中国科学: 数学 2022 年 第 52 卷 第 10 期: 1183 ~ 1202
SCIENTIA SINICA Mathematica
论 文

基于投影的两样本分布相等的非参数检验

许凯[1], 朱利平[2*]

1. 安徽师范大学数学与统计学院, 芜湖 241002;
2. 中国人民大学统计与大数据研究院, 北京 100872
E-mail: tjxxukai@163.com, zhu.liping@ruc.edu.cn

收稿日期: 2020-11-03; 接受日期: 2021-08-16; 网络出版日期: 2022-03-10; * 通信作者
国家自然科学基金 (批准号: 11901006, 11731011 和 11931014) 和安徽省自然科学基金 (批准号: 1908085QA06) 资助项目

摘要 本文通过利用投影推广经典的 Cramér-von Mises 度量, 研究能够适应于高维数据的两样本分布相等的非参数检验. 本文所提出的新度量是非负的, 且新度量等于零当且仅当两总体具有相同分布, 这确保相应的检验能够探测任意的备择假设. 本文所推荐的检验统计量具有精确的代数表示, 不依赖于任何冗余参数. 由于新度量的定义不需要任何的矩条件, 本文所提出的检验方法能够对数据中强影响值和异常值稳健. 在传统的 "大样本、固定维数" 和新型的 "固定样本、高维数" 框架下, 研究了新检验统计量的渐近性质. 在传统的 "大样本、固定维数" 框架下, 证明所推荐的检验功效不依赖于两样本量的比值的大小, 这确保新检验可以适用于非平衡样本的数据. 在新型的 "固定样本、高维数" 框架下, 证明所推荐的检验功效主要由两总体的位置和尺度差异决定. 在这一框架下, 本文进一步修正所推荐的检验统计量使其在探测两总体的位置和尺度差异时具有更高的功效. 数值研究表明本文所提出的检验是有效和切实可行的.

关键词 Cramér-von Mises 检验 分布相等 高维 投影 两样本问题

MSC (2020) 主题分类 62H15, 62G10, 62G20

1 引言

考虑随机向量 $\boldsymbol{x}=(X_1,\ldots,X_p)^{\mathrm{T}}\in\mathbb{R}^p$ 和 $\boldsymbol{y}=(Y_1,\ldots,Y_p)^{\mathrm{T}}\in\mathbb{R}^p$, F 和 G 分别为 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的分布函数且未知. $\{\boldsymbol{x}_i,i=1,\ldots,m\}$ 和 $\{\boldsymbol{y}_i,i=1,\ldots,n\}$ 分别是来自 F 和 G 的两个简单随机样本. 本文感兴趣的问题是检验

$$H_0:F=G \quad \text{vs.} \quad H_1:F\neq G. \tag{1.1}$$

检验两独立样本的异质性是统计学中最基本的问题之一 (参见文献 [1,2]). 在正态假设下, Student t 和 Hotelling T^2 [3] 这两类似然比检验方法常用来检验 $p=1$ 和 $p>1$ 时两总体的一阶矩 (均值) 是否相

英文引用格式: Xu K, Zhu L P. Nonparametric two-sample tests for equality of distributions using projections (in Chinese). Sci Sin Math, 2022, 52: 1183–1202, doi: 10.1360/SSM-2020-0317

等. 当维数为 1 或固定且样本趋于无穷大时它们也是最有效的. 但当维数发散时, Hotelling T^2 检验会失效. 文献 [4–7] 理论上证明了高维数据下 Hotelling T^2 检验失效的原因, 并对它进行修正. 两个总体的一阶矩并不能完全描述两个总体分布间的差异. 为此, 研究者有时候对两个总体的二阶矩是否相等较为感兴趣, 即两样本协方差相等检验问题. 对于固定和发散维数的两样本协方差相等检验的研究已有大量的文献, 可参见文献 [3,8–12]. 上述的均值、协方差相等检验是一种受限制的两样本分布相等检验. 众所周知, 对多元正态总体而言, 均值和协方差足够刻画两样本间的差异. 但对多元非正态总体而言, 一、二阶矩并不能完全刻画两样本间的差异.

检验 (1.1) 已受到国内外统计学者的广泛关注. 在一维情形下, Kolmogorov-Smirnov (KS) [13] 检验和 Cramér-von Mises (CvM) [14, 15] 检验是最著名的两个基于经验分布函数的方法. 由于相应的检验统计量仅与样本的秩有关, 当 $p = 1$ 时, KS 和 CvM 检验统计量的极限零分布不依赖于任何冗余参数, 不需要任何矩条件. 但随着维数 p 的增加, 它们会遭遇所谓的维数灾难问题 (参见文献 [16]). 在多维情形下, 主要存在 4 类检验. 从密度函数的角度, 文献 [17] 通过比较两总体密度的差值构造检验统计量, 文献 [18] 通过比较两总体密度的商值构造检验统计量. 这些方法是非参数光滑检验 (smooth test, ST). 因而, ST 需要考虑密度函数估计至关重要的窗宽或者节点数的选择等问题, 且无法在高维数据下直接使用. 从特征函数的角度, 文献 [19] 推荐势能检验统计量 (energy statistic, ES). ES 是一个很受欢迎的方法, 文献 [20] 基于文献 [21] 证明两样本 ES 检验能够应用到高维数据中. 但遗憾的是, 势能检验需要总体分布满足某些矩条件因而对厚尾数据或者含异常值点的数据不够稳健. 从两总体分布最大均值偏差 (maximum mean discrepancy, MMD) 的角度, 文献 [22] 提出了基于正定核的检验法, 因而 MMD 亦需要选择合理的正定核及窗宽, 且高维数据下, MMD 的有效性问题尚未得到解决. 从图结构 (graph constructions, GC) 的角度, 文献 [23] 基于最小生成树构造检验统计量, 文献 [24] 基于最临近法构造检验统计量, 文献 [25] 基于组合样本的秩构造检验统计量, 文献 [26] 通过比较样本间 Euclid 距离的临近值构造检验统计量, 文献 [27] 基于最短 Hamilton 路径构造检验统计量. 尽管这些 GC 方法很有用甚至它们中某些方法在高维数据情形下也可使用, 但 GC 方法需要选择一些调节参数, 如最临近数、最优权重和最优路径等. 最近, 文献 [28] 在可分离的 Banach 空间上, 提出利用球偏离系数 (ball divergence, BD) 构造检验统计量. 基于 BD 的检验法, 对于非平衡数据和尺度差异的备择结构非常有效, 但在高维情形下这一方法的表现有待进一步研究.

通过利用投影推广经典的 Cramér-von Mises (CvM) 度量, 本文研究能够适应于高维数据的两样本分布相等的非参数检验. 利用投影方法去检验单样本的多元分布拟合优度问题很早就受到国内学者的关注 (参见文献 [29,30]). 但鲜有文献在两样本/高维两样本框架下考虑如何应用投影方法, 本文希冀解决这个问题. 通过投影 $\boldsymbol{x} \in \mathbb{R}^p$ 和 $\boldsymbol{y} \in \mathbb{R}^p$ 到单位球空间 $\mathcal{S}^{p-1} \overset{\text{def}}{=} \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| = 1\}$, 可以基于如下广义的 CvM 统计量:

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}} \left[\int_{-\infty}^{\infty} \{\widehat{F}_{\boldsymbol{\alpha}}(t) - \widehat{G}_{\boldsymbol{\alpha}}(t)\}^2 d\widehat{H}_{\boldsymbol{\alpha}}(t) \right] d\boldsymbol{\alpha} \tag{1.2}$$

去比较 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的分布差异, 其中 $\widehat{F}_{\boldsymbol{\alpha}}(t)$ 和 $\widehat{G}_{\boldsymbol{\alpha}}(t)$ 分别为两投影样本 $\{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i), i = 1, \ldots, m\}$ 和 $\{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i), i = 1, \ldots, n\}$ 的经验分布函数, $\widehat{H}_{\boldsymbol{\alpha}}(t) \overset{\text{def}}{=} \{m\widehat{F}_{\boldsymbol{\alpha}}(t) + n\widehat{G}_{\boldsymbol{\alpha}}(t)\}/(m+n)$. 使用 (1.2) 去检验 (1.1) 有如下几个优点:

(1) 我们将证明统计量 (1.2) 具有显式代数表示, 不依赖于投影方向选择.

(2) 令 $\tau \overset{\text{def}}{=} m/(m+n)$ 和 $H_{\boldsymbol{\alpha}}(t) \overset{\text{def}}{=} \tau F_{\boldsymbol{\alpha}}(t) + (1-\tau)G_{\boldsymbol{\alpha}}(t)$, 其中 $F_{\boldsymbol{\alpha}}(t)$ 和 $G_{\boldsymbol{\alpha}}(t)$ 分别表示 $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x})$

和 $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y})$ 的分布函数. 统计量 (1.2) 对应的总体度量

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\left[\int_{-\infty}^{\infty}\{F_{\boldsymbol{\alpha}}(t)-G_{\boldsymbol{\alpha}}(t)\}^2dH_{\boldsymbol{\alpha}}(t)\right]d\boldsymbol{\alpha} \tag{1.3}$$

是非负的且等于零时当且仅当两总体具有相同分布, 这确保相应的检验能够探测任意的备择假设.

(3) 由于 (1.3) 的定义不需要任何的矩条件, 本文所推荐的检验方法能够对数据中强影响值和异常值稳健.

我们所提出的投影方法与文献 [30] 中的投影方法有密切联系, 为进一步加强文章的动机和背景, 我们指出如下 3 点不同之处: (i) 方法上, 本文研究的是两样本问题, 而文献 [30] 考虑的是具有冗余参数的单样本问题; (ii) 计算上, 文献 [30] 的方法需要利用 Monte Carlo 近似选择若干投影方向去计算他们的检验统计量, 而本文借助于文献 [16] 中的计算方法成功避开了投影方向的选择问题; (iii) 第 3 个不同是, 在维数发散的情形下, 本文研究所提出统计量的高维性质, 并进一步考虑如何在高维情形下对所提出检验进行改进修正, 而文献 [30] 考虑的是固定维数下的问题.

本文余下内容的结构安排如下. 第 2 节详细地研究基于投影的总体 CvM 度量 (1.3) 和样本 CvM 度量 (1.2) 的代数表示问题. 第 3 节在传统 “大样本、固定维数” 框架下, 研究所推荐的检验统计量的大样本性质. 特别地, 从理论角度解释了新检验可以适用于非平衡样本的数据. 第 4 节在新型的 “固定样本、高维数” 框架下, 证明所推荐的检验功效主要依赖于两总体的位置和尺度差异. 在这一框架下, 本文进一步修正所推荐的检验统计量使其在探测两总体的位置和尺度差异时具有更高的功效. 第 5 节利用数值模拟实验验证本文所提出的检验方法的有效性和可行性. 相关定理的证明可参见附录 A.

2 检验统计量

本节构造一个检验统计量去检验 (1.1), 并解释其合理性. 利用特征函数这一工具, 可以证明 $F=G$ 当且仅当对所有 $\boldsymbol{\alpha}\in\mathcal{S}^{p-1}$, $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x})$ 和 $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y})$ 具有相同分布函数. 给定 $\boldsymbol{\alpha}\in\mathcal{S}^{p-1}$, $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x})$ 和 $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y})$ 具有相同分布函数当且仅当

$$\int_{-\infty}^{\infty}\{F_{\boldsymbol{\alpha}}(t)-G_{\boldsymbol{\alpha}}(t)\}^2dH_{\boldsymbol{\alpha}}(t)=0. \tag{2.1}$$

为把所有的投影方向考虑进来, 对 (2.1) 左边的 $\boldsymbol{\alpha}\in\mathcal{S}^{p-1}$ 进行积分, 获得基于投影的 CvM 度量 (1.3). 更重要的是, (1.3) 是非负的, 且 (1.3) 等于零当且仅当 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 具有相同分布.

接下来推导出 (1.2) 和 (1.3) 的精确代数表示. 根据 $H_{\boldsymbol{\alpha}}(t)$ 的定义, (1.3) 等价表示为

$$\begin{aligned}&\tau\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\left[\int_{-\infty}^{\infty}\{F_{\boldsymbol{\alpha}}(t)-G_{\boldsymbol{\alpha}}(t)\}^2dF_{\boldsymbol{\alpha}}(t)\right]d\boldsymbol{\alpha}\\&\quad+(1-\tau)\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\left[\int_{-\infty}^{\infty}\{F_{\boldsymbol{\alpha}}(t)-G_{\boldsymbol{\alpha}}(t)\}^2dG_{\boldsymbol{\alpha}}(t)\right]d\boldsymbol{\alpha}.\end{aligned} \tag{2.2}$$

假设 $m,n\geqslant 3$. 由 $F_{\boldsymbol{\alpha}}(t)$ 和 $G_{\boldsymbol{\alpha}}(t)$ 的定义, 有

$$\begin{aligned}&\int_{-\infty}^{\infty}\{F_{\boldsymbol{\alpha}}(t)-G_{\boldsymbol{\alpha}}(t)\}^2dF_{\boldsymbol{\alpha}}(t)\\&\quad=\mathrm{E}[\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_1\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)-I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_1\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)\}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_2\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)-I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_2\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)\}].\end{aligned}$$

利用 Fibini 定理知, (2.2) 的第一项等于

$$\tau \mathrm{E}\bigg[\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_1\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)-I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_1\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)\}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_2\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)-I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_2\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_3)\}d\boldsymbol{\alpha}\bigg]. \quad (2.3)$$

根据文献 [31, 引理 1], (2.3)/τ 具有如下显式表示:

$$\begin{aligned}&c_p[\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_3,\boldsymbol{y}_2-\boldsymbol{x}_3)+\mathrm{ang}(\boldsymbol{y}_1-\boldsymbol{x}_3,\boldsymbol{x}_2-\boldsymbol{x}_3)\\&\quad-\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_3,\boldsymbol{x}_2-\boldsymbol{x}_3)-\mathrm{ang}(\boldsymbol{y}_1-\boldsymbol{x}_3,\boldsymbol{y}_2-\boldsymbol{x}_3)],\end{aligned}$$

其中, $c_p\overset{\mathrm{def}}{=}\pi^{p/2-1}/\Gamma(p/2)$, $\Gamma(\cdot)$ 是 Γ 函数, $\mathrm{ang}(\boldsymbol{x},\boldsymbol{y})$ 表示向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间的角度, 即

$$\mathrm{ang}(\boldsymbol{a},\boldsymbol{b})\overset{\mathrm{def}}{=}\arccos\bigg(\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}}{\|\boldsymbol{a}\|\cdot\|\boldsymbol{b}\|}\bigg),\quad \|\boldsymbol{a}\|=(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})^{1/2},$$

$\arccos(\cdot)$ 是反三角余弦函数. 记 $T_1\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_3,\boldsymbol{y}_2-\boldsymbol{x}_3)\}$, $T_2\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{x}_3,\boldsymbol{x}_2-\boldsymbol{x}_3)\}$ 和 $T_3\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{y}_1-\boldsymbol{x}_3,\boldsymbol{y}_2-\boldsymbol{x}_3)\}$. 从而, (2.3) 等于 $c_p\tau(2T_1-T_2-T_3)$. 进一步, 记 $T_4\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{y}_3,\boldsymbol{y}_2-\boldsymbol{y}_3)\}$, $T_5\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{x}_1-\boldsymbol{y}_3,\boldsymbol{x}_2-\boldsymbol{y}_3)\}$ 和 $T_6\overset{\mathrm{def}}{=}\mathrm{E}\{\mathrm{ang}(\boldsymbol{y}_1-\boldsymbol{y}_3,\boldsymbol{y}_2-\boldsymbol{y}_3)\}$. 类似讨论知 (2.2) 的第 2 项等价表示为 $c_p(1-\tau)(2T_4-T_5-T_6)$. 因此, (1.3) 的显式形式是上述两个结果的总和, 严格的表述参见定理 1. 为记号方便, 定义

$$T\overset{\mathrm{def}}{=}\tau(2T_1-T_2-T_3)+(1-\tau)(2T_4-T_5-T_6).$$

定理 1 令 F 和 G 分别表示 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的分布函数且假设 $m,n\geqslant 3$, 则 (1.3) 等于 c_pT, T 是非负的, 且 T 等于零当且仅当 $F=G$.

根据定理 1, 可以利用 T 的样本形式 (1.2) 去检验 (1.1). 基于样本 $\{\boldsymbol{x}_i,i=1,\ldots,m\}$ 和 $\{\boldsymbol{y}_i,i=1,\ldots,n\}$, 利用如下形式的 V 统计量去估计 T_k:

$$\begin{aligned}&\widehat{T}_1\overset{\mathrm{def}}{=}\frac{1}{nm^2}\sum_{i,j=1}^{m}\sum_{k=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{y}_k-\boldsymbol{x}_j),\quad \widehat{T}_2\overset{\mathrm{def}}{=}\frac{1}{m^3}\sum_{i,j,k=1}^{m}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{x}_k-\boldsymbol{x}_j),\\&\widehat{T}_3\overset{\mathrm{def}}{=}\frac{1}{n^2m}\sum_{i,j=1}^{n}\sum_{k=1}^{m}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{x}_k,\boldsymbol{y}_j-\boldsymbol{x}_k),\quad \widehat{T}_4\overset{\mathrm{def}}{=}\frac{1}{mn^2}\sum_{i=1}^{m}\sum_{j,k=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{y}_k,\boldsymbol{y}_j-\boldsymbol{y}_k),\\&\widehat{T}_5\overset{\mathrm{def}}{=}\frac{1}{nm^2}\sum_{i,j=1}^{m}\sum_{k=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{y}_k,\boldsymbol{x}_j-\boldsymbol{y}_k),\quad \widehat{T}_6\overset{\mathrm{def}}{=}\frac{1}{n^3}\sum_{i,j,k=1}^{n}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{y}_k,\boldsymbol{y}_j-\boldsymbol{y}_k).\end{aligned}$$

记 $\widehat{T}=\tau(2\widehat{T}_1-\widehat{T}_2-\widehat{T}_3)+(1-\tau)(2\widehat{T}_4-\widehat{T}_5-\widehat{T}_6)$. 类似于定理 1, 如下的定理 2 建立了 (1.2) 与 $\widehat{T}$ 之间的等价性.

定理 2 对任意 $m,n\geqslant 3$, (1.2) 等于 $c_p\widehat{T}$, $\widehat{T}$ 是非负的.

根据定理 1 和 2, 可以基于统计量 $\widehat{T}$ 去构造检验函数.

3 “大样本、固定维数” 框架下的渐近性质

记 $F_{\widehat{T}}(t)=\mathrm{pr}(\widehat{T}\leqslant t)$. 在显著水平 α 给定下, 若 $\widehat{T}$ 大于临界值 $c_\alpha=\inf\{t\in\mathbb{R}^1:1-F_{\widehat{T}}(t)\leqslant\alpha\}$, 则拒绝 H_0, 否则就接受它. 为了获得 c_α, 需要研究 $\widehat{T}$ 的渐近性质. 首先在 “大样本、固定维数” 框架下研究 $\widehat{T}$ 的收敛性. 换句话讲, 首先考虑 $\min(m,n)\to\infty$ 但 p 固定. 假设当 $\min(m,n)$ 趋向于 ∞ 时 $\tau\to\tau_0\geqslant 0$. $\tau_0=0$ 或 $\tau_0=1$ 暗示着观测样本是非平衡的.

定理 3 (1) 在 H_0 下, 当 $\min(m,n)$ 趋向于 ∞ 时,

$$\left\{\frac{mn}{m+n}\right\}\widehat{T}\xrightarrow{d}\sum_{k=1}^{\infty}\lambda_k Z_k^2,$$

其中, $\{\lambda_k\}$ 是非负的且依赖于分布函数 $F=G$, $\{Z_k\}$ 是独立的且具有标准正态分布随机序列.

(2) 在 H_1 下, 当 $\min(m,n)$ 趋向于 ∞ 时,

$$\{9n^{-1}\sigma_{01}^2+9m^{-1}\sigma_{10}^2\}^{-1/2}(\widehat{T}-T)\xrightarrow{d}N(0,1),$$

其中 σ_{01}^2 和 σ_{10}^2 分别定义在 (A.2) 和 (A.3) 中, 且 $\{mn/(m+n)\}\widehat{T}$ 依概率收敛到 ∞.

定理 3 表明利用统计量 $\widehat{T}$ 构造的检验是相合的. 遵照文献 [32] 的备择结构 (3.2), 下面研究统计量 $\widehat{T}$ 在相邻备择结构:

$$H_{1n}: f_{12}(\boldsymbol{x},\boldsymbol{y})=f_1(\boldsymbol{x})f_2(\boldsymbol{y},\widehat{\theta})$$

下的功效, 这里 $\widehat{\theta}\to 0$ (随着 $\min(m,n)\to\infty$), f_{12}、f_1 和 f_2 分别表示 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的联合及边缘密度函数且 $f_1(\cdot)=f_2(\cdot,0)$. 为此, 需要如下 3 个条件.

(C1) 对于一切 $\boldsymbol{y}\in\mathbb{R}^p$, $f_2(\boldsymbol{y},\theta)$ 关于 $\theta\in\Theta$ 是绝对连续的, 这里 Θ 是有界的参数空间.

(C2) 对于一切 $\theta\in\Theta$, $\partial f_2(\boldsymbol{y},\theta)/\partial\theta$ 对几乎一切 $\boldsymbol{y}\in\mathbb{R}^p$ 存在.

(C3) 对于一切 $\theta\in\Theta$, Fisher 信息量 $\boldsymbol{I}_{f_2}(\theta)=\mathrm{E}[\partial\log\{f_2(\boldsymbol{y},\theta)\}/\partial\theta]^2$ 存在, $\boldsymbol{I}_{f_2}(0)>0$ 且 $\boldsymbol{I}_{f_2}(\theta)$ 在 $\theta=0$ 处连续.

条件 (C1)–(C3) 分别对应于文献 [32] 中的条件 (A1)–(A3). 这些条件要求总体的密度存在并满足一定的规则假设. 值得注意的是, 统计量 $\widehat{T}$ 在 H_0 和 H_1 下的收敛性并不需要这些条件.

命题 1 在 H_{1n} 下, 假设条件 (C1)–(C3) 都满足, 当 $n^{1/2}\widehat{\theta}\to\infty$ (随着 $\min(m,n)\to\infty$) 且 $\tau_0>0$ 时, $\{mn/(m+n)\}\widehat{T}$ 依概率收敛到 ∞.

由命题 1 易知, 基于 $\widehat{T}$ 的检验能够探测以参数收敛速度 $O(n^{-1/2})$ 收敛到原假设的相邻备择假设.

根据定理 3 知, 本文所推荐的检验能够探测任意的备择假设. 临界值 c_α 是由 $\{\lambda_k\}$ 决定的, 但渐近零分布中的 $\{\lambda_k\}$ 依赖于未知分布 $F=G$. 为了获得可行的检验函数, 我们推荐如下的排序抽样法去近似渐近零分布 $\sum_{k=1}^{\infty}\lambda_k Z_k^2$.

(1) 随机地对 $\{\boldsymbol{x}_i, i=1,\ldots,m\}$ 和 $\{\boldsymbol{y}_j, j=1,\ldots,n\}$ 合并的样本进行排序;

(2) 选择前 m 个观测记为 $\{\boldsymbol{x}_i^b, i=1,\ldots,m\}$, 剩余的 n 个观测记为 $\{\boldsymbol{y}_j^b, j=1,\ldots,n\}$;

(3) 基于随机的排序样本 $\{\boldsymbol{x}_i^b, i=1,\ldots,m\}$ 和 $\{\boldsymbol{y}_j^b, j=1,\ldots,n\}$, 再次计算统计量 $\widehat{T}^b$;

(4) 重复上述步骤 B 次, 获得基于排序样本的统计量 $\widehat{T}^b$ $(b=1,\ldots,B)$. 所推荐检验的临界值 c_α 可近似为

$$\inf\{t\in\mathbb{R}^1: 1-\widehat{F}_B(t)\leqslant\alpha\},$$

其中 $\widehat{F}_B(t)=B^{-1}\sum_{b=1}^{B}I(\widehat{T}^b\leqslant t)$. 如下的定理 4 研究了上述排序方法的相合性.

定理 4 (1) 在 H_0 下, 当 $\min(m,n)$ 趋向于 ∞ 时,

$$\mathrm{pr}(\widehat{T}^b\leqslant t\mid\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)-\mathrm{pr}(\widehat{T}\leqslant t)\xrightarrow{\mathrm{pr}}0,\quad\text{对所有 } t\in\mathbb{R}^+.$$

(2) 在 H_1 下, 当 $\min(m,n)$ 趋向于 ∞ 时,

$$\mathrm{pr}(\widehat{T}^b\geqslant\widehat{T}\mid\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)\xrightarrow{\mathrm{pr}}0.$$

在 H_0 下, 基于随机排序样本的统计量 $\widehat{T}^b$ 与基于原随机样本的统计量 $\widehat{T}$ 具有相同分布. 定理 4 的第 1 个结论表明对于足够大的 B, $\inf\{t\in\mathbb{R}^1: 1-\widehat{F}_B(t)\leqslant\alpha\}$ 是临界值 c_α 一个好的近似. 在 H_1 下, $\widehat{T}$ 依概率大于 0, 而 $\widehat{T}^b$ 依概率收敛到 0. 定理 4 的第 2 个结论表明所推荐的检验能够探测任意的备择假设.

接下来, 定理 5 将证明所推荐检验的相合性与 τ 无关. 这确保新检验可以适用于非平衡样本的数据. 为了记号方便, 令

$$D_1 \overset{\text{def}}{=} \lim_{\min(m,n)\to\infty}\{\mathrm{var}_{H_1}(\widehat{T})\} \quad 和 \quad D_2 \overset{\text{def}}{=} \liminf_{\min(m,n)\to\infty}\{\mathrm{E}_{H_1}(\widehat{T})-\mathrm{E}_{H_0}(\widehat{T})\},$$

其中, var_{H_1} 表示在 H_1 下计算的方差, E_{H_0} 和 E_{H_1} 分别表示在 H_0 和 H_1 下计算的期望.

定理 5 对于 H_0 和任意备择 H_1, 有 $D_1=0$ 和 $D_2>0$. 更重要的是, D_2 不依赖于$\tau=m/(m+n)$.

4 “固定样本、高维数” 框架下的渐近性质

本节考虑 m 和 n 都是固定的但 p 趋于 ∞ 时, $\widehat{T}$ 及其修正统计量 $\widehat{T}^*$ 的渐近性质. 令 $\boldsymbol{u}_1$ 和 $\boldsymbol{\Sigma}_1$ 分别表示 $\boldsymbol{x}$ 的均值和协方差, $\boldsymbol{u}_2$ 和 $\boldsymbol{\Sigma}_2$ 分别表示 $\boldsymbol{y}$ 的均值和协方差. 为了数理分析, 我们作如下假设.

(A1) 对于 $k=1,\ldots,p$, X_k 和 Y_k 的四阶矩是一致有界的, 即

$$\limsup_{p\to\infty}\max_{k=1,\ldots,p}\mathrm{E}(X_k^4)<\infty \quad 和 \quad \limsup_{p\to\infty}\max_{k=1,\ldots,p}\mathrm{E}(Y_k^4)<\infty.$$

(A2) 对于 $i=1,\ldots,n$ 和 $k=1,\ldots,p$, 记 U_k 和 V_k 表示 $X_{i,k}$ 或 $Y_{i,k}$. 假设

$$\sum_{k\neq l}|\mathrm{corr}\{(U_k-V_k)(U_k-V_k),(U_l-V_l)(U_l-V_l)\}|=o(p^2).$$

(A3) 当 p 趋于 ∞ 时, 假设 $\mathrm{tr}(\boldsymbol{\Sigma}_1)/p$、$\mathrm{tr}(\boldsymbol{\Sigma}_2)/p$ 和 $\|\boldsymbol{u}_1-\boldsymbol{u}_2\|^2/p$ 的极限存在. 它们的极限值分别表示为 σ_1^2、σ_2^2 和 ν^2, 即

$$\sigma_k^2 \overset{\text{def}}{=} \lim_{p\to\infty}\frac{\mathrm{tr}(\boldsymbol{\Sigma}_k)}{p}, \quad k=1,2 \quad 和 \quad \nu^2 \overset{\text{def}}{=} \lim_{p\to\infty}\frac{\|\boldsymbol{u}_1-\boldsymbol{u}_2\|^2}{p}.$$

在 “固定样本、高维数” 框架下, 假设条件 (A1) 成立. 这是高维渐近分析所必须付出的代价, 参见文献 [21]. 条件 (A2) 要求 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的成分不是强相依的. 条件 (A3) 是一个标准的高维假设, 参见文献 [20, 21, 27, 33]. 事实上, 如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的分量都是独立同分布的, 条件 (A1) 可以进一步弱化而条件 (A2) 和 (A3) 自然成立.

对于 $k=1,2$, 令

$$T_{k,1} \overset{\text{def}}{=} \arccos\left[\frac{\sigma_k}{\{2(\sigma_1^2+\sigma_2^2+\nu^2)\}^{1/2}}\right] \quad 和 \quad T_{k,2} \overset{\text{def}}{=} \arccos\left\{\frac{\sigma_k^2+\nu^2}{\sigma_1^2+\sigma_2^2+\nu^2}\right\}.$$

进一步表示

$$\gamma_0 \overset{\text{def}}{=} \tau\left(2T_{1,1}-T_{1,2}-\frac{\pi}{3}\right)+(1-\tau)\left(2T_{2,1}-T_{2,2}-\frac{\pi}{3}\right).$$

引理 1 假设条件 (A1)–(A3) 成立. 当 p 趋于 ∞ 时, 有 $\lim_{p\to\infty} T=\gamma_0$, $\widehat{T}\xrightarrow{\text{pr}}\gamma_0$ 且 $\gamma_0\geqslant 0$. 更重要的是, $\sigma_1^2=\sigma_2^2$ 和 $\nu^2=0$ 当且仅当 $\gamma_0=0$.

引理 1 的发现与文献 [34, 35] 中的结果是一致的, 即比较两个高维随机向量投影的分布差异仅需要比较它们的均值和方差的差异. 由于 $\gamma_0 = 0$ 当且仅当 $\sigma_1^2 = \sigma_2^2$ 和 $\nu^2 = 0$, 因此, 当 p 趋于 ∞ 时, 所推荐的检验可以探测 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 均值和方差的差异.

在 "固定样本、高维数" 框架下, 我们仍然使用第 3 节的排序抽样法去决定所推荐检验的 p- 值

$$B^{-1}\sum_{b=1}^{B} I(\widehat{T}^b \geqslant \widehat{T}).$$

$C(m+n, n)$ 表示从 $m+n$ 个不同元素中选取 n 个不同元素的组合数.

定理 6 假设条件 (A1)−(A3) 成立. 当 p 趋于 ∞ 时, 如果 $\nu^2 > 0$ 或 $\sigma_1^2 \neq \sigma_2^2$, 则由排序抽样法所决定的极限 p- 值依概率收敛到 $2/C(m+n, n)$.

给定显著水平 α, 定理 6 表明当 p 趋于 ∞ 时, 只需要 $2/C(m+n, n) \leqslant \alpha$, 则所推荐检验的功效依概率收敛到 1. 该条件 $2/C(m+n, n) \leqslant \alpha$ 是十分自然的. 例如, 对于 $\alpha = 0.05$ 且 $m = n$, 我们仅需每个随机样本量不小于 4.

在 "固定样本、高维数" 框架下, 我们的检验可以进一步改进. 为了解释, 考虑一个特殊情形: $\sigma_1^2 \geqslant \sigma_2^2 + \nu^2$. 一方面, 注意到 $\sigma_1^2/(\sigma_1^2 + \sigma_2^2 + \nu^2) \geqslant 1/2$, 从而 $T_{1,1} \leqslant \pi/3$. 另一方面, 注意到对于 $2^{-1} \leqslant u \leqslant 2^{-1/2}$, $\arccos(u) \geqslant \arccos(2u^2)$, 从而 $T_{1,1} \geqslant T_{1,2}$. 因此, $T_{1,1}$ 是 $T_{1,2}$ 和 $\pi/3$ 中间的一个量. 显然, γ_0 定义的第 1 项 $\tau(2T_{1,1} - T_{1,2} - \pi/3)$ 小于 $\tau(T_{1,1} - T_{1,2} + |T_{1,1} - \pi/3|)$. 注意到 γ_0 是 p 趋于 ∞ 时的极限. 为了增加所推荐检验的功效, 我们可以修正 $\widehat{T}$ 使得修正后的统计量的首项具有极限 $|T_{1,1} - T_{1,2}| + |T_{1,1} - \pi/3|$ 或 $(T_{1,1} - T_{1,2})^2 + (T_{1,1} - \pi/3)^2$. 注意到随着 $\widehat{T}$ 增大, 我们拒绝 H_0. 这种修正会使得改进后的统计量值变得比 $\widehat{T}$ 更大, 从而增加了检验的功效.

引理 2 $H_0 : F = G$ 成立当且仅当 $T_1 = T_2 = T_3$ 和 $T_4 = T_5 = T_6$.

由引理 2 知, 为了 (1.1), 我们仅需要检验 $T_1 = T_2$ 和 $T_1 = T_3$ 以及 $T_4 = T_5$ 和 $T_4 = T_6$ 是否相等. 基于引理 2, 考虑如下统计量:

$$\widehat{T}^* = \tau\{(\widehat{T}_1^* - \widehat{T}_2^*)^2 + (\widehat{T}_1^* - \widehat{T}_3^*)^2\} + (1-\tau)\{(\widehat{T}_4^* - \widehat{T}_5^*)^2 + (\widehat{T}_4^* - \widehat{T}_6^*)^2\}, \tag{4.1}$$

这里 $\widehat{T}_k^*$ 是 T_k 的无偏估计, 分别定义为

$$\begin{aligned}
\widehat{T}_1^* &\stackrel{\text{def}}{=} \{nm(m-1)\}^{-1}\sum_{i\neq j}^{m}\sum_{k=1}^{n}\operatorname{ang}(\boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{y}_k - \boldsymbol{x}_j),\\
\widehat{T}_2^* &\stackrel{\text{def}}{=} \{m(m-1)(m-2)\}\sum_{i\neq j, i\neq k, j\neq k}^{m}\operatorname{ang}(\boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{x}_k - \boldsymbol{x}_j),\\
\widehat{T}_3^* &\stackrel{\text{def}}{=} \{nm(n-1)\}^{-1}\sum_{i\neq j}^{n}\sum_{k=1}^{m}\operatorname{ang}(\boldsymbol{y}_i - \boldsymbol{x}_k, \boldsymbol{y}_j - \boldsymbol{x}_k),\\
\widehat{T}_4^* &\stackrel{\text{def}}{=} \{mn(n-1)\}^{-1}\sum_{i=1}^{m}\sum_{j\neq k}^{n}\operatorname{ang}(\boldsymbol{x}_i - \boldsymbol{y}_k, \boldsymbol{y}_j - \boldsymbol{y}_k),\\
\widehat{T}_5^* &\stackrel{\text{def}}{=} \{nm(m-1)\}^{-1}\sum_{i,j=1}^{m}\sum_{k=1}^{n}\operatorname{ang}(\boldsymbol{x}_i - \boldsymbol{y}_k, \boldsymbol{x}_j - \boldsymbol{y}_k),\\
\widehat{T}_6^* &\stackrel{\text{def}}{=} \{n(n-1)(n-2)\}\sum_{i\neq j, i\neq k, j\neq k}^{n}\operatorname{ang}(\boldsymbol{y}_i - \boldsymbol{y}_k, \boldsymbol{y}_j - \boldsymbol{y}_k).
\end{aligned}$$

记 $\gamma_0^* \overset{\text{def}}{=} \tau\{(T_{1,1}-T_{1,2})^2+(T_{1,1}-\pi/3)^2\}+(1-\tau)\{(T_{2,1}-T_{2,2})^2+(T_{2,1}-\pi/3)^2\}$. 类似于引理 1, 有如下引理.

引理 3 假设条件 (A1)–(A3) 成立. 当 p 趋于 ∞ 时, 有 $\widehat{T}^* \xrightarrow{\text{pr}} \gamma_0^*$, $\gamma_0^* \geqslant 0$. 更重要的是, $\sigma_1^2=\sigma_2^2$ 和 $\nu^2=0$ 当且仅当 $\gamma_0^*=0$.

根据定义知

$$\widehat{T} \overset{\text{def}}{=} \tau\{(\widehat{T}_1-\widehat{T}_2)+(\widehat{T}_1-\widehat{T}_3)\}+(1-\tau)\{(\widehat{T}_4-\widehat{T}_5)+(\widehat{T}_4-\widehat{T}_6)\}.$$

统计量 $\widehat{T}_1-\widehat{T}_2$、$\widehat{T}_1-\widehat{T}_3$、$\widehat{T}_4-\widehat{T}_5$ 和 $\widehat{T}_4-\widehat{T}_6$ 可以为正, 也可以为负. 不同于统计量 $\widehat{T}$, 修正后的 $\widehat{T}^*$ 中的每一个项都是正的. 下面的定理 7 在 "固定样本、高维数" 框架下说明了基于统计量 $\widehat{T}^*$ 检验的有效性.

定理 7 假设条件 (A1)–(A3) 成立. 当 p 趋于 ∞ 时, 如果 $\nu^2>0$ 或 $\sigma_1^2\neq\sigma_2^2$, 则由排序抽样法所决定的修正检验统计量的极限 p- 值依概率收敛到 $2/C(m+n,n)$.

根据定义, 本文所推荐的统计量 $\widehat{T}$ 和 $\widehat{T}^*$ 的计算复杂度是 $O\{(m+n)^3\}$. 随着计算机计算能力的提升, 这样的复杂度是可以忍受的.

5 数值研究

本节通过数值模拟展示本文所提出检验的有效性. 为了解释方便, 将基于统计量 $\widehat{T}$ 的检验记为 XZ1, 将基于修正检验统计量 $\widehat{T}^*$ 的检验记为 XZ2. 为了比较, 也考虑如下已有的 9 个检验:

(1) 文献 [36] 的 k 临近检验 (Henze, H).

(2) 文献 [24] 的修正的 k 临近检验 (Mondal-Biswas-Ghosh, MBG). 根据文献 [24,36], 考虑 $k=3$.

(3) 文献 [25] 的排序检验 (Hall-Tajvidi, HT). 根据文献 [25], 选择 $\gamma=2$ 和 $w_1(j)=w_2(j)=1$.

(4) 文献 [26] 的精确检验 (Rosenbaum, R). 使用 `R` 软件包 "`crossmatch`" 去计算该检验的 p- 值.

(5) 文献 [27] 的游程检验 (Biswas-Mukhopadhyay-Ghosh, BMG). 使用 `R` 软件包 "`TSP`" 去计算该检验的 p- 值.

(6) 文献 [19] 的检验 (Baringhaus-Franz, BF).

(7) 文献 [20] 的检验 (Biswas-Ghosh, BG).

(8) 文献 [28] 的基于球偏离系数的检验 (Pan-Tian-Wang-Zhang, PTWZ).

(9) 文献 [22] 基于核方法的检验 (kernel maximum mean discrepancy, KMMD). 使用 `R` 软件包中的 "`kmmd`" 函数去计算该检验的 p- 值. 这里选择 "Gaussian" 核.

(10) 文献 [7] 的两样本均值相等检验 (Chang-Zheng-Zhou-Zhou, CZZZ).

(11) 文献 [12] 的两样本协方差相等检验 (Chang-Zhou-Zhou-Wang, CZZW). 使用 `R` 软件包 "`HDtest`" 去实施 CZZZ 检验和 CZZW 检验.

假设样本 $\{\boldsymbol{x}_i, i=1,\ldots,m\}$ 和 $\{\boldsymbol{y}_i, i=1,\ldots,n\}$ 分别来自于 $t_d(\boldsymbol{u}_1,\boldsymbol{\Sigma}_1)$ 和 $t_d(\boldsymbol{u}_2,\boldsymbol{\Sigma}_2)$. 对于 $k=1,2$, $t_d(\boldsymbol{u}_k,\boldsymbol{\Sigma}_k)$ 表示多元 t 分布, 其中, d 为自由度, $\boldsymbol{u}_k$ 是位置参数, $\boldsymbol{\Sigma}_k$ 是尺度参数. 对每一个实验, 我们重复排序 500 次去获得检验的临界值. 重复实验 1,000 次去计算经验检验水平和功效. 显著水平 $\alpha=0.05$. 考虑 3 个模拟的例子和 1 个实例分析. 具体地, 在传统 "大样本、固定维数" 框架下和新型的 "固定样本、高维数" 框架下研究第 1 和第 2 个模拟的例子. 第 3 个模拟的例子主要关心的是非平衡的数据情形.

例 1 在这个例子中, 设置 $m = n = 20$、$p = 10$、$\boldsymbol{u}_1 = \mathbf{0}_{p\times 1}$、$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_{p\times p}$、$\boldsymbol{u}_2 = \delta\mathbf{1}_{p\times 1}$ 和 $\boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{\Sigma}_1$. 对于 (δ, σ^2), 考虑 4 种组合: $(0, 1.0)$、$(0.5, 1.0)$、$(0.3, 2.0)$ 和 $(0, 3.0)$. 它们分别对应于零假设、位置差异、位置尺度差异和尺度差异. 当自由度趋于无穷大时, 多元 t 分布相应于多元正态分布.

表 1 列出了在例 1 中, 当 $m = n = 20$ 和 $p = 10$ 时的经验检验水平和功效. 模拟结果表明所有检验的第一类错误都能够控制得很好, 很合理地在 0.05 左右波动. 关于不犯第二类错误的比较, 即功效比较, 所推荐的 XZ1 检验和 XZ2 检验具有很强的可比性. 对于位置差异, 我们的 XZ1 检验的功效在大多数情形下是最高的. 对于尺度差异和位置尺度差异, 所提出的修正 XZ2 检验的功效在大多数情形下是最好的. PTWZ 检验能够很好探测尺度差异; 但对于位置差异, PTWZ 方法的检验功效变低. 对于正态情形, BG 检验和 BF 检验具有令人满意的功效; 然而对于厚尾数据, 这两种方法的检验功效表现在减小. 由于 R 检验、BMG 检验、HT 检验、H 检验、MBG 检验和 KMMD 检验会受到相关冗余参数的影响, 在许多情形下, 它们的检验功效比 XZ1 检验、XZ2 检验、PTWZ 检验、BG 检验和 BF 检验的功效低. 另外, CZZZ 方法无法探测尺度差异, 而 CZZW 方法不能够探测位置差异, 这并不奇怪, 因为它们仅对两样本的一阶矩或二阶矩差异有功效, 并不能完全刻画两样本间的分布差异.

例 2 固定 $m = n = 20$, p 分别取 30、60、90、150 和 200. 设置 $\boldsymbol{u}_1 = \mathbf{0}_{p\times 1}$、$\boldsymbol{\Sigma}_1 = (0.5^{|k-l|})_{p\times p}$、$\boldsymbol{u}_2 = \delta\mathbf{1}_{p\times 1}$ 和 $\boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{\Sigma}_1$. 为了评价检验水平和功效, 考虑如下的零假设和备择假设: (δ, σ^2) 分别为 $(0, 1.0)$、$(0.25, 1.0)$、$(0.15, 2.0)$ 和 $(0, 2.5)$. 自由度 d 分别设置为 3、5 和 ∞.

对于不同的自由度 d 和维数 p, 表 2–4 列出了各检验的经验检验水平和功效. 总体而言, 所有检验的模拟显著水平都较为合理地接近 0.05, 这说明各检验具有很好地控制第一类错误的能力. 对于功效比较, 当维数 p 增加时, 所有的检验功效都在提高. 这是因为随着维数的增加, 两总体分布差异性在变大. 在位置差异备择下, XZ1 检验的功效比 XZ2 检验的功效高, 且几乎是所有检验中最好的. 在尺度差异备择下, XZ2 检验的功效比 XZ1 检验的功效高, 且几乎是所有检验中最好的. 从表 2–4 也可观测到 PTWZ 检验、BG 检验和 BF 检验在探测位置差异时的表现相对较弱. 或许是因为 R 检验、BMG 检验、HT 检验、H 检验、MBG 检验和 KMMD 检验的功效与最优的冗余参数选择有关, 对于一些备择, 它们的检验功效会表现不尽人意. 再一次, 我们有理由观测到 CZZZ 检验在尺度差异的情况下失去功效, 而 CZZW 检验在位置差异的情况下失去功效.

例 3 固定 $n = 10$ 和 $p = 100$, 取 $m = 20$ 或 $m = 60$. 这将产生非平衡的样本. 类似于例 1, 设置 $\boldsymbol{u}_1 = \mathbf{0}_{p\times 1}$、$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_{p\times p}$、$\boldsymbol{u}_2 = \delta\mathbf{1}_{p\times 1}$ 和 $\boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{\Sigma}_1$. 对于 (δ, σ^2), 考虑 4 种组合: $(0, 1.0)$、$(0.5, 1.0)$、$(0.3, 2.0)$ 和 $(0, 3.0)$. 自由度 d 设置为 3. 模拟结果见表 5.

从表 5 的模拟结果中可以得出与上述例子类似的结论. 所推荐的 XZ1 检验、BF 检验、KMMD

表 1 在例 1 中, 当 $m = n = 20$ 和 $p = 10$ 时的经验检验水平和功效比较

	t_d	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
$(\delta, \sigma^2) = (0, 1.0)$	t_∞	0.046	0.044	0.056	0.062	0.040	0.050	0.042	0.064	0.068	0.041	0.056	0.069	0.066
	t_3	0.060	0.048	0.041	0.067	0.036	0.050	0.060	0.048	0.046	0.050	0.045	0.073	0.049
$(\delta, \sigma^2) = (0.5, 1.0)$	t_∞	0.958	0.954	0.755	0.438	0.362	0.450	0.964	0.598	0.694	0.456	0.812	0.673	0.066
	t_3	0.814	0.560	0.179	0.372	0.252	0.068	0.703	0.136	0.642	0.334	0.588	0.502	0.049
$(\delta, \sigma^2) = (0.3, 2.0)$	t_∞	0.574	0.990	0.992	0.194	0.156	0.998	0.694	0.990	0.420	0.916	0.575	0.488	0.246
	t_3	0.596	0.672	0.603	0.162	0.171	0.388	0.622	0.420	0.500	0.554	0.490	0.390	0.145
$(\delta, \sigma^2) = (0, 3.0)$	t_∞	0.542	1.000	1.000	0.134	0.327	1.000	0.858	1.000	0.536	0.998	0.790	0.065	0.546
	t_3	0.384	0.893	0.879	0.124	0.140	0.716	0.602	0.728	0.372	0.814	0.541	0.060	0.351

表 2　在例 2 中, 当 $d=\infty$ 时的经验检验水平和功效比较

	p	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
$(\delta,\sigma^2)=(0,1.0)$	30	0.044	0.048	0.054	0.072	0.039	0.042	0.046	0.046	0.052	0.042	0.046	0.040	0.054
	90	0.046	0.046	0.046	0.070	0.036	0.052	0.044	0.048	0.042	0.060	0.050	0.056	0.052
	150	0.044	0.042	0.049	0.070	0.039	0.056	0.042	0.044	0.044	0.046	0.052	0.040	0.057
	200	0.048	0.050	0.051	0.068	0.037	0.060	0.048	0.054	0.040	0.048	0.047	0.042	0.060
$(\delta,\sigma^2)=(0.25,1.0)$	30	0.456	0.424	0.252	0.138	0.108	0.092	0.460	0.186	0.216	0.072	0.379	0.418	0.054
	90	0.796	0.784	0.496	0.262	0.158	0.176	0.804	0.302	0.428	0.148	0.710	0.756	0.052
	150	0.924	0.918	0.592	0.318	0.280	0.286	0.922	0.390	0.548	0.170	0.807	0.898	0.057
	200	0.958	0.950	0.658	0.392	0.330	0.342	0.950	0.442	0.664	0.232	0.862	0.956	0.060
$(\delta,\sigma^2)=(0.15,2.0)$	30	0.418	1.000	1.000	0.122	0.158	1.000	0.602	1.000	0.202	1.000	0.569	0.106	0.264
	90	0.804	1.000	1.000	0.148	0.284	1.000	0.952	1.000	0.010	1.000	0.864	0.160	0.358
	150	0.944	1.000	1.000	0.140	0.440	1.000	0.996	1.000	0.000	1.000	0.950	0.200	0.402
	200	0.980	1.000	1.000	0.158	0.548	1.000	1.000	1.000	0.000	1.000	1.000	0.250	0.446
$(\delta,\sigma^2)=(0,2.5)$	30	0.578	1.000	1.000	0.130	0.314	1.000	0.850	1.000	0.174	1.000	0.806	0.044	0.422
	90	0.954	1.000	1.000	0.122	0.604	1.000	1.000	1.000	0.000	1.000	1.000	0.054	0.476
	150	0.996	1.000	1.000	0.110	0.780	1.000	1.000	1.000	0.000	1.000	1.000	0.038	0.526
	200	1.000	1.000	1.000	0.112	0.904	1.000	1.000	1.000	0.000	1.000	1.000	0.036	0.602

表 3　在例 2 中, 当 $d=5$ 时的经验检验水平和功效比较

	p	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
$(\delta,\sigma^2)=(0,1.0)$	30	0.060	0.052	0.048	0.076	0.032	0.048	0.056	0.034	0.068	0.060	0.042	0.034	0.076
	90	0.056	0.043	0.065	0.062	0.036	0.062	0.058	0.052	0.056	0.062	0.048	0.050	0.084
	150	0.055	0.059	0.040	0.074	0.036	0.050	0.044	0.054	0.060	0.045	0.056	0.044	0.082
	200	0.062	0.058	0.056	0.072	0.044	0.058	0.062	0.058	0.062	0.058	0.053	0.056	0.096
$(\delta,\sigma^2)=(0.25,1.0)$	30	0.236	0.114	0.062	0.122	0.070	0.048	0.212	0.040	0.192	0.088	0.207	0.132	0.076
	90	0.522	0.104	0.060	0.254	0.110	0.052	0.424	0.058	0.346	0.060	0.295	0.288	0.084
	150	0.698	0.096	0.054	0.350	0.136	0.050	0.554	0.060	0.444	0.048	0.421	0.430	0.082
	200	0.758	0.110	0.063	0.428	0.188	0.060	0.594	0.058	0.466	0.072	0.503	0.496	0.096
$(\delta,\sigma^2)=(0.15,2.0)$	30	0.436	0.800	0.772	0.098	0.120	0.686	0.556	0.584	0.250	0.764	0.445	0.074	0.146
	90	0.768	0.875	0.823	0.148	0.202	0.722	0.810	0.642	0.198	0.768	0.708	0.122	0.174
	150	0.832	0.904	0.860	0.172	0.242	0.716	0.848	0.622	0.208	0.744	0.725	0.132	0.194
	200	0.860	0.922	0.866	0.180	0.278	0.692	0.856	0.644	0.146	0.756	0.762	0.178	0.240
$(\delta,\sigma^2)=(0,2.5)$	30	0.512	0.946	0.936	0.096	0.144	0.888	0.688	0.812	0.238	0.912	0.549	0.038	0.188
	90	0.810	0.974	0.967	0.114	0.240	0.896	0.890	0.862	0.102	0.932	0.772	0.046	0.226
	150	0.856	0.958	0.972	0.116	0.276	0.908	0.906	0.844	0.052	0.918	0.789	0.048	0.238
	200	0.878	0.978	0.976	0.138	0.300	0.888	0.902	0.880	0.038	0.938	0.801	0.062	0.282

检验和 H 检验都能很好地应对非平衡的样本情形. 我们的 XZ1 检验几乎在所有的情形下表现得都很优秀. 对于尺度差异备择, PTWZ 方法表现出比较不错的功效表现. 但对于位置差异备择, PTWZ 检验、BG 检验和 CZZW 检验具有明显的劣势.

表 4 在例 2 中, 当 $d=3$ 时的经验检验水平和功效比较

	p	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
$(\delta,\sigma^2)=(0,1.0)$	30	0.062	0.054	0.037	0.042	0.030	0.060	0.056	0.042	0.060	0.066	0.049	0.046	0.042
	90	0.046	0.048	0.062	0.074	0.035	0.042	0.046	0.056	0.050	0.054	0.055	0.046	0.090
	150	0.060	0.054	0.057	0.064	0.047	0.070	0.066	0.060	0.052	0.060	0.053	0.050	0.102
	200	0.052	0.056	0.049	0.075	0.037	0.068	0.056	0.058	0.044	0.046	0.047	0.044	0.094
$(\delta,\sigma^2)=(0.25,1.0)$	30	0.230	0.100	0.064	0.156	0.066	0.058	0.166	0.050	0.182	0.068	0.164	0.088	0.042
	90	0.354	0.070	0.068	0.216	0.076	0.052	0.210	0.058	0.280	0.064	0.173	0.164	0.090
	150	0.534	0.098	0.080	0.302	0.084	0.074	0.266	0.066	0.416	0.076	0.215	0.230	0.102
	200	0.610	0.090	0.078	0.424	0.148	0.068	0.318	0.062	0.404	0.070	0.282	0.268	0.094
$(\delta,\sigma^2)=(0.15,2.0)$	30	0.360	0.592	0.508	0.094	0.122	0.394	0.430	0.372	0.226	0.550	0.371	0.060	0.076
	90	0.556	0.674	0.612	0.110	0.134	0.390	0.550	0.382	0.202	0.568	0.463	0.098	0.100
	150	0.660	0.698	0.655	0.136	0.168	0.404	0.594	0.388	0.222	0.544	0.505	0.104	0.116
	200	0.694	0.726	0.681	0.178	0.164	0.432	0.620	0.364	0.214	0.540	0.532	0.094	0.142
$(\delta,\sigma^2)=(0,2.5)$	30	0.424	0.790	0.756	0.086	0.124	0.594	0.552	0.616	0.236	0.708	0.459	0.048	0.114
	90	0.624	0.812	0.834	0.102	0.135	0.612	0.688	0.618	0.124	0.766	0.557	0.052	0.124
	150	0.702	0.818	0.842	0.110	0.187	0.620	0.704	0.598	0.104	0.740	0.574	0.062	0.144
	200	0.726	0.844	0.851	0.148	0.165	0.632	0.720	0.618	0.105	0.784	0.605	0.044	0.162

表 5 在例 3 中, 当 $d=3$, $p=100$, $n=10$, $m=20$ 或 $m=60$ 时的经验检验水平和功效比较

	m	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
$(\delta,\sigma^2)=(0,1.0)$	20	0.053	0.060	0.056	0.071	0.037	0.063	0.060	0.054	0.047	0.060	0.060	0.072	0.069
	60	0.048	0.055	0.049	0.082	0.044	0.052	0.057	0.061	0.044	0.053	0.057	0.074	0.058
$(\delta,\sigma^2)=(0.5,1.0)$	20	1.000	0.650	0.124	0.908	0.870	0.092	0.954	0.066	0.914	0.254	0.812	0.901	0.069
	60	1.000	0.574	0.100	0.911	0.966	0.070	1.000	0.080	0.911	0.250	0.947	1.000	0.058
$(\delta,\sigma^2)=(0.3,2.0)$	20	0.982	0.403	0.477	0.436	0.525	0.360	0.896	0.126	0.842	0.218	0.890	0.784	0.241
	60	0.995	0.150	0.548	0.530	0.695	0.417	0.989	0.044	0.770	0.013	0.963	0.899	0.322
$(\delta,\sigma^2)=(0,3.0)$	20	0.783	0.544	0.806	0.196	0.184	0.601	0.708	0.370	0.654	0.426	0.737	0.063	0.395
	60	0.938	0.332	0.975	0.230	0.267	0.739	0.850	0.112	0.711	0.010	0.865	0.070	0.586

例 4 在该例中, 我们使用两个实例数据集: 电离层数据集和克隆数据集, 去展示所推荐方法的实际表现. 电离层数据集记录了 126 个好的雷达和 225 个坏的雷达所返回的数据, 数据的维数是 34. 克隆数据集是一个基因表达的高维数据, 记录了 22 个正常组织细胞和 40 个克隆的非正常癌组织细胞. 每个组织细胞包含了 2,000 个基因表达. 关于这两个数据集更详细的描述, 可参见 R 软件包 `dprep` 及网址 http://www.ics.uci.edu/~mlearn/MLRepository.html 和 http://microarray.princeton.edu/oncology/.

在有监督分类推断中, 这两个数据集已得到广泛的研究. 相关研究已表明这两个数据集分别来自不同的两个总体 (参见文献 [27]). 基于所有样本, 我们也计算出所推荐的检验 p- 值本质上为 0, 这也支持已有的发现. 换句话说, 在这两个数据集中, 备择假设为真. 为了功效比较, 随机地从所有观测样本中选出 N 个观测子样本, 即 $n=m=N$. 重复该随机抽样 1,000 次, 计算所有检验的经验功效. 模拟结果见表 6. 从表 6 中的模拟结果可以看到, 对于电离层数据集, 所推荐的 XZ2 检验功效表现最好, 紧接着是 MBG 检验和 BG 检验, 其余检验也有可比较的表现. 对于克隆数据集, 所推荐的 XZ1 检验

表 6　在电离层数据集和克隆数据集分析中, 当 $\alpha = 0.05$ 时的经验检验功效比较

	N	XZ1	XZ2	PTWZ	R	BMG	BG	BF	HT	H	MBG	KMMD	CZZZ	CZZW
电离层数据集	6	0.358	0.412	0.352	0.215	0.367	0.446	0.361	0.395	0.334	0.432	0.250	0.343	0.401
	7	0.443	0.542	0.452	0.269	0.445	0.528	0.464	0.447	0.391	0.502	0.371	0.405	0.448
	8	0.486	0.671	0.514	0.307	0.525	0.618	0.513	0.406	0.473	0.595	0.445	0.479	0.576
	9	0.612	0.787	0.606	0.338	0.686	0.689	0.634	0.495	0.584	0.677	0.598	0.590	0.651
	10	0.694	0.831	0.639	0.391	0.753	0.726	0.710	0.544	0.642	0.748	0.645	0.669	0.711
	11	0.725	0.860	0.657	0.426	0.772	0.738	0.742	0.498	0.712	0.788	0.682	0.730	0.822
克隆数据集	6	0.283	0.261	0.137	0.086	0.104	0.115	0.244	0.121	0.250	0.113	0.288	0.372	0.097
	9	0.667	0.494	0.286	0.145	0.267	0.172	0.477	0.080	0.541	0.246	0.416	0.688	0.136
	12	0.896	0.705	0.379	0.221	0.383	0.243	0.755	0.110	0.833	0.542	0.733	0.912	0.130
	15	0.965	0.823	0.425	0.279	0.483	0.235	0.863	0.086	0.961	0.755	0.825	0.990	0.154
	18	0.995	0.967	0.607	0.325	0.697	0.387	0.982	0.103	1.000	0.964	0.886	0.997	0.173
	21	1.000	1.000	0.778	0.414	0.774	0.395	1.000	0.107	1.000	1.000	0.963	1.000	0.265

和 CZZZ 检验的功效比其他检验要高, 尽管 H 检验、XZ2 检验、BF 检验和 KMMD 检验也有令人满意的功效表现.

致谢　感谢审稿人对本文提出的许多优秀的修改意见.

参考文献

1 Lehmann E L, Romano J P. Testing Statistical Hypotheses, 3rd ed. New York: Springer, 2005
2 Thas O. Comparing Distributions. New York: Springer, 2010
3 Anderson T W. An Introduction to Multivariate Statistical Analysis, 3rd ed. New York: Wiley, 2003
4 Bai Z, Sarandasa H. Effect of high dimension: By an example of a two sample problem. Statist Sinica, 1996, 6: 311–329
5 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
6 Cai T T, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. J R Stat Soc Ser B Stat Methodol, 2014, 76: 349–372
7 Chang J, Zheng C, Zhou W X, et al. Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. Biometrics, 2017, 73: 1300–1310
8 Schott J R. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Comput Statist Data Anal, 2007, 51: 6535–6542
9 Li J, Chen S X. Two sample tests for high-dimensional covariance matrices. Ann Statist, 2012, 40: 908–940
10 Cai T T, Liu W, Xia Y. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. J Amer Statist Assoc, 2013, 108: 265–277
11 Cai T T, Liu W. Large-scale multiple testing of correlations. J Amer Statist Assoc, 2016, 111: 229–240
12 Chang J, Zhou W, Zhou W X, et al. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. Biometrics, 2017, 73: 31–41
13 Smirnov N V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Moscow Univ Math Bull, 1939, 2: 3–14
14 Anderson T W. On the distribution of the two-sample Cramér-von Mises criterion. Ann of Math Stud, 1962, 33: 1148–1159
15 Rosenblatt M. Limit theorems associated with variants of the von Mises statistic. Ann of Math Stud, 1952, 23: 617–623
16 Escanciano J C. A consistent diagnostic test for regression models using projections. Econometric Theory, 2006, 22: 1030–1051
17 Anderson N H, Hall P, Titterington D M. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. J Multivariate Anal, 1994, 50: 41–54
18 Zhou W X, Zheng C, Zhang Z. Two-sample smooth tests for the equality of distributions. Bernoulli, 2017, 23: 951–989

19 Baringhaus L, Franz C. On a new multivariate two-sample test. J Multivariate Anal, 2004, 88: 190–206
20 Biswas M, Ghosh A K. A nonparametric two-sample test applicable to high dimensional data. J Multivariate Anal, 2014, 123: 160–171
21 Hall P, Marron J S, Neeman A. Geometric representation of high dimension, low sample size data. J R Stat Soc Ser B Stat Methodol, 2005, 67: 427–444
22 Gretton A, Borgwardt K, Rasch M, et al. A kernel two sample test. J Mach Learn Res, 2012, 13: 723–773
23 Friedman J H, Rafsky L C. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann Statist, 1979, 7: 697–717
24 Mondal P K, Biswas M, Ghosh A K. On high dimensional two-sample tests based on nearest neighbors. J Multivariate Anal, 2015, 141: 168–178
25 Hall P, Tajvidi N. Permutation tests for equality of distributions in high-dimensional settings. Biometrika, 2002, 89: 359–374
26 Rosenbaum P R. An exact distribution-free test comparing two multivariate distributions based on adjacency. J R Stat Soc Ser B Stat Methodol, 2005, 67: 515–530
27 Biswas M, Mukhopadhyay M, Ghosh A K. A distribution-free two-sample run test applicable to high-dimensional data. Biometrika, 2014, 101: 913–926
28 Pan W, Tian Y, Wang X, et al. Ball divergence: Nonparametric two sample test. Ann Statist, 2018, 46: 1109–1137
29 Cui H. Average projection type weighted Cramér-von Mises statistics for testing some distributions. Sci China Ser A, 2002, 45: 562–577
30 Zhu L X, Fang K T, Bhatti M I. On estimated projection pursuit-type Crámer-von Mises statistics. J Multivariate Anal, 1997, 63: 1–14
31 Zhu L, Xu K, Li R, et al. Projection correlation between two random vectors. Biometrika, 2017, 104: 829–843
32 Chikkagoudar M S, Bhat B V. Limiting distribution of two-sample degenerate U-statistic under contiguous alternatives and applications. J Appl Stat Sci, 2016, 22: 127–139
33 Baringhaus L, Franz C. Rigid motion invariant two-sample tests. Statist Sinica, 2010, 20: 1333–1361
34 Diaconis P, Freedman D. Asymptotics of graphical projection pursuit. Ann Statist, 1984, 12: 793–815
35 Hall P, Li K C. On almost linearity of low dimensional projections from high dimensional data. Ann Statist, 1993, 21: 867–889
36 Henze N. A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann Statist, 1988, 16: 772–783
37 Dudley R M. Central limit theorems for empirical measures. Ann Probab, 1978, 6: 899–929
38 van der Vaart A W, Wellner J A. Weak Convergence and Empirical Processes. New York: Springer, 1996
39 Chang M N. Weak convergence of a self-consistent estimator of the survival function with doubly censored data. Ann Statist, 1990, 18: 391–404
40 Lee A J. U-statistics: Theory and Practice. Statistics: Textbooks and Monographs, vol. 110. Boca Raton-London-New York: CRC Press, 1990
41 Serfling R L. Approximation Theorems in Mathematical Statistics. New York: Wiley, 1980

附录 A

附录 A 包含了定理 1–7 和引理 1–4 的证明. 由于定理 1 的证明类似于定理 2, 为了避免重复, 我们省略定理 1 的证明.

定理 2 的证明 定义

$$I_1 \overset{\text{def}}{=} m^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_j \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha},$$

$$I_2 \overset{\text{def}}{=} n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_j \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha},$$

$$I_3 \overset{\text{def}}{=} m^{-1}n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}} I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_j \leqslant \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha}.$$

通过简单的代数运算, 有

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\left\{m^{-1}\sum_{i=1}^{m}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)-n^{-1}\sum_{i=1}^{n}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)\right\}^2 d\boldsymbol{\alpha}=:I_1+I_2-2I_3.$$

由文献 [31, 引理 1] 知

$$\begin{aligned}
I_1&=c_pm^{-2}\sum_{i=1}^{m}\sum_{j=1}^{m}\{\pi-\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_r,\boldsymbol{x}_j-\boldsymbol{x}_r)\},\\
I_2&=c_pn^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\pi-\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{x}_r,\boldsymbol{y}_j-\boldsymbol{x}_r)\},\\
I_3&=c_pm^{-1}n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}[\pi-\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_r,\boldsymbol{y}_j-\boldsymbol{x}_r)].
\end{aligned}$$

从而,

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{\widehat{F}_{\boldsymbol{\alpha}}(t)-\widehat{G}_{\boldsymbol{\alpha}}(t)\}^2d\widehat{F}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}=c_p(2\widehat{T}_1-\widehat{T}_2-\widehat{T}_3).$$

类似地, 可以证明

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{\widehat{F}_{\boldsymbol{\alpha}}(t)-\widehat{G}_{\boldsymbol{\alpha}}(t)\}^2d\widehat{F}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}=c_p(2\widehat{T}_4-\widehat{T}_5-\widehat{T}_6).$$

联合上述两个结果, 可得

$$\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{\widehat{F}_{\boldsymbol{\alpha}}(t)-\widehat{G}_{\boldsymbol{\alpha}}(t)\}^2d\widehat{H}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}=c_p\widehat{T},$$

且 $\widehat{T}$ 是非负的. □

定理 3 的证明 首先证明定理的第一部分. 定义经验过程

$$U_m(\boldsymbol{\alpha},t)=m^{-1/2}\sum_{i=1}^{m}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant t)-\mathrm{pr}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}\leqslant t)\},\quad(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}$$

和

$$V_n(\boldsymbol{\alpha},t)=n^{-1/2}\sum_{i=1}^{n}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant t)-\mathrm{pr}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}\leqslant t)\},\quad(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}.$$

考虑加权的经验过程

$$\begin{aligned}
W_{m,n}(\boldsymbol{\alpha},t)&=(1-\tau)^{1/2}U_m(\boldsymbol{\alpha},t)-\tau^{1/2}V_n(\boldsymbol{\alpha},t)\\
&=\left\{\frac{mn}{m+n}\right\}^{1/2}\left\{m^{-1}\sum_{i=1}^{m}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant t)-n^{-1}\sum_{i=1}^{n}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant t)\right\}.
\end{aligned}$$

由定理 2 知统计量 $\{mn/(m+n)\}\widehat{T}$ 可等价表示为

$$c_p^{-1}\left[\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{W_{m,n}(\boldsymbol{\alpha},t)\}^2d\widehat{H}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}\right].$$

定义函数类 $\mathcal{F}\stackrel{\text{def}}{=}\{I(\boldsymbol{\alpha}^{\mathrm{T}}\cdot\leqslant t):(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$. 根据文献 [37,38] 知, $\mathcal{F}$ 是一个 Vapnik-Červonenkis 类且满足 Donsker 条件. 在 H_0 下, 即 $F=G=H$, 经验过程 $\{U_m(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$ 和 $\{V_n(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$ 依分布都收敛到随机过程 $\{B_H(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$, 其中 $B_H(\boldsymbol{\alpha},t)$ 是 H-Brown 桥, 即一个均值为 0、协方差为

$$\mathrm{cov}\{B_H(\boldsymbol{\alpha},t),B_H(\boldsymbol{\beta},s)\}=\mathrm{pr}_H(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}\leqslant t,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}\leqslant s)-\mathrm{pr}_H(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}\leqslant t)\mathrm{pr}_H(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}\leqslant s)$$

的 Gauss 过程.

因而, 经验过程 $\{W_{m,n}(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$ 依分布收敛到随机过程 $\{(1-\tau_0)^{1/2}B_H(\boldsymbol{\alpha},t)+\tau_0^{1/2}\widetilde{B}_H(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$, 其中 $\widetilde{B}_H(\cdot,\cdot)$ 是独立于 Gauss 过程 $B_H(\cdot,\cdot)$ 的 H-Brown 桥. 由 Gauss 过程的独立性知, 随机过程 $\{(1-\tau_0)^{1/2}B_H(\boldsymbol{\alpha},t)+\tau_0^{1/2}\widetilde{B}_H(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$ 与 Gauss 过程 $\{B_H(\boldsymbol{\alpha},t);(\boldsymbol{\alpha},t)\in\mathcal{S}^{p-1}\times\mathbb{R}\}$ 具有相同分布. 由 Glivenko-Cantelli 定理知, 当 $\tau\to\tau_0$ 时, 经验过程 $\widehat{H}_{\boldsymbol{\alpha}}(\cdot)$ 几乎处处收敛到随机过程 $H_{\boldsymbol{\alpha}}(\cdot)$. 结合连续映射定理 [39] 知, 对每一个实值的 a, 有

$$\widehat{T}_a=c_p^{-1}\bigg[\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-a}^{+a}\{W_{m,n}(\boldsymbol{\alpha},t)\}^2d\widehat{H}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}\bigg].$$

依分布收敛到

$$T_a=c_p^{-1}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-a}^{+a}B_H^2(\boldsymbol{\alpha},t)dH_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}.$$

另一方面, 注意到 $c_p^{-1}\mathrm{E}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}B_H^2(\boldsymbol{\alpha},t)dH_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}$ 等于

$$\begin{aligned}
&c_p^{-1}\tau\mathrm{E}\bigg[\mathrm{E}\bigg\{\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha}\;\bigg|\;\boldsymbol{x}_r\bigg\}\\
&\quad-\mathrm{E}\bigg\{\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_j\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha}\;\bigg|\;\boldsymbol{x}_r\bigg\}\bigg]\\
&\quad+c_p^{-1}(1-\tau)\mathrm{E}\bigg[\mathrm{E}\bigg\{\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r)d\boldsymbol{\alpha}\;\bigg|\;\boldsymbol{x}_r\bigg\}\\
&\quad-\mathrm{E}\bigg\{\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_j\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r)d\boldsymbol{\alpha}\;\bigg|\;\boldsymbol{y}_r\bigg\}\bigg].
\end{aligned}$$

根据事实

$$\begin{aligned}
&\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha}=\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r)d\boldsymbol{\alpha}=c_p\pi,\\
&\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_j\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}_r)d\boldsymbol{\alpha}=c_p\{\pi-\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_r,\boldsymbol{x}_j-\boldsymbol{x}_r)\},\\
&\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}I(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_i\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_j\leqslant\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y}_r)d\boldsymbol{\alpha}=c_p\{\pi-\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{y}_r,\boldsymbol{y}_j-\boldsymbol{y}_r)\},
\end{aligned}$$

上式进一步等于

$$S\stackrel{\text{def}}{=}\tau T_2+(1-\tau)T_6.$$

显然, $0\leqslant S\leqslant\pi<\infty$. 因此, T_∞ 几乎处处有限. 进一步, $\widehat{T}_a$ 是关于 a 的单调增函数. 因而, 由单调收敛定理知, $mn\{(m+n)\}^{-1}\widehat{T}$ 依分布收敛到

$$c_p^{-1}\bigg[\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{B_H(\boldsymbol{\alpha},t)\}^2dH_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}\bigg].$$

在 H_0 下, $\widehat{T}$ 是一个退化的 V- 统计量. 通过退化的 V- 统计量的谱分解知, 存在无穷多个非负的常数 λ_k^* 使得 $mn\{(m+n)\}^{-1}\widehat{T}$ 依分布收敛到

$$\sum_{k=1}^{\infty}\lambda_k^* Z_k^2, \tag{A.1}$$

这里 Z_k 是独立同分布的标准正态随机序列, λ_k^* 依赖于未知分布 $F=G$. 接下来, 计算 $\sum_{k=1}^{\infty}\lambda_k^*$. 因为 $c_p^{-1}\int_{\|\boldsymbol{\alpha}\|=1}\int_{-\infty}^{+\infty}B_H^2(\boldsymbol{\alpha},t)dHd\boldsymbol{\alpha}$ 和 $\sum_{k=1}^{\infty}\lambda_k^* Z_k^2$ 具有相同分布, 所以

$$\mathrm{E}\Big\{c_p^{-1}\int_{\|\boldsymbol{\alpha}\|=1}\int_{-\infty}^{+\infty}B_H^2(\boldsymbol{\alpha},t)dHd\boldsymbol{\alpha}\Big\}=\mathrm{E}\Big(\sum_{k=1}^{\infty}\lambda_k^* Z_k^2\Big)=\sum_{j=1}^{\infty}\lambda_j^*.$$

因此, $\sum_{j=1}^{\infty}\lambda_j^*=S<\infty$. 由强大数定理和 Slutsky 定理, 我们完成定理第一部分的证明.

下面证明第二部分. 定义

$$\begin{aligned}&\psi(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2},\boldsymbol{x}_{i_3};\boldsymbol{y}_{j_1},\boldsymbol{y}_{j_2},\boldsymbol{y}_{j_3})\\&\quad=\tau_0\{2\mathrm{ang}(\boldsymbol{x}_{i_1}-\boldsymbol{x}_{i_3},\boldsymbol{y}_{j_2}-\boldsymbol{x}_{i_3})-\mathrm{ang}(\boldsymbol{x}_{i_1}-\boldsymbol{x}_{i_3},\boldsymbol{x}_{i_2}-\boldsymbol{x}_{i_3})-\mathrm{ang}(\boldsymbol{y}_{j_1}-\boldsymbol{x}_{i_3},\boldsymbol{y}_{j_2}-\boldsymbol{x}_{i_3})\}\\&\qquad+(1-\tau_0)\{2\mathrm{ang}(\boldsymbol{x}_{i_1}-\boldsymbol{y}_{j_3},\boldsymbol{y}_{j_2}-\boldsymbol{y}_{j_3})-\mathrm{ang}(\boldsymbol{x}_{i_1}-\boldsymbol{y}_{j_3},\boldsymbol{x}_{i_2}-\boldsymbol{y}_{j_3})-\mathrm{ang}(\boldsymbol{y}_{j_1}-\boldsymbol{y}_{j_3},\boldsymbol{y}_{j_2}-\boldsymbol{y}_{j_3})\}.\end{aligned}$$

令 $\mathcal{P}_3$ 表示由 1、2 和 3 所有排序构成的集合. 对 ψ 进行对称化, 即

$$h_\psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3;\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3)=36^{-1}\sum_{(i_1,i_2,i_3)\in\mathcal{P}_3}\sum_{(j_1,j_2,j_3)\in\mathcal{P}_3}\psi(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2},\boldsymbol{x}_{i_3};\boldsymbol{y}_{j_1},\boldsymbol{y}_{j_2},\boldsymbol{y}_{j_3}).$$

注意到 $\widehat{T}$ 也是 $\mathrm{E}\{h_\psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3;\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3)\}$ 的 V 型估计量. 利用标准非退化的 V 型统计量和 U 型统计量理论 [40, 41], 知 $\mathrm{var}^{-1/2}(\widehat{T})(\widehat{T}-T)\xrightarrow{d}N(0,1)$ 且 $\mathrm{var}(\widehat{T})=\{1+o(1)\}\{9n^{-1}\sigma_{01}^2+9m^{-1}\sigma_{10}^2\}$, 其中,

$$\sigma_{01}^2=\mathrm{var}[\mathrm{E}\{h_\psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3;\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3)\mid\boldsymbol{y}_1\}], \tag{A.2}$$

$$\sigma_{10}^2=\mathrm{var}[\mathrm{E}\{h_\psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3;\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3)\mid\boldsymbol{x}_1\}]. \tag{A.3}$$

根据 V- 统计量的强大数定理知, $\widehat{T}$ 几乎处处收敛到

$$\tau_0(2T_1-T_2-T_3)+(1-\tau_0)(2T_4-T_5-T_6).$$

由定理 1 知, 在 H_1 下, $2T_1-T_2-T_3>0$ 且 $2T_4-T_5-T_6>0$. 因此, 由 Slutsky 定理并结合 $mn/(m+n)\to\infty$ 和 $\widehat{S}$ 几乎处处收敛到 $S>0$ 的事实, 我们完成定理第二部分的证明. □

命题 1 的证明 定义

$$\widetilde{T}:=\tau(2T_1^*-T_2^*-T_3^*)+(1-\tau)(2T_4^*-T_5^*-T_6^*),$$

其中

$$T_1^*=\{nm(m-1)\}^{-1}\sum_{i\neq j}^{m}\sum_{k=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{y}_k-\boldsymbol{x}_j),$$

$$\widehat{T}_2^*=\{m(m-1)(m-2)\}\sum_{i\neq j,i\neq k,j\neq k}^{m}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{x}_k-\boldsymbol{x}_j),$$

$$\widehat{T}_3^* = \{nm(n-1)\}^{-1}\sum_{i\neq j}^{n}\sum_{k=1}^{m}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{x}_k,\boldsymbol{y}_j-\boldsymbol{x}_k),$$

$$\widehat{T}_4^* = \{mn(n-1)\}-1\sum_{i=1}^{m}\sum_{j\neq k}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{y}_k,\boldsymbol{y}_j-\boldsymbol{y}_k),$$

$$\widehat{T}_5^* = \{nm(m-1)\}^{-1}\sum_{i,j=1}^{m}\sum_{k=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{y}_k,\boldsymbol{x}_j-\boldsymbol{y}_k),$$

$$\widehat{T}_6^* = \{n(n-1)(n-2)\}\sum_{i\neq j,i\neq k,j\neq k}^{n}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{y}_k,\boldsymbol{y}_j-\boldsymbol{y}_k).$$

易知, $\widetilde{T}$ 是 T 的两样本 U 型估计量. 记 $\widehat{\theta}=cn^{-1/2}$. 在条件 (C1)–(C3) 下, 利用文献 [32, 定理 3.1], 可以证明存在二阶矩有限的随机变量 $\mathcal{X}$、$\mathcal{Y}$ 和 $\mathcal{Z}$ 使得 $\{mn/(m+n)\}\widetilde{T}$ 依分布收敛到 $c^2\mathcal{X}+c\mathcal{Y}+\mathcal{Z}$, 其中 $\mathcal{X}$ 是非负的且非退化的. 从而, 当相邻备择结构满足 $n^{1/2}\widehat{\theta}=c\to\infty$ 时, 易证 $\{mn/(m+n)\}\widetilde{T}$ 依概率收敛到 ∞. 结合 U 型统计量和 V 型统计量标准理论, $\{mn/(m+n)\}(\widehat{T}-\widetilde{T})$ 依概率收敛到 $\sum_{k=1}^{\infty}\lambda_k$. 由 Slutsky 定理可以得出命题成立. □

定理 4 的证明 记 $H_\tau=\tau F+(1-\tau)G$ 和 $H_{\boldsymbol{\alpha},0}=\tau_0F_{\boldsymbol{\alpha}}+(1-\tau_0)G_{\boldsymbol{\alpha}}$, 其中 τ_0 是 τ 的极限值. 重复定理 2 和 3 的证明知, 给定样本 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ 条件下, $\widehat{T}^b$ 依分布收敛到

$$c_p^{-1}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{(1-\tau)^{1/2}B_{H_\tau}(\boldsymbol{\alpha},t)-\tau^{1/2}\widetilde{B}_{H_\tau}\}^2(\boldsymbol{\alpha},t)d\widehat{H}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}.$$

注意到 $(1-\tau_0)^{1/2}B_{H_{\tau_0}}-\tau_0^{1/2}\widetilde{B}_{H_{\tau_0}}$ 本身也是 H_{τ_0}-Brown 桥. 根据 Glivenko-Cantelli 定理知

$$c_p^{-1}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}\{(1-\tau)^{1/2}B_{H_\tau}(\boldsymbol{\alpha},t)-\tau^{1/2}\widetilde{B}_{H_\tau}\}^2(\boldsymbol{\alpha},t)d\widehat{H}_{\boldsymbol{\alpha}}(t)d\boldsymbol{\alpha}$$

依分布进一步收敛到

$$c_p^{-1}\int_{\boldsymbol{\alpha}\in\mathcal{S}^{p-1}}\int_{-\infty}^{+\infty}B^2_{H_{\tau_0}}(\boldsymbol{\alpha},t)dH_{\boldsymbol{\alpha},0}(t)d\boldsymbol{\alpha}. \tag{A.4}$$

在 H_0 下, 有 $H_{\tau_0}=F=G$ 和 $\tau_0F_{\boldsymbol{\alpha}}(t)+(1-\tau_0)G_{\boldsymbol{\alpha}}(t)=F_{\boldsymbol{\alpha}}(t)=G_{\boldsymbol{\alpha}}(t)$. 这表明在零假设下, (A.4) 和 (A.1) 具有相同的分布. 因此, 定理的第一部分成立. 另一方面, 由 (A.4) 知, 在 H_1 下, $\widehat{T}^b=O_p^*(1)$. 然而, 由定理 3(2) 知, 在 H_1 下, $\widehat{T}$ 依概率收敛到 ∞. 从而, 相应的 p- 值依概率收敛到 0. 因此, 定理的第二个陈述成立. □

定理 5 的证明 记

$$\widehat{R}_1 = -(mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{x}_j,\boldsymbol{y}_i-\boldsymbol{x}_j)+\widehat{T}_3^*,$$

$$\widehat{R}_2 = -(mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{y}_j,\boldsymbol{x}_i-\boldsymbol{y}_j)+\widehat{T}_5^*,$$

$$\widehat{R}_3 = +m^{-2}\sum_{i\neq j}^{m}\mathrm{ang}(\boldsymbol{x}_i-\boldsymbol{x}_j,\boldsymbol{x}_i-\boldsymbol{x}_j)+n^{-2}\sum_{i\neq j}^{n}\mathrm{ang}(\boldsymbol{y}_i-\boldsymbol{y}_j,\boldsymbol{y}_i-\boldsymbol{y}_j)$$
$$-2\widehat{T}_1^*-(3m-2)m^{-1}\widehat{T}_2^*-2\widehat{T}_4^*-(3n-2)n^{-1}\widehat{T}_6^*.$$

由 $\widehat{T}$ 的定义及简单计算知

$$\widehat{T}=\tau\widehat{U}_1+(1-\tau)\widehat{U}_2+m\{(m+n)n\}^{-1}\widehat{R}_1+n\{(m+n)m\}^{-1}\widehat{R}_2+(m+n)^{-1}\widehat{R}_3,$$

其中 $\widehat{U}_1=2\widehat{T}_1^*-\widehat{T}_2^*-\widehat{T}_3^*$, $\widehat{U}_2=2\widehat{T}_4^*-\widehat{T}_5^*-\widehat{T}_6^*$. 注意到 $\widehat{U}_1$、$\widehat{U}_2$、$\widehat{R}_1$、$\widehat{R}_2$ 和 $\widehat{R}_3$ 都是标准 U- 统计量. 根据 U- 统计量的 Hoeffding 分解及事实 $0\leqslant\mathrm{ang}(\cdot,\cdot)\leqslant\pi$, 有 $\mathrm{var}_{H_1}(\widehat{T})=O[m^2(m+n)^{-2}+n^2(m+n)^{-2}+m^2\{(m+n)n\}^{-2}+n^2\{(m+n)m\}^{-2}+(m+n)^{-2}]$. 因此, $\lim_{n\to\infty}\mathrm{var}_{H_1}(\widehat{T})=0$.

另一方面, 注意到在 H_1 下, $\mathrm{E}\widehat{U}_1=2T_1-T_2-T_3\stackrel{\mathrm{def}}{=}\delta_1>0$ 和 $\mathrm{E}\widehat{U}_2=2T_4-T_5-T_6\stackrel{\mathrm{def}}{=}\delta_2>0$. 因而, 对任意 $m,n\geqslant 3$, 有 $\tau\mathrm{E}\widehat{U}_1+(1-\tau)\mathrm{E}\widehat{U}_2=m(m+n)^{-1}\delta_1+n(m+n)^{-1}\delta_2\geqslant\min(\delta_1,\delta_2)>0$. 由 $0\leqslant\mathrm{ang}(\cdot,\cdot)\leqslant\pi$ 知, 在 H_0 和 H_1 下, 存在某个正常数 c 使得 $m\{(m+n)n\}^{-1}\mathrm{E}\widehat{R}_1+n\{(m+n)m\}^{-1}\mathrm{E}\widehat{R}_2+(m+n)^{-1}\mathrm{E}\widehat{R}_3\leqslant c[m\{(m+n)n\}^{-1}+n\{(m+n)m\}^{-1}+(m+n)^{-1}]=c\{m^{-1}+n^{-1}-(m+n)^{-1}\}$. 因此, $\mathrm{E}_{H_1}\widehat{T}\geqslant\min(\delta_1,\delta_2)\{1+o(1)\}>0$. 注意到, 在零假设下 $\delta_1=\delta_2=0$, 即 $\mathrm{E}_{H_0}\widehat{T}=o(1)$. 因此, $\liminf_{n\to\infty}(\mathrm{E}_{H_1}\widehat{T}-\mathrm{E}_{H_0}\widehat{T})$ 是正的, 仅依赖于 δ_1 和 δ_2, 且与比率 τ 无关. □

引理 4 及其证明 在条件 (A1)–(A3) 下, 有如下引理 4.

引理 4 假设条件 (A1)–(A3) 成立. 当 p 趋于 ∞ 时, 对所有 $i\neq j$, $j\neq k$ 和 $i\neq k$, 有

(1) $(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{y}_k-\boldsymbol{x}_j)/p\xrightarrow{\mathrm{pr}}\sigma_1^2$ 和 $(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{x}_j)/p\xrightarrow{\mathrm{pr}}\sigma_1^2$;

(2) $(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{y}_j)/p\xrightarrow{\mathrm{pr}}\sigma_2^2$ 和 $(\boldsymbol{y}_i-\boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{y}_k-\boldsymbol{y}_j)/p\xrightarrow{\mathrm{pr}}\sigma_2^2$;

(3) $(\boldsymbol{y}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{y}_k-\boldsymbol{x}_j)/p\xrightarrow{\mathrm{pr}}\sigma_1^2+\nu^2$ 和 $(\boldsymbol{x}_i-\boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{y}_j)/p\xrightarrow{\mathrm{pr}}\sigma_2^2+\nu^2$;

(4) $\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/p\xrightarrow{\mathrm{pr}}2\sigma_1^2$ 和 $\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2/p\xrightarrow{\mathrm{pr}}2\sigma_2^2$;

(5) $\|\boldsymbol{x}_i-\boldsymbol{y}_j\|^2/p\xrightarrow{\mathrm{pr}}\sigma_1^2+\sigma_2^2+\nu^2$.

由于相关收敛结果的证明是类似的. 为了避免重复, 我们仅处理 $(\boldsymbol{y}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{y}_k-\boldsymbol{x}_j)/p$. 在条件 (A1) 和 (A2) 下, 当 $p\to\infty$ 时, 对所有的 $i\neq j$, $j\neq k$ 和 $i\neq k$, 有 $(\boldsymbol{y}_i-\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{y}_j-\boldsymbol{x}_k)/p-\mathrm{E}\{(\boldsymbol{y}_i-\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{y}_j-\boldsymbol{x}_k)\}/p\xrightarrow{\mathrm{pr}}0$. 通过条件 (A3), 当 $p\to\infty$ 时, $\mathrm{E}\{(\boldsymbol{y}_i-\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{y}_j-\boldsymbol{x}_k)\}/p=\mathrm{tr}(\boldsymbol{\Sigma}_1)/p+\|\boldsymbol{u}_1-\boldsymbol{u}_2\|^2/p\to\sigma_1^2+\nu^2$. 综合以上结果知 $(\boldsymbol{y}_i-\boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{y}_j-\boldsymbol{x}_k)/p\xrightarrow{\mathrm{pr}}\sigma_1^2+\nu^2$. 证毕. □

引理 1 的证明 根据引理 4 易知, 当 p 趋于 ∞ 时, 有 $\lim_{p\to\infty}T=\gamma_0$ 和 $\widehat{T}\xrightarrow{\mathrm{pr}}\gamma_0$. 定义 $h(x)\stackrel{\mathrm{def}}{=}2\arccos(x)-\arccos(2x^2)-\pi/3$. 简单计算知, $h(x)$ 在区间 $0\leqslant x\leqslant 1/2$ 上是严格递减函数, 在区间 $1/2\leqslant x\leqslant 2^{-1/2}$ 上是严格递增函数. 从而, 在区间 $0\leqslant x\leqslant 2^{-1/2}$ 上, $h(x)\geqslant h(1/2)=0$ 且等号成立当且仅当 $x=1/2$. 结合 γ_0 的定义并注意到 $T_{1,2}\leqslant T_{1,1}$ 和 $T_{2,2}\leqslant T_{2,1}$, 可得

$$\gamma_0\geqslant\tau\left\{2\arccos(x_1)-\arccos(2x_1^2)-\frac{\pi}{3}\right\}+(1-\tau)\left\{2\arccos(x_2)-\arccos(2x_2^2)-\frac{\pi}{3}\right\}\geqslant 0,$$

其中 $1/2\leqslant x_1=[\sigma_1^2/\{2(\sigma_1^2+\sigma_2^2+\nu^2)\}]^{1/2}\leqslant 2^{-1/2}$ 和 $1/2\leqslant x_2=[\sigma_2^2/\{2(\sigma_1^2+\sigma_2^2+\nu^2)\}]^{1/2}\leqslant 2^{-1/2}$. 因此, $\gamma_0=0$ 当且仅当 $\nu^2=0$ 和 $\sigma_1^2=\sigma_2^2$. □

定理 6 的证明 不失一般性, 考虑 $m=n$. 在一个排序样本中, 假设 $n-s$ 个观测 $(s=0,1,\ldots,n)$ 来自于 F, s 个观测来自于 G, 另一个排序样本来自于剩余样本. 由引理 1 及 $T_{k,1}$ 和 $T_{k,2}$ 的定义知, 当 $p\to\infty$ 时, 统计量依概率收敛到

$$2c_{s,1}(T_{11}+T_{21})-c_{s,2}\left(T_{12}+T_{22}+\frac{2\pi}{3}\right)\stackrel{\mathrm{def}}{=}\gamma_s,$$

其中 $c_{s,1}=\{s^2+(n-s)^2\}/n^2-2s(n-s)/\{(n-1)n\}$ 和 $c_{s,2}=-2s(n-s)/n^2+(n-s)(n-s-1)/\{(n-1)n\}+s(s-1)/\{(n-1)n\}$. 进一步, 注意到

$$n(n-1)=2n(n-s)+(n-s)(n-s-1)+s(s-1)\quad\text{和}\quad s^2+(n-s)^2=n^2-2s(n-s).$$

因此, $\gamma_s=\gamma_0[\{(n-s)^2+s^2\}/n^2-s(n-s)/C(n,2)]$. 因为 $\{(n-s)^2+s^2\}/n^2-s(n-s)/C(n,2)\leqslant 1$ 和 $\gamma_0>0$, 所以对所有的 s 有 $\gamma_s\leqslant\gamma_0$, 且等号成立当且仅当 $s=0$ 或 $s=n$. 当 $p\to\infty$ 时, $\widehat{T}$ 的排序分布趋于具有 $n+1$ 个离散点 $\gamma_0,\gamma_1,\ldots,\gamma_n$ 且对应概率分别为 $C(n,n)C(n,0)/C(2n,n), C(n,n-1)C(n,1)/C(2n,n),\ldots,C(n,0)C(n,n)/C(2n,n)$ 的多点分布. 因此, 当 p 趋于无穷时, 基于排序分布知, 统计量 $\widehat{T}$ 大于或等于 γ_0 的概率趋于 $2/C(2n,n)$. □

引理 2 的证明 由 $F=G$ 立即知 $T_1=T_2=T_3$ 和 $T_4=T_5=T_6$. 另一方面, 若 $T_1=T_2=T_3$ 和 $T_4=T_5=T_6$, 则 $2T_1-T_2-T_3=0$ 和 $2T_4-T_5-T_6=0$. 根据定理 1 知 $2T_1-T_2-T_3\geqslant 0$ 和 $2T_4-T_5-T_6\geqslant 0$, 其中等号成立当且仅当 F 和 G 是相同的. □

引理 3 的证明 结论是显然的, 因为 γ_0^* 中的每一项都是非负的. □

定理 7 的证明 不失一般性, 考虑 $m=n$. 首先, 注意到

$$\begin{aligned}\gamma_0^*=&\,8^{-1}\left\{\left(2T_{11}-T_{12}+2T_{21}-T_{22}-\frac{2\pi}{3}\right)^2+\left(T_{12}+T_{22}-\frac{2\pi}{3}\right)^2\right\}\\&+8^{-1}\{(2T_{11}+T_{22}-2T_{21}-T_{12})^2+(T_{22}-T_{12})^2\}\end{aligned}$$

和

$$\begin{aligned}\widehat{T}^*=&\,8^{-1}\{(2\widehat{T}_1^*-\widehat{T}_2^*-\widehat{T}_3^*+2\widehat{T}_4^*-\widehat{T}_5^*-\widehat{T}_6^*)^2+(\widehat{T}_2^*+\widehat{T}_5^*-\widehat{T}_3^*-\widehat{T}_6^*)^2\}\\&+8^{-1}\{(2\widehat{T}_1^*+\widehat{T}_5^*+\widehat{T}_6^*-2\widehat{T}_4^*-\widehat{T}_2^*-\widehat{T}_3^*)^2+(\widehat{T}_2^*+\widehat{T}_6^*-\widehat{T}_3^*-\widehat{T}_5^*)^2\}.\end{aligned}$$

在一个排序样本中, 假设 $n-s$ 个观测 $(s=0,1,\ldots,n)$ 来自 F, s 个观测来自 G, 另一个排序样本来自剩余样本. 基于这些观测, 由定理 6 知, $(2\widehat{T}_1^*-\widehat{T}_2^*-\widehat{T}_3^*+2\widehat{T}_4^*-\widehat{T}_5^*-\widehat{T}_6^*)^2\xrightarrow{\text{pr}}\gamma_0^2[\{(n-s)^2+s^2\}/n^2-s(n-s)/C(n,2)]^2$ 和 $(2\widehat{T}_1^*+\widehat{T}_5^*+\widehat{T}_6^*-2\widehat{T}_4^*-\widehat{T}_2^*-\widehat{T}_3^*)^2\xrightarrow{\text{pr}}(2T_{11}+T_{22}-2T_{21}-T_{12})^2[\{(n-s)^2+s^2\}/n^2-s(n-s)/C(n,2)]^2$, 其中 $\gamma_0=2T_{11}-T_{12}+2T_{21}-T_{22}-2\pi/3$, 现需要处理 $(\widehat{T}_2^*+\widehat{T}_5^*-\widehat{T}_3^*-\widehat{T}_6^*)^2$ 和 $(\widehat{T}_2^*+\widehat{T}_6^*-\widehat{T}_3^*-\widehat{T}_5^*)^2$. 因为处理这两项是类似的, 为了避免重复, 我们考虑前者. 根据引理 4 知, 当 $p\to\infty$ 时, $\widehat{T}_2^*+\widehat{T}_6^*\xrightarrow{\text{pr}}2[s(n-s)/\{(n-1)n\}]T_{1,1}+[(n-s)(n-s-1)/\{(n-1)n\}]T_{1,2}+[s(n-s)/\{(n-1)n\}]T_{2,1}+[s(s-1)/\{(n-1)n\}]T_{2,2}+s(s-1)\pi/\{3(n-1)n\}+(n-s)(n-s-1)\pi/\{3(n-1)n\}$ 和 $\widehat{T}_3^*+\widehat{T}_5^*\xrightarrow{\text{pr}}2[s(n-s)/\{(n-1)n\}]T_{1,1}+[s(s-1)/\{(n-1)n\}]T_{1,2}+[s(n-s)/\{(n-1)n\}]T_{2,1}+[(n-s)(n-s-1)/\{(n-1)n\}]T_{2,2}+s(s-1)\pi/\{3(n-1)n\}+(n-s)(n-s-1)\pi/\{3(n-1)n\}$. 因此, 当 p 趋于无穷时, $(\widehat{T}_2^*+\widehat{T}_6^*-\widehat{T}_3^*-\widehat{T}_5^*)^2\xrightarrow{\text{pr}}[\{C(n-s,2)-C(s,2)\}^2/\{C(n,2)\}^2](T_{22}-T_{12})^2$. 类似地, 当 p 趋于无穷时, $(\widehat{T}_2^*+\widehat{T}_5^*-\widehat{T}_3^*-\widehat{T}_6^*)^2\xrightarrow{\text{pr}}[\{C(n-s,2)-C(s,2)\}^2/\{C(n,2)\}^2](T_{12}+T_{22}-2\pi/3)^2$.

综合上述结果知, 当 $p\to\infty$ 时, $\widehat{T}^*$ 依概率收敛到 γ_s^*, 其中

$$\begin{aligned}\gamma_s^*=&\,8^{-1}\left\{\left(2T_{11}-T_{12}+2T_{21}-T_{22}-\frac{2\pi}{3}\right)^2+(2T_{11}+T_{22}-2T_{21}-T_{12})^2\right\}\\&\times\left[\frac{(n-s)^2+s^2}{n^2}-\frac{s(n-s)}{C(n,2)}\right]^2\\&+8^{-1}\left\{\left(T_{12}+T_{22}-\frac{2\pi}{3}\right)^2+(T_{22}-T_{12})^2\right\}\left\{\frac{C(n-s,2)-C(s,2)}{C(n,2)}\right\}^2\\\leqslant&\,\gamma_0^*\max\left(\left[\frac{(n-s)^2+s^2}{n^2}-\frac{s(n-s)}{C(n,2)}\right]^2,\left\{\frac{C(n-s,2)-C(s,2)}{C(n,2)}\right\}^2\right).\end{aligned}$$

因为 $\max([\frac{\{(n-s)^2+s^2\}}{n^2}-\frac{s(n-s)}{C(n,2)}]^2,[\frac{\{C(n-s,2)-C(s,2)\}}{C(n,2)}]^2)\leqslant 1$ 且当 $s=0$ 或 $s=n$ 时等号成立, 当 p 趋于无穷时, $\widehat{T}^*$ 的排序分布趋于具有 $n+1$ 个离散点 $\gamma_0^*,\gamma_1^*,\ldots,\gamma_n^*$ 且对应概率分别为 $C(n,n)C(n,0)/C(2n,n)$,

$C(n,n-1)C(n,1)/C(2n,n),\ldots,C(n,0)C(n,\ n)/C(2n,n)$ 的多点分布. 从而, 当 p 趋于无穷时, 基于排序分布知, 统计量 $\widehat{T}^*$ 大于或等于 γ_0^* 的概率趋于 $2/C(2n,n)$. 证毕. □

Nonparametric two-sample tests for equality of distributions using projections

Kai Xu & Liping Zhu

Abstract We propose a nonparametric two-sample test, which generalizes the Cramér-von Mises test through projections, to test for equality of two distributions in high dimensions. The population version of our proposed generalized Cramér-von Mises statistic is nonnegative and equals zero if and only if the two distributions are identical, ensuring that our proposed test is consistent against all the fixed alternatives. In addition, our proposed test statistic has an explicit form and is completely free of tuning parameters. It requires no moment conditions and hence is robust to the presence of outliers and heavy-tail observations. We study the asymptotic behaviors of our proposed test under both the "large sample size, fixed dimension" and the "fixed sample size, large dimension" paradigms. In the former paradigm, we show that the asymptotic power of our proposed test does not depend on the size ratio of the two random samples. This ensures that our proposed test can be readily applied to imbalanced samples. In the latter paradigm, we observe that, surprisingly, the two distributions are equal if and only if their first two moments are equal. In this paradigm, we suggest tailoring our proposed test to detect location shifts and scale differences, which further enhances the power performance of our proposed test significantly. Numerical studies confirm that our proposals are superior to many existing tests in high-dimensional two-sample test problems.

Keywords **Cramér-von Mises test, equality of distributions, high-dimension, projections, two-sample problem**

MSC(2020) **62H15, 62G10, 62G20**

doi: 10.1360/SSM-2020-0317