2qn_{ij}}}.\tag{2.34}$$

对于任何显著性水平 $\alpha>0$, 令 $\chi_d^2(\alpha)$ 表示 χ_d^2 分布的上 α 分位数. 注意到尽管 $\hat{\beta}$ 和 $\hat{d}$ 分别是 β 和 d 的有偏估计, 它们却可以是比率相合的, 正如下面的定理所示.

定理 2.3 令 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}$、$\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}$、$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ 和 $\hat{\kappa}_{ij}$ 分别为 $\mathrm{tr}(\boldsymbol{\Sigma})$、$\mathrm{tr}^2(\boldsymbol{\Sigma})$、$\mathrm{tr}(\boldsymbol{\Sigma}^2)$ 和 κ_{ij} $(i=1,\ldots,a,\ j=1,\ldots,b)$ 的比率相合估计, 随着 $n,p\to\infty$, 对于定义于 (2.34) 的 $\hat{\beta}$ 和 $\hat{d}$, 有 $\hat{\beta}/\beta\xrightarrow{P}1$, $\hat{d}/d\xrightarrow{P}1$, 且 $[\hat{\beta}\chi_{\hat{d}}^2(\alpha)]/[\beta\chi_d^2(\alpha)]\xrightarrow{P}1$.

使用 (2.34) 中的估计量, 对双因素 MANOVA 问题 (2.3), 就可以通过近似临界值 $\hat{\beta}\chi_{\hat{d}}^2(\alpha)$ 或近似 P 值 $\Pr(\chi_{\hat{d}}^2\geqslant T_n/\hat{\beta})$ 得到基于 L^2 范数统计量和两阶矩匹配卡方近似的检验结果.

当近似自由度 d 很大时, 也可以使用定理 2.2 中的正态近似. 使用 (2.31) 给出的估计量 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}$, 以及 (2.30) 和 (2.33) 给出的估计量 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$, 得到以下推论.

推论 2.3 假设定理 2.2 或推论 2.2 的条件成立, 则随着 $n,p\to\infty$, 有

$$\frac{T_{n0}-q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)\quad\text{和}\quad\frac{\beta\chi^2_d-q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1). \tag{2.35}$$

对于任意 $\alpha>0$, 令 z_α 和 $\Phi(\cdot)$ 分别表示标准正态分布 $\mathcal{N}(0,1)$ 的上 α 分位数和累积分布函数. 利用推论 2.3, 当 d 很大时, 可以使用近似临界值 $q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}+[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}z_\alpha$ 或近似 P 值 $1-\Phi([T_n-q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}]/[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2})$ 给出基于 L^2 范数统计量和正态近似的检验结果.

3 统计量的一些不变性质

本节说明我们所提出的解决双因素 MANOVA 模型背景下的 GLHT 问题 (2.12) 的检验方法具有一些很好的不变性质. 下面的定理总结了这些不变性质.

定理 3.1 (1) 定义于 (2.16) 的检验统计量 T_n 关于 GLHT 问题 (2.12) 非奇异线性变换不变. 具体地, 对于任意的 $(qp)\times(qp)$ 维非奇异矩阵 $\boldsymbol{B}$, 若 (2.12) 中的常数阵 $\boldsymbol{C}$ 被替换为 $\boldsymbol{BC}$, 则 T_n 的值不变.

(2) (2.34) 中的参数估计量 $\hat\beta$ 和 $\hat d$ 在下述样本的变换下保持不变:

$$\boldsymbol{y}_{ijk}\to\boldsymbol{R}\boldsymbol{y}_{ijk}+\boldsymbol{\xi},\quad k=1,\dots,n_{ij},\quad i=1,\dots,a,\quad j=1,\dots,b,$$

其中 $\boldsymbol{R}$ 是任何 $p\times p$ 维的正交矩阵, $\boldsymbol{\xi}$ 是任何 $p\times1$ 维的向量.

(3) 统计量 T_n、参数估计量 $\hat\beta$ 和 $\hat d$ 是关于因素的各个水平标号次序 $i=1,\dots,a$ 或 $j=1,\dots,b$ 置换不变的.

定理 3.1(1) 意味着我们提出的检验方法不依赖于所选择的设计对比矩阵, 也就是说 (2.4) 中的矩阵 $\boldsymbol{H}_a$、$\boldsymbol{H}_b$ 和 $\boldsymbol{H}_{ab}$ 可以被任何定义了相同假设的等价设计矩阵所取代. 定理 3.1(2) 表明参数估计量 $\hat\beta$ 和 $\hat d$ 与参数 β 和 d 具有相同的不变性质, 即它们都是关于样本的正交 - 平移变换不变量.

4 检验的近似和渐近功效函数

本节研究检验统计量 T_n 的功效函数. 令 $S_n=(\boldsymbol{Cx})^\top(\boldsymbol{CDC}^\top)^{-1}(\boldsymbol{C\mu})$. 根据 (2.18), 有

$$T_n\overset{d}{=}T_{n0}+2S_n+(\boldsymbol{C\mu})^\top(\boldsymbol{CDC}^\top)^{-1}(\boldsymbol{C\mu}), \tag{4.1}$$

其中注意到 T_{n0} 和零假设 H_0 下的 T_n 具有相同的分布. 经过一些简单的计算, 可以得到

$$\mathrm{E}(S_n)=0,\quad \mathrm{Var}(S_n)=\boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}(\boldsymbol{A}_0\otimes\boldsymbol{\Sigma})\boldsymbol{D}^{-1/2}\boldsymbol{\mu}. \tag{4.2}$$

为简单起见, 在以下条件成立时考虑 T_n 的功效函数:

$$\text{随着 } n,p\to\infty,\quad \boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}(\boldsymbol{A}_0\otimes\boldsymbol{\Sigma})\boldsymbol{D}^{-1/2}\boldsymbol{\mu}=o[\mathrm{tr}(\boldsymbol{\Sigma}^2)]. \tag{4.3}$$

这个条件意味着 S_n 代表的信息相对于 T_{n0} 所包含的信息可以被忽略. 在条件 (4.3) 下, 有 $S_n/\mathrm{Var}(T_{n0})\xrightarrow{P}0$.

随着 $p\to\infty$, 当 d^* 趋向于一个有限数时, 首先考虑检验的近似功效函数. 在这种情况下, 我们所提出的基于 L^2 范数的检验是通过使用两阶矩匹配卡方近似的近似临界值 $\hat{\beta}\chi^2_{\hat{d}}(\alpha)$ 来给出检验结果的. T_n 的相应功效函数可以表示为 $\Pr[T_n/\hat{\beta}\geqslant\chi^2_{\hat{d}}(\alpha)]$. 有以下定理.

定理 4.1 假设随着 $p\to\infty$, d^* 趋于一个有限数. 在假设 (A1) 和定理 2.3 成立的条件下, 随着 $n,p\to\infty$, T_n 的近似功效函数为

$$\Pr\left[\frac{T_n}{\hat{\beta}}\geqslant\chi^2_{\hat{d}}(\alpha)\right]\approx\Pr\left[\chi^2_d\geqslant\chi^2_d(\alpha)-\frac{n}{\beta}\Delta\right],\tag{4.4}$$

其中 $\Delta=(\boldsymbol{C\mu})^\top(\boldsymbol{CTC}^\top)^{-1}(\boldsymbol{C\mu})$, 矩阵 $\boldsymbol{T}=\mathrm{diag}(\tau_{11}^{-1},\ldots,\tau_{1b}^{-1},\ldots,\tau_{a1}^{-1},\ldots,\tau_{ab}^{-1})$.

现在考虑当 $p\to\infty$ 和 $d^*\to\infty$ 时 T_n 的功效函数. 在这种情况下, 我们所提出的基于 L^2 范数的检验可以通过正态近似的近似临界值 $q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}+[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}z_\alpha$ 给出检验结果. T_n 的相应功效函数可以表示为 $\Pr([T_n-q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}]/[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}\geqslant z_\alpha)$. 有以下定理.

定理 4.2 在假设 A 成立的条件下, 随着 $n,p\to\infty$, T_n 的渐近功效函数为

$$\Pr\left(\frac{T_n-q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}]^{1/2}}\geqslant z_\alpha\right)=\Phi\left(-z_\alpha+\frac{n\Delta}{\sqrt{2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}}\right)+o(1).\tag{4.5}$$

5 模拟实验

本节使用模拟实验来检查新提出的检验方法的性能. 对于给定的组样本数 $\boldsymbol{n}=[n_{11},n_{12},\ldots,n_{ab}]$ 和协方差矩阵 $\boldsymbol{\Sigma}$, 首先基于以下模型生成 ab 组多元随机样本:

$$\boldsymbol{y}_{ijk}=\boldsymbol{\mu}_{ij}+\boldsymbol{\Sigma}_{ij}^{1/2}\boldsymbol{z}_{ijk},\quad k=1,\ldots,n_{ij},\tag{5.1}$$

其中随机向量 $\boldsymbol{z}_{ijk}$ ($k=1,\ldots,n_{ij}$, $i=1,\ldots,a$, $j=1,\ldots,b$) 分别为下列设定生成的独立同分布样本:

M1: 对 $t=1,\ldots,p$, $z_{ijkt}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$, 其中 z_{ijkt} 是 $\boldsymbol{z}_{ijk}$ 的第 t 个元素.

M2: 对 $t=1,\ldots,p$, $z_{ijkt}=\sqrt{\frac{1}{2}}w_{ijkt}$, 其中 $w_{ijkt}\overset{\text{i.i.d.}}{\sim}t(4)$ (自由度为 4 的 t 分布).

M3: 对 $t=1,\ldots,p$, $z_{ijkt}=\sqrt{\frac{1}{2}}(w_{ijkt}-1)$, 其中 $w_{ijkt}\overset{\text{i.i.d.}}{\sim}\chi^2_1$. 令组均值向量 $\boldsymbol{\mu}_{ij}=\boldsymbol{\mu}_{11}+ij\delta\boldsymbol{h}/(ab)$, $i=1,\ldots,a$, $j=1,\ldots,b$, 其中 $\boldsymbol{\mu}_{11}=\boldsymbol{0}$ 为第一组 (因素 A 的第一个水平) 的均值向量, $\boldsymbol{h}=\boldsymbol{u}/\|\boldsymbol{u}\|$ 为一个指定组均值差异方向的常量单位向量, $\boldsymbol{u}=[1,\ldots,p]^\top$, δ 为控制组均值差异程度的可调整参数. 对于一些特别挑选的 δ 值, 使用模拟的样本计算出检验方法的统计量值和相应的 P 值. 其中, 在每种模拟设置下, 除了选取 $\delta=0$, 还选取两个非零的 δ 值并尽量使得功效最大的那个检验方法的经验功效分别达到 50% 和 90%. 当 P 值小于显著性水平 $\alpha=5\%$ 时, 拒绝原假设. 根据计算的 P 值可以得到 $N=10{,}000$ 次重复实验下零假设被拒绝的比例, 该比例即检验的经验显著性水平 (当 $\delta=0$ 时) 或经验功效 (当 $\delta>0$ 时). 本文仅展示检验交互效应的模拟结果, 因为检验主效应的结果是非常相似的.

在模拟 1 中, 主要研究维数效应. 所考虑的检验包括经典的 WLR 检验、LHT 检验和 BNP 迹检验, 以及我们的方法. 对于所有的经典检验, 都使用 F 分布近似逼近其零分布. 对于我们的检验, 同时用卡方近似和正态近似方法逼近零分布, 并分别记这两种近似方法对应的检验为 Tchi 和 Tnm. 令协方差矩阵 $\boldsymbol{\Sigma}=\mathrm{diag}(\boldsymbol{d}/\|\boldsymbol{d}\|)$, 即对角元素为 $\boldsymbol{d}/\|\boldsymbol{d}\|$ 的对角矩阵, 其中 $\boldsymbol{d}=[1,2^3,\ldots,p^3]$, $\|\boldsymbol{d}\|=\sqrt{1^6+2^6+\cdots+p^6}$. 设置两个因素 A 和 B 分别有 $a=3$ 和 $b=2$ 个水平. 为了简单起见, 令组样本数仅对于因素 A 的 a 个水平是不同的, 但对因素 B 的 b 个水平是相同的, 即对 $i=1,\ldots,a$, $j=1,\ldots,b$,

有 $n_{ij} = n_{i1}$. 上述组样本数设定对我们的结论没有实质影响. 考虑组样本数 $\boldsymbol{n} = [n_{11}, n_{21}, n_{31}] = [10, 12, 15]$ 及对应的两个不同维数 $p = 25$ 和 $p = 65$, 以及 $\boldsymbol{n} = [50, 60, 70]$ 及对应的维数 $p = 150$ 和 $p = 350$. 根据 (2.11), 检验交互效应不依赖于权重设置方式, 所以, 直接采用等权重方法为各组加权. 注意除我们的检验 Tchi 和 Tnm 之外的其他检验方法都是在低维数据场景中提出的, 并且要求合并样本协方差矩阵是非奇异的. 因此, 对于模拟 1 中的所有情况, 数据维数都小于样本的自由度, 以确保待比较的经典检验方法可以使用.

模拟 1 中各个检验的经验水平和功效在表 1 中给出. 可以看出, 当数据维数相对较小时 ($\boldsymbol{n} = [10, 12, 15]$ 且 $p = 25$, 以及 $\boldsymbol{n} = [50, 60, 70]$ 且 $p = 150$), 经典 WLR、LHT 和 BNP 迹检验比我们的检验 Tchi 和 Tnm 在功效上要强很多. 然而, 当数据维数相对很大时 ($\boldsymbol{n} = [10, 12, 15]$ 且 $p = 65$, 以及 $\boldsymbol{n} = [50, 60, 70]$ 且 $p = 350$), 我们的检验 Tchi 和 Tnm 比经典 WLR、LHT 和 BNP 迹检验的功效明显强很多. 数据维数相对很大时, 经典的检验功效较低的原因是在这种情况下数据的维数接近于样本自由度, WLR、LHT 和 BNP 迹检验中使用的合并样本协方差阵是接近奇异的. 还能看到检验方法 Tchi 可以比 Tnm 更好地控制第一类错误, 这意味着在该模拟设置下卡方近似方法比正态近似方法更为精确. 此外, 在三个经典检验中, LHT 检验是功效最强的, BNP 迹检验则是功效相对较弱的, 这是因为模拟设置中的协方差矩阵具有非常不一样的特征值, 这也与经典的结论一致 (参见文献 [5, 第 8.6.5 小节]). 该模拟表明, 经典检验的功效会受到数据维度的极大影响, 并且当维数 (小于但) 与总样本量相当时可能表现不佳. 而我们的检验对维数效应是稳健的, 并且当维数适中但比总样本量稍小时, 要优于经典检验方法.

在模拟 2 中, 着重研究高维数据背景下新检验方法的表现. 与模拟 1 一样, 我们同时考虑基于卡方近似的检验 Tchi 和基于正态近似的检验 Tnm. 还把这两个方法与文献 [19,35] 中各自提出的高维一般线性假设检验方法作比较. 这两个用来作比较的方法分别记作 T_{SF} 和 T_{SK}. 注意到与 Tnm 一样, T_{SF} 和 T_{SK} 也是基于正态近似来逼近统计量的零分布. 令两个因素的水平分别为 $a = 3$ 个和 $b = 5$ 个. 与模拟 1 类似, 设置组样本数仅对于因素 A 的 a 个水平是不同的, 并记 $\boldsymbol{n} = [n_{11}, n_{21}, n_{31}]$. 设置协方差阵为 $\boldsymbol{\Sigma} = \sigma^2[(1-\rho)\boldsymbol{I}_p + \rho\boldsymbol{J}_p]$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 其中 $\boldsymbol{J}_p$ 表示一个 $p \times p$ 维的元素全为 1 的矩阵. 该形式的协方差阵依赖于参数 σ 和 ρ, 它们分别控制着数据的整体波动程度和相关程度. 令 $\sigma^2 = 2$,

表 1　模拟 1 中各检验的经验水平 ($\boldsymbol{\delta = 0}$) 和功效 ($\boldsymbol{\delta > 0}$) (%)

$\boldsymbol{n}$	p	δ	WLR	LHT	BNP	Tchi	Tnm
$[10, 12, 15]^\top$	25	0.0	5.28	5.60	4.98	5.38	6.70
		2.0	53.05	57.51	47.03	16.44	18.99
		3.2	96.90	98.31	91.52	43.93	48.24
	65	0.0	4.72	5.21	4.67	5.33	6.35
		3.2	23.62	21.72	22.05	47.79	50.90
		5.0	47.19	49.35	32.73	95.37	96.06
$[50, 60, 70]^\top$	150	0.0	4.62	4.67	4.47	4.80	5.58
		0.6	34.93	36.40	33.73	8.68	9.61
		1.0	92.70	94.75	89.55	20.51	22.16
	350	0.0	5.42	6.17	5.28	4.98	5.34
		1.5	19.88	19.42	19.70	54.07	55.92
		2.1	34.22	36.48	27.99	94.69	95.05

考虑不同的数据相关程度 $\rho=0.1$、$\rho=0.5$ 和 $\rho=0.9$, 以及样本量较小的情况 $\boldsymbol{n}=\boldsymbol{n}_1=[30,40,50]$、样本量适中的情况 $\boldsymbol{n}=\boldsymbol{n}_2=[60,80,100]$ 和样本量较大的情况 $\boldsymbol{n}=\boldsymbol{n}_3=[90,120,150]$, 并分别考虑不同的数据维数 $p=50$、$p=500$ 和 $p=1{,}000$, 以便于研究维数较低, 适中及较高的情况. 其余的相关设置与模拟 1 一样.

模拟 2 中各个检验的经验水平和功效分别由表 2 和 3 给出. 可以看出, 当 $\rho=0.1$ 时, T_{SK} 非常有效, 但当 $\rho=0.5$ 和 $\rho=0.9$ 时, 它变得相当保守. 当数据相关性较小时, T_{SK} 具有比其他检验更好的第一类错误控制和与其他方法相当的检验功效. 但当数据相关性为中等或较高时, T_{SK} 的经验水平接近 0 且只有非常低的检验功效. 在所有情况下, T_{SF} 和 Tnm 的表现相似. 当 $\rho=0.1$ 时, Tchi 具有与 T_{SF} 和 Tnm 相似的功效, 但有着稍好的第一类错误控制. 当 $\rho=0.5$ 和 $\rho=0.9$ 时, Tchi 对第一类错误的控制明显优于 T_{SF} 和 Tnm. 模拟结果中 Tnm 的经验功效都比 Tchi 的要大一些, 但注意到 Tnm

表 2 模拟 2 中各检验的经验水平 (%)

			$\rho=0.1$				$\rho=0.5$				$\rho=0.9$			
模型	p	$\boldsymbol{n}$	T_{SF}	T_{SK}	Tchi	Tnm	T_{SF}	T_{SK}	Tchi	Tnm	T_{SF}	T_{SK}	Tchi	Tnm
M1	50	$\boldsymbol{n}_1$	5.37	4.14	5.09	5.44	7.13	1.54	6.05	7.30	7.40	0.19	5.58	7.48
		$\boldsymbol{n}_2$	6.02	4.56	5.61	6.06	7.06	1.42	5.86	7.18	6.97	0.19	5.40	7.06
		$\boldsymbol{n}_3$	5.46	4.17	5.13	5.54	6.82	1.41	5.67	6.81	6.89	0.10	5.00	6.94
	500	$\boldsymbol{n}_1$	7.14	5.16	6.94	7.19	7.06	0.22	6.11	7.14	7.29	0.01	5.69	7.55
		$\boldsymbol{n}_2$	6.61	4.92	6.42	6.67	6.47	0.21	5.40	6.56	7.17	0.00	5.68	7.26
		$\boldsymbol{n}_3$	6.24	4.64	5.98	6.29	7.19	0.18	6.19	7.24	6.87	0.00	5.16	6.87
	1,000	$\boldsymbol{n}_1$	6.59	4.68	6.44	6.72	6.63	0.08	5.62	6.77	7.25	0.00	5.73	7.36
		$\boldsymbol{n}_2$	6.90	4.75	6.70	6.96	6.62	0.05	5.75	6.69	6.52	0.00	4.97	6.59
		$\boldsymbol{n}_3$	6.66	4.70	6.51	6.72	7.36	0.05	6.18	7.41	6.83	0.00	5.13	6.91
M2	50	$\boldsymbol{n}_1$	5.64	4.30	4.98	5.40	6.99	1.29	5.82	7.07	6.88	0.14	5.26	6.96
		$\boldsymbol{n}_2$	5.84	4.32	5.23	5.74	6.81	1.36	5.76	6.89	6.74	0.16	5.27	6.92
		$\boldsymbol{n}_3$	5.83	4.39	5.19	5.73	6.72	1.32	5.70	6.73	6.64	0.15	5.08	6.66
	500	$\boldsymbol{n}_1$	6.65	5.07	6.39	6.64	7.10	0.18	6.04	7.18	7.13	0.00	5.36	7.39
		$\boldsymbol{n}_2$	6.83	4.92	6.49	6.85	6.53	0.15	5.59	6.58	6.90	0.00	4.96	6.98
		$\boldsymbol{n}_3$	6.53	4.94	6.33	6.49	6.51	0.23	5.60	6.57	6.83	0.00	5.39	6.83
	1,000	$\boldsymbol{n}_1$	6.74	4.67	6.59	6.79	7.08	0.09	6.04	7.24	6.84	0.00	5.23	6.96
		$\boldsymbol{n}_2$	6.67	4.91	6.46	6.68	6.98	0.08	5.96	7.02	7.32	0.00	5.59	7.37
		$\boldsymbol{n}_3$	6.36	4.39	6.17	6.44	6.99	0.05	6.00	7.04	6.66	0.00	5.18	6.72
M3	50	$\boldsymbol{n}_1$	5.67	4.39	4.84	5.33	6.92	1.49	5.81	6.93	7.01	0.20	5.29	7.18
		$\boldsymbol{n}_2$	6.05	4.58	5.34	5.80	6.70	1.34	5.68	6.71	6.73	0.14	4.98	6.78
		$\boldsymbol{n}_3$	5.83	4.38	5.32	5.73	6.46	1.49	5.44	6.50	6.82	0.13	5.22	6.82
	500	$\boldsymbol{n}_1$	6.88	5.17	6.60	6.85	6.65	0.23	5.59	6.73	6.94	0.01	5.37	7.11
		$\boldsymbol{n}_2$	6.67	5.03	6.35	6.68	7.13	0.16	6.09	7.17	6.53	0.00	4.96	6.59
		$\boldsymbol{n}_3$	6.82	5.18	6.53	6.81	7.07	0.13	5.98	7.11	6.58	0.00	5.09	6.57
	1,000	$\boldsymbol{n}_1$	6.70	4.58	6.51	6.78	6.88	0.14	5.96	7.06	6.99	0.00	5.52	7.16
		$\boldsymbol{n}_2$	6.74	4.70	6.53	6.79	6.84	0.09	5.78	6.93	6.17	0.00	4.93	6.28
		$\boldsymbol{n}_3$	6.77	4.66	6.54	6.76	6.97	0.08	5.99	7.00	6.37	0.00	4.88	6.43

表 3　模拟 2 中各检验的经验功效 (%)

				$\rho=0.1$				$\rho=0.5$				$\rho=0.9$			
模型	p	$\boldsymbol{n}$	δ	T_{SF}	T_{SK}	Tchi	Tnm	T_{SF}	T_{SK}	Tchi	Tnm	T_{SF}	T_{SK}	Tchi	Tnm
M1	50	$\boldsymbol{n}_1$	10	97.20	96.33	97.03	97.27	44.84	21.27	42.04	45.24	26.19	2.28	22.25	26.55
		$\boldsymbol{n}_2$	7	96.91	96.07	96.74	96.94	43.72	20.86	40.57	43.96	24.54	2.04	20.69	24.78
		$\boldsymbol{n}_3$	6	98.16	97.57	98.01	98.17	48.31	23.46	44.78	48.48	28.29	2.37	24.03	28.44
	500	$\boldsymbol{n}_1$	28	99.01	98.44	98.95	99.02	35.92	4.05	33.38	36.22	21.03	0.04	18.01	21.25
		$\boldsymbol{n}_2$	19	98.24	97.43	98.20	98.25	33.42	3.43	30.47	33.49	20.24	0.03	16.58	20.35
		$\boldsymbol{n}_3$	16	98.94	98.29	98.86	98.95	35.37	3.86	32.63	35.46	21.02	0.06	17.54	21.25
	1,000	$\boldsymbol{n}_1$	36	97.50	96.10	97.41	97.51	30.34	1.45	28.06	30.64	18.94	0.00	15.59	19.21
		$\boldsymbol{n}_2$	26	97.92	96.65	97.78	97.93	31.44	1.57	28.52	31.52	18.51	0.02	15.37	18.64
		$\boldsymbol{n}_3$	21	97.68	96.19	97.56	97.69	30.79	1.43	27.98	30.92	19.13	0.00	15.68	19.19
M2	50	$\boldsymbol{n}_1$	10	96.99	96.57	96.56	96.74	44.77	21.20	41.81	45.09	26.03	2.24	22.00	26.54
		$\boldsymbol{n}_2$	7	96.66	95.97	96.22	96.48	43.75	20.86	40.53	43.95	25.15	1.92	21.23	25.24
		$\boldsymbol{n}_3$	6	98.34	98.06	98.15	98.30	47.94	23.03	44.67	48.03	28.30	2.26	24.18	28.37
	500	$\boldsymbol{n}_1$	28	99.01	98.64	98.94	99.00	36.35	4.37	33.47	36.67	21.50	0.05	18.03	21.70
		$\boldsymbol{n}_2$	19	98.26	97.56	98.19	98.27	33.81	3.32	31.05	33.91	19.40	0.02	15.79	19.53
		$\boldsymbol{n}_3$	16	98.91	98.51	98.85	98.91	34.71	3.70	31.52	34.88	20.23	0.04	17.08	20.42
	1,000	$\boldsymbol{n}_1$	36	97.39	95.98	97.30	97.39	29.86	1.50	27.39	30.20	18.08	0.00	14.90	18.44
		$\boldsymbol{n}_2$	26	97.97	96.80	97.85	97.96	31.19	1.46	28.77	31.36	19.28	0.02	15.55	19.42
		$\boldsymbol{n}_3$	21	97.78	96.47	97.69	97.79	30.70	1.32	27.86	30.87	18.59	0.02	15.14	18.75
M3	50	$\boldsymbol{n}_1$	10	97.09	96.52	96.72	97.01	44.61	21.37	41.56	45.18	25.28	2.01	21.46	25.82
		$\boldsymbol{n}_2$	7	96.70	95.88	96.39	96.69	44.01	20.36	40.97	44.35	26.20	1.73	22.19	26.39
		$\boldsymbol{n}_3$	6	98.42	97.88	98.16	98.39	49.13	23.79	45.96	49.33	28.27	2.22	23.67	28.43
	500	$\boldsymbol{n}_1$	28	99.04	98.73	98.98	99.04	36.51	4.27	33.72	36.88	20.71	0.02	17.25	21.02
		$\boldsymbol{n}_2$	19	98.30	97.45	98.21	98.31	33.15	3.31	30.43	33.28	20.18	0.05	16.50	20.38
		$\boldsymbol{n}_3$	16	98.94	98.49	98.87	98.93	34.80	3.48	31.72	34.77	21.20	0.05	17.28	21.38
	1,000	$\boldsymbol{n}_1$	36	97.54	96.03	97.42	97.53	29.63	1.54	27.10	30.11	18.28	0.00	15.15	18.63
		$\boldsymbol{n}_2$	26	98.02	96.92	97.88	98.04	31.57	1.64	29.12	31.63	19.31	0.01	15.81	19.39
		$\boldsymbol{n}_3$	21	98.11	96.79	97.97	98.12	30.57	1.47	27.71	30.62	18.17	0.00	14.84	18.34

的经验水平也都比 Tchi 的要大一些, 所以, 该结果很可能是 Tnm 对第一类错误控制不足导致的. 数值仿真模拟结果表明, 在两种逼近检验统计量 T_n 零分布的方法中, 卡方近似更为准确, 尤其是在数据相关性较高时. 因此在实际数据分析中, 建议使用基于 L^2 范数统计量和两阶矩匹配卡方近似的检验方法 Tchi.

6　实际数据应用

6.1　矫正器数据

在这个实际问题研究中, 将检验方法 Tnm 和 Tchi, 以及 T_{SF} 和 T_{SK} 应用到第 1 节中介绍的矫

正器数据, 来检验交互作用、实验条件和个体因素对膝盖处的平均力矩的影响. 对于这个数据, 因为各组样本数相同, 可以直接使用第 2.1 小节中提到的等权重方法来设置各组权重.

检验结果由表 4 给出. 在表 4 中, "统计量" 一列所给出的使用正态近似的检验 Tnm 的统计量值同时也是中心且标准化后的 L^2 范数统计量的值 $(T_n/\hat{\beta}-\hat{d})/(2\hat{d})^{1/2}$. 从表 4 可以看出, 交互作用、实验条件和个体因素对组平均值具有显著影响, 因为相应的 P 值都非常小. 表 4 在后两行附带也给出了两种事后对比检验的结果. 正如第 2.2 小节中提到过的, 这两行的检验结果是通过选取相应的对比矩阵得到的. 各检验都认为两种事后对比中因素效应的差异是显著的. 从自由度的估计值 $\hat{d}$ 可以看出, 除了在检验交互和个体效应时 $\hat{d}$ 值较大, 在其他几项检验中 $\hat{d}$ 值都很小.

6.2 乳腺癌数据

在本应用中, 研究由实验特征导致的微阵列数据集的异质性对数据质量的影响. 该数据集首先由文献 [36] 发布, 并被文献 [37] 用来研究系统效应的修正方法. 我们使用由 J. S. Marron 教授热情提供的预处理后的数据集, 该数据集也可以在 https://genome.unc.edu/pubsup/dwd/# 下载. 预处理的数据集已消除了缺失值, 并过滤到只包含 5,961 个基因. 该数据集被文献 [36] 标注了 5 个癌症类别标签 (分别是 luminal A、luminal B、normal-like、ERBB2 和 basal), 并且具有两个已知的可以引起系统性偏差的异质性标记. 一个异质性标记是 RNA 的不同来源 (源效应), 另一个是微阵列的不同实验批次 (批次效应). 如果考虑数据中某种癌症类别的源效应和批次效应对微阵列实验中基因表达水平的影响, 则该数据可以在高维双因素 MANOVA 问题背景下进行研究. 由于在此背景下一些癌症类别的样本量相当小, 只考虑癌症类别 luminal A 且具有源标签 $i=1,2$ 和批次标签 $j=2,3$ 的部分数据. 相应的组样本大小分别为 $n_{12}=9$、$n_{13}=11$、$n_{22}=5$ 和 $n_{23}=4$.

分别使用等权重方法和样本量适应权重方法来为各组加权. 检验结果见表 5, 其中两种加权方法

表 4 矫正器数据中实验条件效应和个体效应的检验结果

	统计量				P 值				参数估计值 (Tchi)	
效应	T_{SF}	T_{SK}	Tnm	Tchi	T_{SF}	T_{SK}	Tnm	Tchi	$\hat{\beta}$	$\hat{d}$
交互	49.556	29.322	38.610	2,234,379.141	0.000	0.000	0.000	0.000	3,874.821	82.048
实验	218.446	133.025	170.168	3,456,760.448	0.000	0.000	0.000	0.000	3,775.451	14.035
个体	480.689	273.808	375.881	10,766,713.643	0.000	0.000	0.000	0.000	3,795.325	27.922
对照 + 矫正器 vs. 弹簧 1 + 2	368.311	223.362	286.590	3,321,499.254	0.000	0.000	0.000	0.000	3,762.202	4.695
对照 vs. 矫正器	9.831	6.105	7.498	104,064.902	0.000	0.000	0.000	0.000	3,758.788	4.699

表 5 乳腺癌数据中被标记为类别 luminal A 的部分数据的实验效应的检验结果

		统计量				P 值				参数估计值 (Tchi)	
权重	效应	T_{SF}	T_{SK}	Tnm	Tchi	T_{SF}	T_{SK}	Tnm	Tchi	$\hat{\beta}$	$\hat{d}$
等权	交互	2.015	1.764	1.606	3,241.404	0.022	0.039	0.054	0.063	45.289	54.764
	来源	16.431	10.407	14.068	9,147.760	0.000	0.000	0.000	0.000	45.289	54.764
	批次	8.410	4.665	7.134	5,861.256	0.000	0.000	0.000	0.000	45.289	54.764
适应	交互	2.015	1.764	1.606	3,241.404	0.022	0.039	0.054	0.063	45.289	54.764
	来源	16.584	10.494	14.005	9,210.388	0.000	0.000	0.000	0.000	46.557	53.272
	批次	10.565	5.991	9.782	6,744.459	0.000	0.000	0.000	0.000	38.310	64.740

给出了类似的结论. 从 P 值可以看出, 在 0.05 的显著性水平下, Tnm 和 Tchi 认为交互作用效应并不特别显著, 但 T_{SF} 和 T_{SK} 认为交互作用效应显著. 而在 0.1 的显著性水平下, 所有检验都认为交互作用效应显著. 此外, 所有的检验方法都表明来源或批次因素对组平均值有非常显著的影响. 检验结果从高维双因素 MANOVA 角度证实, 实验条件如 RNA 来源和实验批次可能对微阵列实验中基因表达水平造成较大的系统偏差.

我们还将各个检验应用到了经文献 [37] 提出的 "距离加权判别" (distance weighted discrimination, DWD) 系统偏差修正方法处理过的数据. 检验结果展示在表 6 中. 对于修正后的数据, 我们看到源效应和批次效应对应的 P 值都很大, 而交互效应的 P 值与原始数据结果相比几乎没有变化. 结果表明, DWD 修正方法在消除实验系统偏差的主效应方面非常有效, 但可能无法消除交互效应带来的偏差. 注意到, 对于这两个版本的数据和本小节所考虑的各项检验, 近似自由度的估计值都很大, 因此, 基于正态近似的 Tnm 和基于卡方近似的 Tchi 给出了较为接近的 P 值.

6.3 眼角膜数据

在这个实际应用中, 研究了来自眼科学的眼角膜表面数据集[38]. 原始数据集包含 256×27 个极坐标网格点上的角膜表面曲率测量值. 网格共有 256 个径向, 每个径向包含 27 个测量位置. 为了消除缺失值并减少测量误差, 在 Legendre-Fourier 回归模型下使用最小二乘拟合重建原始数据集并得到包含 100×20 个极坐标网格点上的拟合值的新数据集. 处理后的数据总样本大小 $n=135$, 维数 $p=2{,}000$. 该数据有两个因素, 其中因素 Ker 有 4 个水平, 因素 LR (left or right) 有 2 个水平. 因素 Ker 表示圆锥角膜 (keratoconus, 一种角膜畸变的疾病) 的程度, 其 4 个水平分别是正常 (normal, 缩写为 Nor)、单侧可疑 (unilateral suspect, 缩写为 Un)、双侧可疑 (suspect, 缩写为 Sus) 和临床确诊圆锥角膜 (clinical keratoconus, 缩写为 Cli). 因子 LR 表示样本是来自左眼或者右眼. 每个组的样本大小分别为 $n_{11}=22$, $n_{12}=21$, $n_{21}=7$, $n_{22}=7$, $n_{31}=11$, $n_{32}=10$, $n_{41}=37$, $n_{42}=35$.

表 7 给出了对数据的双因素 MANOVA 分析结果. 可以看出, 各检验方法都认为交互效应不显著, 且认为因素 Ker 对角膜表面曲率均值有显著影响. 除 T_{SK} 外的所有检验方法都认为因素 LR 对角膜表面曲率均值有显著影响. T_{SK} 对因素 LR 的效应判断略微保守, 并且在两种不同的加权重方法下都给出大于 0.05 的 P 值. 注意到近似自由度的估计值都较小, 所以方法 Tchi 给出的结果更为可信. 此外, 从 P 值的大小可知因素 Ker 比因素 LR 对角膜表面曲率均值的影响更显著. 可以从检验结果推断, 平均意义下, 不同程度的圆锥角膜具有非常不同的角膜表面曲率, 且左角膜表面曲率也不同于右角膜表面曲率.

表 6 修正后的乳腺癌数据中被标记为类别 luminal A 的部分数据的实验效应的检验结果

		统计量				P 值				参数估计值 (Tchi)	
权重	效应	T_{SF}	T_{SK}	Tnm	Tchi	T_{SF}	T_{SK}	Tnm	Tchi	$\hat{\beta}$	$\hat{d}$
等权	交互	2.015	1.764	1.606	3,241.531	0.022	0.039	0.054	0.063	45.289	54.764
	来源	0.867	0.785	0.614	2,771.018	0.193	0.216	0.270	0.256	45.289	54.764
	批次	−0.137	0.302	−0.255	2,359.555	0.555	0.381	0.600	0.577	45.289	54.764
适应	交互	2.015	1.764	1.606	3,241.531	0.022	0.039	0.054	0.063	45.289	54.764
	来源	1.060	0.903	0.770	2,850.242	0.145	0.183	0.221	0.212	46.557	53.272
	批次	0.029	0.421	−0.120	2,427.837	0.488	0.337	0.548	0.525	38.310	64.740

表 7 眼角膜数据中各因素效应的检验结果

		统计量				P 值				参数估计值 (Tchi)	
权重	效应	T_{SF}	T_{SK}	Tnm	Tchi	T_{SF}	T_{SK}	Tnm	Tchi	$\hat{\beta}$	$\hat{d}$
等权	交互	−0.022	−0.025	−0.026	43.691	0.509	0.510	0.511	0.448	4.8827	9.060
	Ker	26.637	7.354	24.851	560.742	0.000	0.000	0.000	0.000	4.8827	9.060
	LR	5.409	1.412	5.070	63.150	0.000	0.079	0.000	0.000	4.7721	2.318
适应	交互	−0.022	−0.025	−0.026	43.691	0.509	0.510	0.511	0.448	4.8827	9.060
	Ker	26.734	7.355	24.671	562.633	0.000	0.000	0.000	0.000	4.9906	8.864
	LR	5.658	1.614	5.601	65.549	0.000	0.053	0.000	0.001	4.2798	2.584

7 总结

本文提出并研究了基于 L^2 范数的高维双因素方差分析方法, 探讨了基于两阶矩匹配的卡方近似和正态近似两种用来近似检验统计量零分布的方法. 此外, 我们研究了所提出的检验方法的近似和渐近功效. 模拟实验和实际数据实例表明, 所提出的检验方法对高维数据有较好的表现. 本文主要在各组同协方差阵假设下研究了高维双因素方差分析问题. 如何将本文提出的方法推广到异协方差阵的情况和多因素方差分析, 以及如何通过改进统计量的形式进一步提高检验的功效, 都是未来值得继续研究的问题.

参考文献

1 Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis, 6th ed. Upper Saddle River: Pearson Prentice Hall, 2007
2 Xu J, Cui X. Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. Bioinformatics, 2008, 24: 1056–1062
3 Tsai C A, Chen J J. Multivariate analysis of variance test for gene set analysis. Bioinformatics, 2009, 25: 897–903
4 中国科学院数学研究所统计组. 方差分析. 北京: 科学出版社, 1977
5 Anderson T W. An Introduction to Multivariate Statistical Analysis, 3rd ed. New York: John Wiley & Sons, 2003
6 方开泰, 张尧庭. 广义多元分析. 北京: 科学出版社, 1993
7 Zhang J T. Two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. Technometrics, 2011, 53: 426–439
8 Harrar S W, Bathke A C. A modified two-factor multivariate analysis of variance: Asymptotics and small sample approximations. Ann Inst Statist Math, 2012, 64: 135–165
9 Xu L W. Parametric bootstrap approaches for two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. J Multivariate Anal, 2015, 133: 291–303
10 Zhang J T, Zhou B, Guo J, et al. A modified Bartlett test for heteroscedastic two-way MANOVA. J Adv Stat, 2016, 1: 94–108
11 Zhang J T. Analysis of Variance for Functional Data. Boca Raton: CRC Press, 2013
12 Dempster A P. A high dimensional two sample significance test. Ann Math Statist, 1958, 29: 995–1010
13 Dempster A P. A significance test for the separation of two highly multivariate small samples. Biometrics, 1960, 16: 41–50
14 Bai Z, Saranadasa H. Effect of high dimension: By an example of a two sample problem. Statist Sinica, 1996, 6: 311–329
15 Srivastava M S, Du M. A test for the mean vector with fewer observations than the dimension. J Multivariate Anal, 2008, 99: 386–402
16 Chen S X, Qin Y L. A two-sample test for high-dimensional data with applications to gene-set testing. Ann Statist,

2010, 38: 808–835
17 Fujikoshi Y, Himeno T, Wakaki H. Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size. J Japan Statist Soc, 2004, 34: 19–26
18 Schott J R. Some high-dimensional tests for a one-way MANOVA. J Multivariate Anal, 2007, 98: 1825–1839
19 Srivastava M S, Kubokawa T. Tests for multivariate analysis of variance in high dimension under non-normality. J Multivariate Anal, 2013, 115: 204–216
20 Zhang J T, Guo J, Zhou B. Linear hypothesis testing in high-dimensional one-way MANOVA. J Multivariate Anal, 2017, 155: 200–216
21 Zhang J T, Xu J F. On the k-sample Behrens-Fisher problem for high-dimensional data. Sci China Ser A, 2009, 52: 1285–1304
22 Bennett B M. Note on a soulution of the generalized Behrens-Fisher problem. Ann Inst Statist Math, 1950, 2: 87–90
23 Yamada T, Himeno T. Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. J Multivariate Anal, 2015, 139: 7–27
24 Hu J, Bai Z, Wang C, et al. On testing the equality of high dimensional mean vectors with unequal covariance matrices. Ann Inst Statist Math, 2017, 69: 365–387
25 Zhou B, Guo J, Zhang J T. High-dimensional general linear hypothesis testing under heteroscedasticity. J Statist Plann Inference, 2017, 188: 36–54
26 Fujikoshi Y. Two-way ANOVA models with unbalanced data. Discrete Math, 1993, 116: 315–334
27 张尧庭, 方开泰. 多元统计分析引论. 北京: 科学出版社, 1982
28 Himeno T, Yamada T. Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. J Multivariate Anal, 2014, 130: 27–44
29 Zhang J T, Guo J, Zhou B, et al. A simple two-sample test in high dimensions based on L^2-norm. J Amer Statist Assoc, 2019, doi: 10.1080/01621459.2019.1604366
30 Zhang J T. Approximate and asymptotic distributions of chi-squared-type mixtures with applications. J Amer Statist Assoc, 2005, 100: 273–285
31 Box G E P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. Ann Math Statist, 1954, 25: 290–302
32 Welch B L. The generalization of "Student's" problem when several different population variances are involved. Biometrika, 1947, 34: 28–35
33 Satterthwaite F E. An approximate distribution of estimates of variance components. Biometrics Bull, 1946, 2: 110–114
34 Zhang J T, Xiao S. A note on the modified two-way MANOVA tests. Statist Probab Lett, 2012, 82: 519–527
35 Srivastava M S, Fujikoshi Y. Multivariate analysis of variance with fewer observations than the dimension. J Multivariate Anal, 2006, 97: 1927–1940
36 Perou C M, Sørlie T, Eisen M B, et al. Molecular portraits of human breast tumours. Nature, 2000, 406: 747–752
37 Benito M, Parker J, Du Q, et al. Adjustment of systematic microarray data biases. Bioinformatics, 2004, 20: 105–114
38 Locantore N, Marron J S, Simpson D G, et al. Robust principal component analysis for functional data. Test, 1999, 8: 1–73
39 Li J, Chen S X. Two sample tests for high-dimensional covariance matrices. Ann Statist, 2012, 40: 908–940

附录 A 相关结论的证明

(2.9) 的证明 注意到 $w_{ij} = u_i v_j$, 所以, $w_{i\cdot} = \sum_{j=1}^{b} w_{ij} = u_i \sum_{j=1}^{b} v_j = u_i$, $w_{\cdot j} = \sum_{i=1}^{a} w_{ij} = (\sum_{i=1}^{a} u_i)v_j = v_j$, $i = 1, \ldots, a$, $j = 1, \ldots, b$. 由 (2.1) 和 (2.2) 可知, $\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}$. 约束 (2.6) 和 (2.8) 可推出 $\sum_{j=1}^{b} v_j \boldsymbol{\gamma}_{ij} = \mathbf{0}$, $i = 1, \ldots, a$, 约束 (2.7) 和 (2.8) 可推出 $\sum_{i=1}^{a} u_i \boldsymbol{\gamma}_{ij} = \mathbf{0}$, $j = 1, \ldots, b$. 再根据约束 (2.5)–(2.8), 进一步有

$$\sum_{i=1}^{a} u_i \boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\beta}_j + \sum_{i=1}^{a} u_i \boldsymbol{\gamma}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\beta}_j, \quad j = 1, \ldots, b,$$

$$\sum_{j=1}^{b} v_j\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_i + \sum_{j=1}^{b} v_j\boldsymbol{\gamma}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_i, \quad i = 1, \ldots, a,$$
$$\sum_{i=1}^{a}\sum_{j=1}^{b} u_i v_j \boldsymbol{\mu}_{ij} = \Big(\sum_{i=1}^{a}\sum_{j=1}^{b} u_i v_j\Big)\boldsymbol{\mu}_0 = \boldsymbol{\mu}_0.$$

从而,

$$\boldsymbol{\mu}_0 = \sum_{i=1}^{a}\sum_{j=1}^{b} u_i v_j \boldsymbol{\mu}_{ij},$$
$$\boldsymbol{\alpha}_i = \sum_{j=1}^{b} v_j\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_0, \quad \boldsymbol{\beta}_j = \sum_{i=1}^{a} u_i\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_0,$$
$$\boldsymbol{\gamma}_{ij} = \boldsymbol{\mu}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j - \boldsymbol{\mu}_0, \quad i = 1, \ldots, a, \quad j = 1, \ldots, b.$$

根据 $\boldsymbol{\mu}$、$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的定义, 将上述关系表示为矩阵形式即有 (2.9). □

定理 2.1 的证明 为记号简便起见, 在本节中, 令 $c = ab$, 通过变换 $(i-1)b + j$, 把双下标 (i, j) 转换成单下标 $(i-1)b + j$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, 我们可以重写

$$\boldsymbol{x}_n = [\boldsymbol{x}_{n,1}^\top, \ldots, \boldsymbol{x}_{n,a}^\top, \ldots, \boldsymbol{x}_{n,(a-1)b+1}^\top, \ldots, \boldsymbol{x}_{n,c}^\top]^\top,$$

并可类似地定义 κ_i 和 n_i, $i = 1, \ldots, c$. 记 $a_{0,st} = a_{0,(i_1,j_1)(i_2,j_2)}$, 其中 $s = (i_1-1)b + j_1$, $t = (i_2-1)b + j_2$, $s, t = 1, \ldots, c$; $i_1, i_2 = 1, \ldots, a$, $j_1, j_2 = 1, \ldots, b$. 根据结论 (2.20), $\boldsymbol{x}_{n,i}$ $(i = 1, \ldots, c)$ 相互独立且均值为 $\mathbf{0}$, 协方差阵为 $\boldsymbol{\Sigma}$. 由 (2.21), 注意到 $\boldsymbol{A}_0$ 是对称阵, 有

$$T_{n0} = \sum_{i=1}^{c} a_{0,ii}\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,i} + 2\sum_{1\leqslant i<j\leqslant c} a_{0,ij}\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j}. \tag{A.1}$$

所以,

$$\begin{aligned}
\mathrm{E}(T_{n0}) &= \sum_{i=1}^{c} a_{0,ii}\,\mathrm{tr}(\boldsymbol{\Sigma}) = q\,\mathrm{tr}(\boldsymbol{\Sigma}),\\
\mathrm{Var}(T_{n0}) &= \sum_{i=1}^{c} a_{0,ii}^2\,\mathrm{Var}(\|\boldsymbol{x}_{n,i}\|^2) + 4\sum_{1\leqslant i<j\leqslant c} a_{0,ij}^2\,\mathrm{Var}(\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j})\\
&= \sum_{i=1}^{c} a_{0,ii}^2[2\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + \kappa_i n_i^{-1}] + 4\sum_{1\leqslant i<j\leqslant c} a_{0,ij}^2\,\mathrm{tr}(\boldsymbol{\Sigma}^2)\\
&= 2\sum_{i,j=1}^{c} a_{0,ij}^2\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + \sum_{i=1}^{c} a_{0,ii}^2\kappa_i n_i^{-1}\\
&= 2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + \sum_{i=1}^{c} a_{0,ii}^2\kappa_i n_i^{-1},
\end{aligned}$$

其中我们用到了 $\mathrm{E}(\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,i}) = \mathrm{Cov}(\boldsymbol{x}_{n,i}) = \boldsymbol{\Sigma}$, $\mathrm{tr}(\boldsymbol{A}_0^2) = \mathrm{tr}(\boldsymbol{A}_0) = \sum_{i=1}^{c} a_{0,ii} = q$, 对于 $i \neq j$, $\mathrm{E}(\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j}) = 0$, $\|\boldsymbol{x}_{n,i}\|^2$ 与 $\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j}$ 的协方差为 0, $\mathrm{Var}(\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j}) = \mathrm{E}(\boldsymbol{x}_{n,i}^\top\boldsymbol{x}_{n,j}\boldsymbol{x}_{n,j}^\top\boldsymbol{x}_{n,i}) = \mathrm{tr}(\boldsymbol{\Sigma}^2)$, 且 $\mathrm{Var}(\|\boldsymbol{x}_{n,i}\|^2) = 2\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + \kappa_i n_i^{-1}$ (参见文献 [29, 引理 S.4]). □

推论 2.1 的证明 由 (2.21), 有 $T_{n0}=\boldsymbol{x}_n^\top\boldsymbol{A}\boldsymbol{x}_n$, $\boldsymbol{A}=[\boldsymbol{D}_0^{1/2}\boldsymbol{C}_0^\top(\boldsymbol{C}_0\boldsymbol{D}_0\boldsymbol{C}_0^\top)^{-1}\boldsymbol{C}_0\boldsymbol{D}_0^{1/2}]\otimes\boldsymbol{I}_p$, 且根据 (2.20), 在正态假设下, $\boldsymbol{x}_n\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_{ab}\otimes\boldsymbol{\Sigma})$. 注意到 $\boldsymbol{A}=\tilde{\boldsymbol{A}}^\top(\boldsymbol{B}^{-1}\otimes\boldsymbol{I}_p)\tilde{\boldsymbol{A}}$, 其中 $\tilde{\boldsymbol{A}}=(\boldsymbol{C}_0\boldsymbol{D}_0^{1/2})\otimes\boldsymbol{I}_p$, $\boldsymbol{B}=\boldsymbol{C}_0\boldsymbol{D}_0\boldsymbol{C}_0^\top$, 所以, $\boldsymbol{\delta}=\tilde{\boldsymbol{A}}\boldsymbol{x}_n\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{B}\otimes\boldsymbol{\Sigma})$. 从而,

$$T_{n0}=(\tilde{\boldsymbol{A}}\boldsymbol{x}_n)^\top(\boldsymbol{B}^{-1}\otimes\boldsymbol{I}_p)(\tilde{\boldsymbol{A}}\boldsymbol{x}_n)=[(\boldsymbol{B}^{-1/2}\otimes\boldsymbol{I}_p)\boldsymbol{\delta}]^\top[(\boldsymbol{B}^{-1/2}\otimes\boldsymbol{I}_p)\boldsymbol{\delta}],$$

其中 $(\boldsymbol{B}^{-1/2}\otimes\boldsymbol{I}_p)\boldsymbol{\delta}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_q\otimes\boldsymbol{\Sigma})$. 然后利用文献 [20, 引理 1], 即可得到表达式 (2.24). 根据 (2.24), $\mathrm{E}(T_{n0})$、$\mathrm{Var}(T_{n0})$ 和 $\mathrm{E}[T_{n0}-\mathrm{E}(T_{n0})]^3$ 的表达式由文献 [30, (4)] 可得. □

定理 2.2 的证明 首先, 在假设 (A1) 和 (A2) 下, 通过一些简单的代数运算可以得到 $\kappa_i=[\mathrm{E}(z_{i1}^4)-3]\sum_{r=1}^m\gamma_{rr}^2$, $i=1,\ldots,c$, 其中 γ_{rr} 是 $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}$ 的第 r 个对角元素. 因为 $\mathrm{E}(z_{i1}^4)<\infty$ 且 $\sum_{r=1}^m\sigma_{rr}^2\leqslant\mathrm{tr}[(\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma})^2]=\mathrm{tr}(\boldsymbol{\Sigma}^2)$, 所以,

$$\kappa_i=O[\mathrm{tr}(\boldsymbol{\Sigma}^2)],\quad i=1,\ldots,c. \tag{A.2}$$

从而,

$$\sum_{i=1}^c a_{0,ii}^2\frac{\kappa_i}{2n_i}=O\bigg[\sum_{i=1}^c a_{0,ii}^2\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{2n_i}\bigg]=o[\mathrm{tr}(\boldsymbol{\Sigma}^2)],$$

这里用到了 $q=\mathrm{tr}(\boldsymbol{A}_0^2)\geqslant\sum_{i=1}^c a_{0,ii}^2$. 因此, 随着 $n,p\to\infty$, 有

$$d=\frac{q\,\mathrm{tr}^2(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)+\sum_{i=1}^c a_{0,ii}^2\frac{\kappa_i}{2n_i}}=q\frac{\mathrm{tr}^2(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}[1+o(1)].$$

由假设 (A4), 随着 $n,p\to\infty$, 有 $\mathrm{tr}^2(\boldsymbol{\Sigma})/\mathrm{tr}(\boldsymbol{\Sigma}^2)\geqslant\lambda_{\max}^2/\mathrm{tr}(\boldsymbol{\Sigma}^2)\to\infty$, 所以, $d\to\infty$. 根据中心极限定理, 当 $d\to\infty$ 时, $(\beta\chi_d^2-\beta d)/\sqrt{2\beta^2 d}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1)$, 也即

$$\frac{\beta\chi_d^2-q\,\mathrm{tr}(\boldsymbol{\Sigma})}{[2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)]^{1/2}[1+\sum_{i=1}^a\sum_{j=1}^b a_{0,(i,j)}^2\frac{\kappa_{ij}}{2qn_{ij}}/\mathrm{tr}(\boldsymbol{\Sigma}^2)]^{1/2}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1).$$

所以 (2.29) 中的第二式成立. 在假设 A 下, T_{n0} 的渐近正态性, 也即 (2.29) 中的第一式, 可以用文献 [14] 中相同的证明方法根据鞅中心极限定理得到, 这里不再赘述. □

推论 2.2 的证明 由 (2.24) 可知 T_{n0} 为一个系数皆非负的中心卡方型混合. 第一个结论可由类似文献 [20, 定理 2] 的证明得到. 因为 $\boldsymbol{\Sigma}$ 非负定, 所以有 $d^*\geqslant 0$. 根据 Cauchy-Schwartz 不等式可知, $\mathrm{tr}^2(\boldsymbol{\Sigma}^2)\leqslant\mathrm{tr}(\boldsymbol{\Sigma}^3)\,\mathrm{tr}(\boldsymbol{\Sigma})$. 从而, $\mathrm{tr}(\boldsymbol{\Sigma}^3)\geqslant\mathrm{tr}^2(\boldsymbol{\Sigma}^2)/\mathrm{tr}(\boldsymbol{\Sigma})$. 因此,

$$d^*=q\frac{\mathrm{tr}^3(\boldsymbol{\Sigma}^2)}{\mathrm{tr}^2(\boldsymbol{\Sigma}^3)}\leqslant q\frac{\mathrm{tr}^3(\boldsymbol{\Sigma}^2)}{\mathrm{tr}^4(\boldsymbol{\Sigma}^2)/\mathrm{tr}^2(\boldsymbol{\Sigma})}=d, \tag{A.3}$$

第二个结论成立. 再由第一个结论, 当 $d^*\to\infty$ 时, 有

$$\frac{T_{n0}-q\,\mathrm{tr}(\boldsymbol{\Sigma})}{\sqrt{2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1).$$

同时, 根据第二个结论, 当 $d^*\to\infty$ 时有 $d\to\infty$, 所以,

$$\frac{\beta\chi_d^2-q\,\mathrm{tr}(\boldsymbol{\Sigma})}{\sqrt{2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}}\xrightarrow{d}\frac{\chi_d^2-d}{\sqrt{2d}}\xrightarrow{\mathcal{L}}\mathcal{N}(0,1).$$

证毕. □

定理 2.3 的证明 根据 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}$、$\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}$、$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ 和 $\hat{\kappa}_{ij}$ 的比率相合性, 由 (2.34) 和 (2.26) 易知 $\hat{\beta}/\beta \xrightarrow{P} 1$, $\hat{d}/d \xrightarrow{P} 1$, 所以, $\hat{\beta}\chi^2_{\hat{d}}(\alpha)$ 也为 $\beta\chi^2_d(\alpha)$ 的比率相合估计. □

推论 2.3 的证明 首先证明 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}$ 是 $\mathrm{tr}(\boldsymbol{\Sigma})$ 的比率相合估计. 由文献 [28, (5)] 可知,

$$\mathrm{Var}[\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij})] = 2(n_{ij}-1)^{-1}\,\mathrm{tr}(\boldsymbol{\Sigma}^2) + n_{ij}^{-1}\kappa_{ij}.$$

在定理 2.2 的条件下, 根据 (A.2) 可知 $n_{ij}^{-1}\kappa_{ij}/\,\mathrm{tr}^2(\boldsymbol{\Sigma}^2) = o(1)$; 而在推论 2.2 的条件下, 根据正态分布的性质, $\kappa_{ij} = 0$. 所以总有 $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij})/\,\mathrm{tr}(\boldsymbol{\Sigma}) \xrightarrow{P} 1$. 根据 (2.31), 可得

$$\frac{\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}}{\mathrm{tr}(\boldsymbol{\Sigma})} = (n-k)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-1)\frac{\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij})}{\mathrm{tr}(\boldsymbol{\Sigma})} \xrightarrow{P} 1.$$

另外, 由文献 [39, 定理 2] 或 [25, 定理 2] 可得 $\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}$ 的比率相合性. 最后, 由 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma})}$ 和 $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ 的比率相合性和 (A.2), 利用 Slutsky 定理即可得到结论 (2.35). □

定理 3.1 的证明 (1) 由统计量的定义 (2.16) 即可验证. 对于 (2), 只需要注意到估计量 $\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{ij})$、$\widehat{\mathrm{tr}^2(\boldsymbol{\Sigma})}_{ij}$、$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}_{ij}$ 和 $\hat{\kappa}_{ij}$ 皆为正交 - 平移不变量. 对于 (3), 由统计量 T_n、估计量 $\hat{\beta}$ 和 $\hat{d}$ 的定义易知它们都是组内标号置换不变的. □

(4.2) 的证明 根据定义 (2.17) 和结论 (2.20), 并注意到 $\boldsymbol{x}_n = \boldsymbol{D}^{-1/2}\boldsymbol{x}$, $\boldsymbol{A}_0^2 = \boldsymbol{A}_0$,

$$\begin{aligned}
&\mathrm{E}(S_n) = [\boldsymbol{C}\,\mathrm{E}(\boldsymbol{x})]^\top(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\mu}) = \boldsymbol{0},\\
&\mathrm{Cov}(S_n) = \mathrm{Cov}(\boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{x}) = \boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}\boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{x}_n)\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{\mu}\\
&\qquad = \boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}\boldsymbol{A}(\boldsymbol{I}_{ab}\otimes\boldsymbol{\Sigma})\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{\mu}\\
&\qquad = \boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}(\boldsymbol{A}_0\otimes\boldsymbol{I}_p)(\boldsymbol{I}_{ab}\otimes\boldsymbol{\Sigma})(\boldsymbol{A}_0\otimes\boldsymbol{I}_p)\boldsymbol{D}^{-1/2}\boldsymbol{\mu}\\
&\qquad = \boldsymbol{\mu}^\top\boldsymbol{D}^{-1/2}(\boldsymbol{A}_0\otimes\boldsymbol{\Sigma})\boldsymbol{D}^{-1/2}\boldsymbol{\mu}.
\end{aligned}$$

证毕. □

定理 4.1 的证明 注意到 $\lim_{n\to\infty} n^{-1}(\boldsymbol{C}\boldsymbol{\mu})^\top(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\mu}) = \Delta$. 在给定的条件和假设 (4.3) 下, 根据 (4.1) 和定理 2.3, 有

$$\begin{aligned}
\Pr\left[\frac{T_n}{\hat{\beta}} \geqslant \chi^2_{\hat{d}}(\alpha)\right] &\approx \Pr\left[\frac{T_{n0}}{\beta} \geqslant \frac{\hat{\beta}\chi^2_{\hat{d}}(\alpha)}{\beta\chi^2_d(\alpha)}\chi^2_d(\alpha) - \frac{(\boldsymbol{C}\boldsymbol{\mu})^\top(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\mu})}{\beta}\right]\\
&\approx \Pr\left[\chi^2_d \geqslant \chi^2_d(\alpha) - \frac{n\Delta}{\beta}\right].
\end{aligned}$$

证毕. □

定理 4.2 的证明 在给定的条件和假设 (4.3) 下, 根据 (4.1) 和定理 2.3, 有

$$\begin{aligned}
\Pr\left\{\frac{T_n - q\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}})}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}} \geqslant z_\alpha\right\} &= \Pr\left\{\frac{T_{n0} - q\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}})}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}} \geqslant z_\alpha - \frac{(\boldsymbol{C}\boldsymbol{\mu})^\top(\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\mu})}{[2q\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}]^{1/2}}\right\} + o(1)\\
&= \Phi\left\{-z_\alpha + \frac{n\Delta}{[2q\,\mathrm{tr}(\boldsymbol{\Sigma}^2)]^{1/2}}\right\} + o(1).
\end{aligned}$$

证毕. □

An L^2-norm based test for high-dimensional two-way MANOVA

Bu Zhou, Jia Guo & Jin-Ting Zhang

Abstract In this paper, we propose and study an L^2-norm based test for high-dimensional two-way MANOVA where there are fewer observations than the dimension. The test statistic is constructed by removing the inverse sample covariance matrix in the Wald-type statistic for the general linear hypothesis. We propose to approximate the null distribution of our test statistic by using the well-known Welch-Satterthwaite chi-squared approximation and discuss the relationship between chi-squared approximation and the commonly-used normal approximation. The L^2-norm based test is also shown to admit several invariant properties under certain transformations. The asymptotic and approximate powers of the proposed test are investigated. Simulation studies and real data applications show that the proposed test performs well for high-dimensional data.

Keywords **high-dimensional data, L^2-norm, chi-squared-type mixtures, Welch-Satterthwaite chi-squared approximation, two-way MANOVA, test of linear hypotheses**

MSC(2010) **62H15, 62F05**

doi: 10.1360/N012019-00060