高校应用数学学报
2025, 40(2): 159-170

基于能量距离的k-配对样本的分布差异检验

陈敏琼[1], 谭合理[2*]

(1. 广州新华学院 人工智能与数据科学系, 广东广州 510520;
2. 广东金融学院 金融数学与统计学院, 广东广州 510521)

摘 要: 针对具有相关性的多个总体分布差异的检验问题, 基于能量距离的概念提出了一种新的检验方法. 首先给出度量k个相关变量的分布差异的测度及其样本估计, 该估计式具有V-统计量形式, 利用V-统计量理论讨论了估计量的渐近性质. 然后给出了检验的Bootstrap重抽样程序并验证了程序的合理性. 数值模拟与实例数据分析均表明, 与经典的Hotelling's T^2检验与Friedman检验方法相比, 新的方法能更准确地鉴别k个相关变量在除了位置之外的其他特征上的差异, 并且能应用于多元变量情形. 因此新方法适用于更广泛的数据类型.

关键词: 能量距离; k-配对样本; 分布差异; Hotelling's T^2检验; Friedman检验; V-统计量; Bootstrap
中图分类号: O212.7
文献标识码: A **文章编号**: 1000-4424(2025)02-0159-12

§1 引 言

在临床试验、心理学或社会学等领域中, 研究者经常会接触重复测量的数据, 在这些领域中, 比较多组重复测量数据的分布差异是一个基本的问题. 例如, 在纵向研究中, 对同一组患者服药前及服药后多次测量血压, 以确定某一药物是否对患者的血压有影响; 相同受试对象在相同实验条件下分别接受不同频率声音的刺激, 比较他们的反应的差异[1]; 比较不同的评价体系下排名的差异[2-3]; 比较空气质量随年份变化的差异[4]. 这些问题的共同点是, 不同组观测结果之间不再是独立而是具有相关性的, 用数学语言可表述如下: 假设$(Y_1, Y_2, \cdots, Y_k)$是欧氏空间$\mathbf{R}^p \times \mathbf{R}^p \times \cdots \times \mathbf{R}^p$中的随机向量, 需要通过重复测量的样本去检验这$k$个具有相关性的变量$Y_1, Y_2, \cdots, Y_k$的分布是否存在显著性差异。

收稿日期: 2023-04-08 修回日期: 2024-10-14
*通讯作者, Email: thl0427@163.com
基金项目: 国家自然科学基金青年项目(11801482); 广东省特色创新人才自然科学基金(2022KTSCX180); 广东省重点建设学科科研能力提升项目(2024ZDJS132)

k个独立样本的分布差异检验问题已有大量文献讨论, 如经典的多元方差分析(MANOVA)、非参数的Kruskal-Wallis检验、多元变量情形下基于能量距离的方差分析(DISCO)检验法[5]. 而讨论当k个变量具有相关性时的分布差异检验问题的文献则相对较少. 当$(Y_1, Y_2, \cdots, Y_k)$为多维正态分布时, Hotelling's T^2检验可以用来检验这k个总体均值向量是否有显著性差异; 当$(Y_1, Y_2, \cdots, Y_k)$不满足正态假设但$p = 1$时, 非参数方法的Friedman检验是检验k个总体均值(或中位数)是否存在显著性差异的经典方法, Cochran'Q检验法则针对0-1分布变量检验它们等于1的概率的差异. Martínez-Camblor等[6]基于核密度估计法提出共面积(Common Area, AC)统计量, 用于检验k个独立样本分布差异, 文献[7]将方法拓展到k个具有相关性变量的分布差异检验问题. 该方法采用核密度估计法, 因此涉及窗宽的选择问题, 并且仅适用于一元连续型变量情形. 当$p > 1$, 同时$(Y_1, Y_2, \cdots, Y_k)$不满足正态假设的情形, 目前尚未有文献提及.

能量距离(Energy Distance, ED)是由Székely[8]提出的用于度量两个独立的多维变量的分布差异的测度, 它是通过对两个随机变量的特征函数之差进行加权积分后得到的一个显式表达, 因此ED具有非负性并且等于0当且仅当两个随机变量同分布的良好性质. Székely等人[5,9-12]展示了用ED处理一系列经典统计问题的结果, 并在文献[13]中对这些结果进行了总结. 研究结果表明, 相比传统的方法, ED方法计算简便、适用于更广泛分布类型的数据, 且能处理多变量情况. 文献[14]指出针对两个独立变量的ED的概念对具有相关性的变量仍然适用, 并将ED应用于两个配对样本分布差异的检验问题. 此外, 文献[14]还重点将ED的概念推广到带协变量的情形, 给出了在给定协变量时两个配对变(向)量的条件分布差异检验方法, 并讨论了方法的大样本性质.

本文借鉴文献[5, 14]的思路, 将两个配对样本的分布差异的ED检验法, 推广至多个具有相关性的总体分布差异的检验. 具体地, 首先给出度量k个相关变量的分布差异的测度, 并给出它的样本估计, 该估计式具有V-统计量形式, 利用V-统计量理论讨论估计量的大样本性质. 然后给出检验的重抽样程序及并验证程序的合理性. 数值模拟与实例数据分析均表明, 新的方法能更准确地鉴别k个相关变量在除了位置之外的其他特征比如方差上的差异, 因此能适用于更广泛的数据类型.

§2 方法原理与主要结果

假设(X, Y)为欧氏空间$\mathbf{R}^p \times \mathbf{R}^p$中的随机向量, 其联合分布函数为$F(x, y), (x, y) \in \mathbf{R}^p \times \mathbf{R}^p$. 对于检验问题

$$H_0 : X \overset{d}{=} Y, \tag{1}$$

其中符号"$\overset{d}{=}$"表示左右两边的随机变(向)量服从相同的分布, 设(X', Y')为(X, Y)的一个独立同分布的复制, 对于具有有限的一阶矩的两个相关变量X和Y, 它们的能量距离可定义为[14]

$$V(X, Y) := E|X - Y'| + E|X' - Y| - E|X - X'| - E|Y - Y'|, \tag{2}$$

并且有$V(X, Y) \geq 0$, 等号成立当且仅当X和Y同分布, 其中"$|\cdot|$"表示欧氏空间中的距离范数. 与独立情形相比, 两个相关变量的ED相当于将独立变量的ED定义中的$E|X - Y|$修正为$E|X - Y'|$. 文献[14]给出了$V(X, Y)$的估计量表达及检验的重抽样程序, 估计量具有V-统计量形式. 具体地, 假设$\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$是来自总体分布$(X, Y)$的一个简单随机样

本, 取$V(X,Y)$的估计量

$$\hat{V}_n(X,Y) := \frac{1}{n^2}\sum_{i,j=1}^{n}\big(|X_i-Y_j|+|X_j-Y_i|-|X_i-X_j|-|Y_i-Y_j|\big) = \frac{1}{n^2}\sum_{i,j=1}^{n}h\big((X_i,Y_i),(X_j,Y_j)\big), \tag{3}$$

其中

$$h((x_1,y_1),(x_2,y_2)) = |x_1-y_2|+|x_2-y_1|-|x_1-x_2|-|y_1-y_2|.$$

可以看到, $\hat{V}_n(X,Y)$的计算方式跟两个独立样本的ED的估计量的计算方式是一致的, 但它们的极限性质是不一样的, 因此计算检验p-值的程序是不一样的[8,14]. 另外, $\hat{V}_n(X,Y)$是$V(X,Y)$的有偏估计, 但是在大样本情形下, 文献[14]提到它具有渐近无偏性, 后文命题2.1将给出详细的证明.

显然, 以上关于两个配对变(向)量的分布差异检验方法可以容易地推广到k个配对变(向)量的分布差异检验问题

$$H_0^k : Y_1 \overset{d}{=} Y_2 \overset{d}{=} \cdots \overset{d}{=} Y_k, \tag{4}$$

其中$(Y_1,Y_2,\cdots,Y_k)$为$\mathbf{R}^p\times\mathbf{R}^p\times\cdots\times\mathbf{R}^p$上的随机向量. 事实上, 若记

$$D^{(k)} := \sum_{l<s}V(Y_l,Y_s).$$

其中"$\sum_{l<s}$"表示对$(Y_1,Y_2,\cdots,Y_k)$中两两配对共$k(k-1)/2$对变量的能量距离求和, 则有$D^{(k)}\geq 0$, 并且$D^{(k)}=0$当且仅当$Y_1,Y_2,\cdots,Y_k$都具有相同的分布. 假设$\{(Y_{1i},Y_{2i},\cdots,Y_{ki})\}_{i=1}^n$为$(Y_1,Y_2,\cdots,Y_k)$的简单随机样本, 则可取$D^{(k)}$相应的估计量

$$\hat{D}_n^{(k)} = \sum_{l<s}\hat{V}_n(Y_l,Y_s) \tag{5}$$

作为H_0^k的检验统计量. 下文用符号"$\xrightarrow[n\to\infty]{\text{a.s.}}$"表示当$n$趋向无穷时"几乎处处收敛", 及"$\xrightarrow[n\to\infty]{\mathcal{D}}$"表示"依分布收敛". 下面的命题给出了$\hat{D}_n^{(k)}$的渐近性质.

命题2.1 假设$E|X|<\infty, E|Y|<\infty$, 则

$$\hat{D}_n^{(k)} \xrightarrow[n\to\infty]{\text{a.s.}} D^{(k)}.$$

证 由$\hat{D}_n^{(k)}$的定义可知, 只需要证明$\hat{V}_n(Y_l,Y_s)\xrightarrow[n\to\infty]{\text{a.s.}}V(Y_l,Y_s),\forall l\neq s,l,s=1,2,\cdots,k$即可. 由于

$$\hat{V}_n(Y_l,Y_s) = \frac{1}{n^2}\sum_{i,j=1}^{n}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big) = \frac{2}{n^2}\sum_{i<j}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)+\frac{1}{n^2}\sum_{i=1}^{n}h\big((Y_{li},Y_{si}),(Y_{li},Y_{si})\big) = \frac{2C_n^2}{n^2}U_{lsn}+\frac{2}{n^2}\sum_{i=1}^{n}|Y_{li}-Y_{si}|, \tag{6}$$

其中U_{lsn}表示U-统计量

$$\frac{1}{C_n^2}\sum_{i<j}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big).$$

对于U-统计量U_{lsn}有

$$E\Big[h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)\Big]=V(Y_l,Y_s),$$

并且由Y_l,Y_s具有有限的一阶矩可知

$$E\Big|h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)\Big|=E\Big||Y_{li}-Y_{sj}|+|Y_{lj}-Y_{si}|-|Y_{li}-Y_{lj}|-|Y_{si}-Y_{sj}|\Big|<\infty.$$

从而根据文献[15]中第3章的定理3可得

$$U_{lsn}\xrightarrow[n\to\infty]{\text{a.s.}}V(Y_l,Y_s),\forall l\neq s,l,s=1,2,\cdots,k. \tag{7}$$

另外由强大数定律可知

$$\frac{1}{n}\sum_{i=1}^{n}|Y_{li}-Y_{si}|\xrightarrow[n\to\infty]{\text{a.s.}}E|Y_l-Y_s|,\forall l\neq s,l,s=1,2,\cdots,k.$$

因此有

$$\frac{1}{n^2}\sum_{i=1}^{n}|Y_{li}-Y_{si}|\xrightarrow[n\to\infty]{\text{a.s.}}0,\forall l\neq s,l,s=1,2,\cdots,k. \tag{8}$$

结合(6)-(8)最终可得

$$\hat{V}_n(Y_l,Y_s)\xrightarrow[n\to\infty]{\text{a.s.}}V(Y_l,Y_s),\forall l\neq s,l,s=1,2,\cdots,k.$$

从而有

$$\hat{D}_n^{(k)}=\sum_{l<s}\hat{V}_n(Y_l,Y_s)\xrightarrow[n\to\infty]{\text{a.s.}}\sum_{l<s}V(Y_l,Y_s)=D^{(k)}.$$

命题2.2 当零假设$H_0^k:Y_1\overset{d}{=}Y_2\overset{d}{=}\cdots\overset{d}{=}Y_k$成立时, 有

$$n\hat{D}_n^{(k)}\xrightarrow[n\to\infty]{\mathcal{D}}\sum_{v}\lambda_v^{(k)}Z_v^2, \tag{9}$$

其中$\{Z_v\}_{v=1}^{\infty}$为独立同分布的标准正态分布随机变量序列, $\{\lambda_v^{(k)}\}_{v=1}^{\infty}$为依赖于$(Y_1,Y_2,\cdots,Y_k)$的联合分布的常数序列.

证 显然$\hat{D}_n^{(k)}$可以重写成

$$\begin{aligned}\hat{D}_n^{(k)}&=\sum_{l<s}\hat{V}_n(Y_l,Y_s)=\sum_{l<s}\frac{1}{n^2}\sum_{i,j=1}^{n}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)=\\&\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{l<s}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)=\\&\frac{1}{n^2}\sum_{i,j=1}^{n}\mathcal{H}\big((Y_{1i},Y_{2i},\cdots,Y_{ki}),(Y_{1j},Y_{2j},\cdots,Y_{kj})\big),\end{aligned}$$

其中

$$\mathcal{H}\big((Y_{1i},Y_{2i},\cdots,Y_{ki}),(Y_{1j},Y_{2j},\cdots,Y_{kj})\big)=\sum_{l<s}h\big((Y_{li},Y_{si}),(Y_{lj},Y_{sj})\big)$$

是关于参数对称的函数. 因此$\hat{D}_n^{(k)}$是一个V-统计量, 并且容易验证$\hat{D}_n^{(k)}$在零假设

$$H_0^k:Y_1\overset{d}{=}Y_2\overset{d}{=}\cdots\overset{d}{=}Y_k$$

下是一阶限化的. 从而有

$$n\hat{D}_n^{(k)}\xrightarrow[n\to\infty]{\mathcal{D}}\sum_{v}\lambda_v^{(k)}\mathcal{Z}_v^2,$$

其中$\{\mathcal{Z}_v\}_{v=1}^{\infty}$为独立同分布的标准正态分布随机变量序列, $\{\lambda_v^{(k)}\}_{v=1}^{\infty}$为积分方程

$$\int \mathcal{H}\big((y_1,y_2,\cdots,y_k),(y_1',y_2',\cdots,y_k')\big)\cdot\rho(y_1,y_2,\cdots,y_k)\mathrm{d}F(y_1,y_2,\cdots,y_k)= \lambda\rho(y_1',y_2',\cdots,y_k') \tag{10}$$

的特征根, 其中$F(y_1,y_2,\cdots,y_k)$为$(Y_1,Y_2,\cdots,Y_k)$的联合分布函数.

注意到由于$Y_1,Y_2,\cdots,Y_k$具有相同的分布, 因此

$$\begin{aligned}&F(y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(k)})=P(Y_1\le y_{\pi(1)},Y_2\le y_{\pi(2)},\cdots,Y_k\le y_{\pi(k)})=\\&P(Y_{\pi(1)}\le y_{\pi(1)},Y_{\pi(2)}\le y_{\pi(2)},\cdots,Y_{\pi(k)}\le y_{\pi(k)})=\\&P(Y_1\le y_1,Y_2\le y_2,\cdots,Y_k\le y_k)=F(y_1,y_2,\cdots,y_k),\end{aligned} \tag{11}$$

其中$(\pi(1),\pi(2),\cdots,\pi(k))$为$1,2,\cdots,k$的任一全排列. 因此(10)中的$F(y_1,y_2,\cdots,y_k)$可以改写为

$$F(y_1,y_2,\cdots,y_k)=\frac{1}{k!}\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}F(y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(k)}). \tag{12}$$

§3 重抽样程序

命题2.2表明在零假设(4)成立时, $n\hat{D}_n^{(k)}$依分布收敛于混合卡方分布$\sum_v\lambda_v^{(k)}\mathcal{Z}_v^2$, 而命题2.1说明$n\hat{D}_n^{(k)}$在备择假设下几乎处处收敛于$+\infty$. 因此, 在给定显著性水平为$\alpha$下只要$n\hat{D}_n^{(k)}>c_\alpha$便拒绝零假设(4), 其中$c_\alpha$为$\sum_{v=1}^{\infty}\lambda_v^{(k)}\mathcal{Z}_v^2$分布的上$\alpha$-分位点. 然而, 由于这些$\lambda_v^{(k)}$依赖于未知的$(Y_1,Y_2,\cdots,Y_k)$的联合分布, 因此实际中很难得到确切的$c_\alpha$. 作为选择, 考虑用bootstrap重抽样方法来近似获得$n\hat{D}_n^{(k)}$在零假设下的渐近分布并依此计算它的检验p-值. 具体的程序步骤如下.

- **步骤1** 基于原始样本
$$\mathcal{S}_n=\{(Y_{11},Y_{21},\cdots,Y_{k1}),(Y_{12},Y_{22},\cdots,Y_{k2}),\cdots,(Y_{1n},Y_{2n},\cdots,Y_{kn})\}$$
计算$n\hat{D}_n^{(k)}$, 记为$n\hat{D}_n^{(k)}(\mathcal{S}_n)$.
- **步骤2** 从原始样本$\mathcal{S}_n$中有放回地抽取容量为n的样本, 不妨记为$\{(Y_{1i}^\star,Y_{2i}^\star,\cdots,Y_{ki}^\star),i=1,2,\cdots,n\}$. $\forall i=1,2,\cdots,n$, 产生$(1,2,\cdots,k)$的一个全排列$(\pi_i(1),\pi_i(2),\cdots,\pi_i(k))$, 最后可得到重抽样本
$$\{(Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b)\}_{i=1}^n:=\{(Y_{\pi_i(1)}^\star,Y_{\pi_i(2)}^\star,\cdots,Y_{\pi_i(k)}^\star)\}_{i=1}^n$$
及其相应的$n\hat{D}_n^{(k)}$值, 记为$n\hat{D}_n^{(k,b)}$.
- **步骤3** 重复步骤2 B次(如取$B=199$) 得到$n\hat{D}_n^{(k)}$的bootstrap统计量$\{n\hat{D}_n^{(k,b)},1\le b\le B\}$, 那么$n\hat{D}_n^{(k)}$的$p$-值可由重抽样统计量中$\{n\hat{D}_n^{(k,b)}>n\hat{D}^{(k)}(\mathcal{S}_n)\}$发生的比例来近似, 即
$$p\approx\frac{1+\sum_{b=1}^{B}I\Big(n\hat{D}_n^{(k,b)}>n\hat{D}^{(k)}(\mathcal{S}_n)\Big)}{1+B},$$
其中$I(\cdot)$表示示性函数.

下面的定理证明了上面提出的bootstrap程序在近似$n\hat{D}^{(k)}$在零假设下的渐近分布的合理性.

定理3.1 给定原始样本$\mathcal{S}_n$, bootstrap统计量$n\hat{D}_n^{(k,b)}$依分布收敛于命题2.2中的渐近分布$\sum_{v=1}^{\infty}\lambda_v^{(k)}\mathcal{Z}_v^2$, 即

$$n\hat{D}_n^{(k,b)}\Big|\mathcal{S}_n\xrightarrow[n\to\infty]{\mathcal{D}}\sum_{v=1}^{\infty}\lambda_v^{(k)}\mathcal{Z}_v^2. \tag{13}$$

证 给定原始样本

$$\mathcal{S}_n=\{(Y_{11},Y_{21},\cdots,Y_{k1}),(Y_{12},Y_{22},\cdots,Y_{k2}),\cdots,(Y_{1n},Y_{2n},\cdots,Y_{kn})\}$$

的条件下, $(Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b)$(其中$i=1,2,\cdots,n$)独立且同服从离散的均匀分布

$$P\big((Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b)=(Y_{\pi(1)j},Y_{\pi(2)j},\cdots,Y_{\pi(k)j})\big)=\frac{1}{nk!},j=1,2,\cdots,n.$$

令$E^\star(\cdot)$表示$E(\cdot|\mathcal{S}_n)$, 则对于V-统计量

$$\hat{D}_n^{(k,b)}=\frac{1}{n^2}\sum_{i,j=1}^{n}\mathcal{H}\big((Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b),(Y_{1j}^b,Y_{2j}^b,\cdots,Y_{kj}^b)\big),$$

由于

$$h\big((x,y),(x',y')\big)+h\big((x,y),(y',x')\big)\equiv 0,$$

容易验证

$$E^\star\Big[\mathcal{H}\big((Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b),(Y_{1j}^b,Y_{2j}^b,\cdots,Y_{kj}^b)\big)\big|(Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b)\Big]=$$

$$\frac{1}{nk!}\sum_{j=1}^{n}\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}\mathcal{H}\big((Y_{1i}^b,Y_{2i}^b,\cdots,Y_{ki}^b),(Y_{\pi(1)j},Y_{\pi(2)j},\cdots,Y_{\pi(k)j})\big)=$$

$$\frac{1}{nk!}\sum_{j=1}^{n}\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}\sum_{l<s}h\big((Y_{li}^b,Y_{si}^b),(Y_{\pi(l)j},Y_{\pi(s)j})\big)=0,$$

其中"$\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}$"表示对$(1,2,\cdots,k)$的所有全排列求和. 因此$\hat{D}_n^{(k,b)}$为一阶退化的$V$-统计量. 依据文献[15]的结论有

$$n\hat{D}_n^{(k,b)}\xrightarrow[n\to\infty]{\mathcal{D}}\sum_v\lambda_v^\star\mathcal{Z}_v^2,$$

其中$\mathcal{Z}_v$为一系列独立服从准正态分布的随机变量, $\lambda_v^\star$为积分方程

$$\int\mathcal{H}\big((y_1,y_2,\cdots,y_k),(y_1',y_2',\cdots,y_k')\big)\cdot\rho(y_1,y_2,\cdots,y_k)dF^b(y_1,y_2,\cdots,y_k)=$$

$$\lambda\rho(y_1',y_2',\cdots,y_k')$$

的特征根, $F^b(y_1,y_2,\cdots,y_k)$为$(Y_1^b,Y_2^b,\cdots,Y_k^b)$的联合分布函数极限. 显然

$$F^b(y_1,y_2,\cdots,y_k)=\lim_{n\to\infty}\frac{1}{k!}\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}F_n(y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(k)})=$$

$$\frac{1}{k!}\sum_{\{(\pi(1),\pi(2),\cdots,\pi(k))\}}F(y_{\pi(1)},y_{\pi(2)},\cdots,y_{\pi(k)}),$$

这正是(12)中的$F(y_1,y_2,\cdots,y_k)$的等价表达. 因此, 这些$\lambda_v^\star$实际上就是积分方程(10)的特征根, 这说明$n\hat{D}_n^{(k,b)}$依分布收敛于命题2.2中的$n\hat{D}_n^{(k)}$在零假设下的渐近分布.

§4 数值模拟

本节考察$\hat{D}_n^{(k)}$在检验重复测量数据分布差异中的表现. 分别讨论Y_1,Y_2,Y_3,Y_4为一元与多元变量情形. 在下面的模拟中, 选定显著性水平$\alpha=0.05$, $\hat{D}_n^{(k)}$的p-值采用$B=199$次重抽样计算得到, 每个实验重复1000次以计算各种方法拒绝零假设的比例.

1) $p=1$情形

首先考虑k个相关变量均为一元变量的情形. 假设有$k=4$维的随机向量(Y_1,Y_2,Y_3,Y_4). 选用经典的Hotelling's T^2参数检验与Friedman非参数检验两种方法与本文提出的新方法进行对比. 此外, 也考虑忽略4个变量之间的相关性, 采用文献[5]中的DISCO检验法, 以考察其直接应用于多个配对变量的分布差异检验时的表现, 设定DISCO检验法在计算检验p-值时的置换样本次数为$B=199$. Hotelling's T^2检验针对(Y_1,Y_2,Y_3,Y_4)的联合分布为多元正态分布的情形, 检验四个指标的均值是否相等, 其统计量的具体表达为

$$T^2=n(A\overline{Y})^{\mathrm{T}}(A\hat{\Sigma}A^{\mathrm{T}})^{-1}A\overline{Y}, \tag{14}$$

其中

$$A=\begin{pmatrix}1,-1,0,0\\0,1,-1,0\\0,0,1,-1\end{pmatrix},$$

$\overline{Y},\hat{\Sigma}$分别为$(Y_1,Y_2,Y_3,Y_4)$的样本均值向量与样本协方差阵. 在给定$\alpha$显著性水平下, T^2的拒绝域为

$$W=\Big\{F=\frac{n-r}{(n-1)r}T^2\geq F_{r,n-r}(\alpha)\Big\},$$

其中n为样本容量, r为矩阵A的行数, $F_{r,n-r}(\alpha)$表示自由度为$(r,n-r)$的F分布的上α-分位数.

假设

$$(X_1,X_2,X_3,X_4)\sim N_4(\mu,\Sigma),$$

其中$\mu=(0,0,0,0)',\Sigma=(\sigma_{ij}),\sigma_{ij}=1,i=j,\sigma_{ij}=0.6^2,i\neq j$, 然后考虑以下10个例子.

- **例1** $Y_1=X_1,Y_2=X_2,Y_3=X_3,Y_4=X_4$.
- **例2** $Y_1=X_1,Y_2=X_2,Y_3=X_3,Y_4=X_4+0.3$.
- **例3** $Y_1=X_1,Y_2=X_2,Y_3=X_3,Y_4=\sqrt{2}X_4$.
- **例4** $Y_1=X_1,Y_2=X_2,Y_3=X_3,Y_4=\sqrt{2}X_4+0.3$.
- **例5** $Y_1=X_1,Y_2=X_2,Y_3=X_3,Y_4=1/\sqrt{2}(X_4^2-1)$.
- **例6** $Y_1=X_1^2,Y_2=X_2^2,Y_3=X_3^2,Y_4=X_4^2$.
- **例7** $Y_1=X_1^2,Y_2=X_2^2,Y_3=X_3^2,Y_4=X_4^2+0.5$.
- **例8** $Z\sim N(1/2,1)$, 且Z与(X_1,X_2,X_3,X_4)相互独立, $Y_1=X_1^2,Y_2=X_2^2,Y_3=X_3^2,Y_4=1/2X_4^2+Z$.
- **例9** $Y_1=X_1^2,Y_2=X_2^2,Y_3=X_3^2,Y_4=\sqrt{2}X_4^2$.
- **例10** $Z\sim N(1/2,3/2)$, 且Z与(X_1,X_2,X_3,X_4)相互独立, $Y_1=X_1^2,Y_2=X_2^2,Y_3=X_3^2,Y_4=1/2X_4^2+Z$.

显然, 以上例子中例1和例6为Y_1,Y_2,Y_3,Y_4同分布的情形; 例2与例7为Y_4与其他三个变量有均值上的差异; 例3与例8为Y_4与其他三个变量均值相同, 但有方差上的差异; 例4与例9为Y_4与其他三个变量既有均值上的差异, 又有方差上的差异; 例5与例10为Y_4与其他三个变量有相同的均值与方差, 但分布不同. 每个例子考虑样本容量$n=50,100,150,200$. 下表1给出了四种方法在检验4个相关变量Y_1,Y_2,Y_3,Y_4的分布差异中的表现对比结果.

首先来比较Hotelling's T^2检验、Friedman检验与$\hat{D}_n^{(k)}$检验法的表现. 从表1可以看到, 在例1和例6中, 三种方法都能较好地将第一类错误控制在名义水平0.05左右; 在例2中, 由

表 1 Y_1, Y_2, Y_3, Y_4为一元相关变量时, 例1-例10中$\hat{D}_n^{(k)}$与Hotelling's T^2检验、Friedman检验及DISCO检验法的第一类错误与功效的比较

	例1				例6			
检验方法	$n=50$	$n=100$	$n=150$	$n=200$	$n=50$	$n=100$	$n=150$	$n=200$
Hoteling's T^2	0.052	0.042	0.043	0.050	0.052	0.053	0.054	0.050
Friedman	0.043	0.051	0.047	0.047	0.056	0.045	0.057	0.056
$\hat{D}_n^{(k)}$	0.034	0.053	0.042	0.040	0.043	0.041	0.047	0.055
DISCO	0.007	0.005	0.003	0.006	0.035	0.025	0.029	0.035
	例2				例7			
检验方法	$n=50$	$n=100$	$n=150$	$n=200$	$n=50$	$n=100$	$n=150$	$n=200$
Hoteling's T^2	0.426	0.762	0.916	0.973	0.550	0.868	0.958	0.994
Friedman	0.349	0.659	0.836	0.927	0.961	1.000	1.000	1.000
$\hat{D}_n^{(k)}$	0.339	0.677	0.861	0.955	0.974	1.000	1.000	1.000
DISCO	0.153	0.430	0.662	0.855	0.954	1.000	1.000	1.000
	例3				例8			
检验方法	$n=50$	$n=100$	$n=150$	$n=200$	$n=50$	$n=100$	$n=150$	$n=200$
Hoteling's T^2	0.053	0.049	0.040	0.049	0.054	0.049	0.062	0.048
Friedman	0.048	0.048	0.055	0.051	0.094	0.181	0.255	0.323
$\hat{D}_n^{(k)}$	0.232	0.562	0.824	0.934	0.445	0.894	0.995	1.000
DISCO	0.062	0.256	0.535	0.725	0.401	0.846	0.991	1.000
	例4				例9			
检验方法	$n=50$	$n=100$	$n=150$	$n=200$	$n=50$	$n=100$	$n=150$	$n=200$
Hoteling's T^2	0.228	0.499	0.668	0.812	0.189	0.365	0.573	0.711
Friedman	0.218	0.445	0.618	0.741	0.132	0.230	0.342	0.430
$\hat{D}_n^{(k)}$	0.509	0.882	0.980	0.999	0.191	0.365	0.580	0.699
DISCO	0.263	0.685	0.910	0.983	0.156	0.298	0.508	0.634
	例5				例10			
检验方法	$n=50$	$n=100$	$n=150$	$n=200$	$n=50$	$n=100$	$n=150$	$n=200$
Hoteling's T^2	0.063	0.060	0.051	0.046	0.049	0.045	0.062	0.051
Friedman	0.106	0.190	0.290	0.373	0.065	0.086	0.121	0.148
$\hat{D}_n^{(k)}$	0.658	0.994	1.000	1.000	0.784	0.993	1.000	1.000
DISCO	0.506	0.971	1.000	1.000	0.735	0.988	1.000	1.000

于(Y_1, Y_2, Y_3, Y_4)服从多元正态分布, 并且四个变量只存在均值上的差异, 正如期望所示, 三种方法中Hotelling'sT^2具有最高的功效, $\hat{D}_n^{(k)}$表现次之, Friedman功效略低于前两者; 而在例3中, 四个变量均值相等, 但方差不全相等, 此时, 易知Hotelling'sT^2与Friedman检验法均失去功效, 而$\hat{D}_n^{(k)}$能以较满意的功效鉴别出这种差异出来. 这不难理解, 由于Hotelling'sT^2与Friedman检验法比较的变量之间的均值或其他位置上的差异, 因此无法鉴别出变量在方差上的差异; 例4中, 当四个变量在均值与方差上均有差异时, $\hat{D}_n^{(k)}$表现也是最优的; 例5中, 四个变量有相同的均值与方差, 但分布不全相同, 此时, Hotelling'sT^2仍然失去功效, Friedman检验法具有微弱的功效, 而$\hat{D}_n^{(k)}$在较小的样本量下都能以接近1的功效鉴别分布上的差异. 例7至例10中, Y_1, Y_2, Y_3, Y_4均非正态分布, 三种方法中, $\hat{D}_n^{(k)}$始终表现最优, 特别是在例8与例10中, 四个变量具有相同的均值但方差不同, 或具有相同的均值与方差但分布不同时, $\hat{D}_n^{(k)}$的功效远远高于两种经典的方法.

接下来, 来看看DISCO检验法的表现. 从表1可以看到, DISCO检验法的功效几乎与$\hat{D}_n^{(k)}$检验看齐, 但从例1和例6的结果来看, 它未能准确地将第一类错误控制在给定的名义水平0.05左右, 特别是在例1中, 第一类错误几乎接近0, 这说明在多个变量之间具有相关性时, 忽略变量之间的相关性, 直接采用DISCO检验法来检验它们的分布差异, 得到的结果不准确.

综上所述, 相比经典的Hotelling'sT^2与Friedman检验法, $\hat{D}_n^{(k)}$方法能更准确地捕捉具有相关性的变量在除了均值或中位数以外的其他特征上的变化. 而与DISCO方法直接应用于多个配对变量的分布差异检验相比, $\hat{D}_n^{(k)}$方法能合理地控制第一类错误, 并且具有更高的功效.

2) $p > 1$情形

下面考虑k个相关变量均为多元变量的情形, 具体维数为$p = 2, 4, 6, 8$四种情形, 同样设定$k = 4$. 首先假设$(X_1, X_2, \cdots, X_{4p}) \sim N_{4p}(\mu, \Sigma)$, 其中$\mu = (0, 0, \cdots, 0)', \Sigma = (\sigma_{ij}), \sigma_{ij} = 1, i = j, \sigma_{ij} = 0.3^2, i \neq j$, 然后再以以下方式生成$Y_1, Y_2, Y_3, Y_4$.

$$Y_1 = (\exp(X_1), \cdots, \exp(X_p)),$$

$$Y_2 = (\exp(X_{p+1}), \cdots, \exp(X_{2p})),$$

$$Y_3 = (\exp(X_{2p+1}), \cdots, \exp(X_{3p})),$$

$$Y_4 = (\exp(X_{3p+1} + a), \cdots, \exp(X_{4p} + a)).$$

其中$a = 0, 0.05, 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$. 显然当$a = 0$时, 四个相关向量同分布; 当$a \neq 0$时, Y_4与前三个向量分布不同, 并且差异随a的增大而增大. 考虑样本容量$n = 50$, $n = 100$与$n = 200$三种情形, 图1给出了$\hat{D}_n^{(k)}$与DISCO检验法在不同a取值下的第一类错误与功效.

从图1可以得到以下几点发现. 1. 当$a = 0$, 样本量较少$(n = 50)$时, $\hat{D}_n^{(k)}$与DISCO方法在维数$p > 2$时的第一类错误均偏小于名义水平$\alpha = 0.05$, 样本量增加到200时, $\hat{D}_n^{(k)}$在各维数下的第一类错误均逐渐接近0.05, 而DISCO方法的第一类错误仍低于名义水平, 并且非常接近0; 2. 在相同样本容量下, $\hat{D}_n^{(k)}$与DISCO方法的功效均随a的增大而增大, 呈S型, 这是因为Y_4与Y_1, Y_2, Y_3之间的分布差异随a的增大而增大; 3. 在相同容量与相同的a下, $\hat{D}_n^{(k)}$与DISCO方法的功效随着维数p的增大而增大, 这是合理的, 因为在这里设计的例子中, Y_4与Y_1, Y_2, Y_3差异在于每个分量都增加a, 因此, 维数越高, 分布差异越大, 但第一类错误会越来越低于名义水平, 需要越来越大的样本量来保证第一类错误的准确性.

总体而言, 在本文设计的具有相关性的多维变量的分布差异检验的例子中, 与前面一维情形的结果类似, DISCO方法虽然在功效上与$\hat{D}_n^{(k)}$方法的表现接近, 但它没能将第一类错误合理地控制在名义水平附近. 这再一次说明, 对于具有相关性的随机变量的分布差异检验问题, 忽略变量之间的相关性直接采有DISCO方法是不准确的.

§5 实例分析

HIV数据集记录了120个艾滋病患者基准期与接受3种不同方案的治疗后的第2、4、6、8、10个月的CD4细胞数, 数据共4个变量: id (个体标识), month (月份), CD4 (CD4细胞计数), group (组, 哑元变量, 其中1=控制组, 2=单独药物治疗, 3= 药物混合治疗). 原始数据及说明可分别从`http://www.biostat.jhsph.edu/fdominic/teaching/LDA/hivstudy.raw`与`https://www.biostat.jhsph.edu/fdominic/teaching/LDA/README`下载. 此处, 笔者关心这120个患者的CD4细胞数随月份的变化情况. 从图2可以看到, 患者的CD4细胞数从基准期到治疗后4月这段时间内逐渐下降, 之后6月、8月和10月的CD4细胞数均值或中位数无明显变化, 但很显然, CD4细胞数的方差逐渐增大. 对4月、6月、8月和10月这四次重复测量的CD4细胞数的分布差异进行检验, 表2列出了Hotelling's T^2检验、Friedman检验、DISCO检验法与本文提

出的$\hat{D}_n^{(k)}$检验法的结果, 其中DISCO方法与$\hat{D}_n^{(k)}$检验法的p-值都是基于399次重抽样程序计算得到. 从表2中可以看到, 经典的Hotelling's T^2检验与Friedman检验方法都无法鉴别这4次CD4细胞数在方差上的变化, DISCO方法忽略了4次CD4细胞数之间的相关性, 因此也未能准确地鉴别这4次CD4细胞数在方差上的变化, 而新方法$\hat{D}_n^{(k)}$能准确地检测出这4次CD4细胞数由于方差上的变化而导致的分布差异.

表 2 患者在4月、6月、8月和10月这4次重复测量的CD4细胞数的分布差异检验结果

检验方法	统计量	p-值
Hotelling's T^2	1.319623	0.2714
Friedman	2.9298	0.4026
DISCO	1.638	0.1050
$\hat{D}_n^{(k)}$	2286.35	0.0025

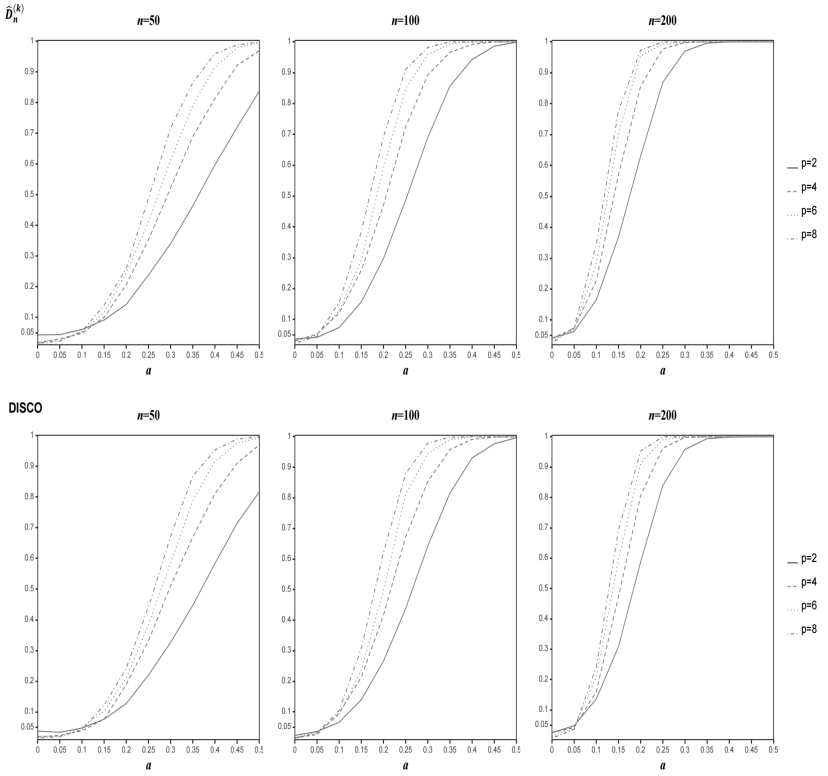

图 1 $p>1$时, $\hat{D}_n^{(k)}$与DISCO方法在检验四个相关变量Y_1, Y_2, Y_3, Y_4分布差异中的第一类错误与功效, 其中上半部分为$\hat{D}_n^{(k)}$方法的结果, 下半部分为DISCO方法的结果

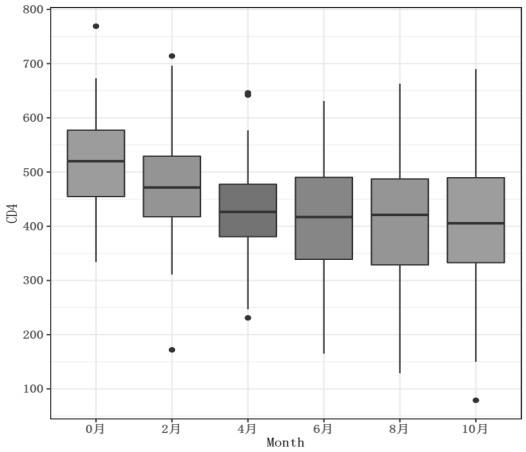

图 2 120个艾滋病患者的CD4 细胞数变化

§6 小结

本文基于能量距离的概念, 提出了一种新的k-配对样本分布差异的检验法, 并讨论了方法的渐近性质. 数值模拟与实际数据分析均表明, 相比传统的Hotelling's T^2检验与Friedman检验方法, 新的方法能更准确地鉴别k个相关变量在除了位置之外的其他特征比如方差上的差异, 并且在检验具有相关性的多维向量的分布差异中表现也不错. 故新方法能适用于更广泛的数据类型.

参考文献:

[1] 申希平, 祁海萍, 刘小宁. Friedman M 检验平均秩的多重比较在SPSS 软件的实现[J]. 中国卫生统计, 2013, 30(4): 611-613.

[2] 马蕊, 张爱霞, 生庆海. Friedman 检验和Kramer 检验在感官排序测试中的比较[J]. 中国乳品工业, 2007, 35(9): 14-16

[3] 李兴国, 赵晓冬. 中国大学评价体系相关性和稳定性的统计学检验[J]. 统计与决策, 2018, 23: 103-105

[4] 梁鑫, 谢佳利, 邵延会. 国内主要城市空气质量统计分析[J]. 数理统计与管理, 2009, 28(3): 550-554.

[5] Rizzo M L, Székely G J. Disco analysis: A nonparametric extension of analysis of variance[J]. The Annals of Applied Statistics, 2010, 4(2): 1034-1055.

[6] Martínez-Camblor P, De Uña-Álvarez J, Corral N. K-Sample test based on the common area of kernel density estimators[J]. Journal of Statistical Planning and Inference, 2008, 138: 4006-4020.

[7] Martínez-Camblor P. Nonparametric k-sample test based on kernel density estimator for paired design[J]. Computational Statistics and Data Analysis, 2010, 54(8): 2035-2045.

[8] Székely G J. E-Statistics: The energy of statistical samples[R]. Bowling Green State University, Department of Mathematics and Statistics Technical Report, 2003 (03-05): 2000-2003.

[9] Rizzo M L. A test of homogeneity for two multivariate populations[C]. Proceedings of the American Statistical Association, Physical and Engineering Sciences Section, 2002.

[10] Székely G J, Rizzo M L. A new test for multivariate normality[J]. Journal of Multivariate Analysis, 2005, 93(1): 58-80.
[11] Székely G J, Rizzo M L. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method[J]. Journal of Classification, 2005, 22(2): 151-183.
[12] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances[J]. The Annals of Statistics, 2007, 35(6): 2769-2794.
[13] Székely G J, Rizzo M L. Energy statistics: A class of statistics based on distances[J]. Journal of Statistical Planning and Inference, 2013, 143(8): 1249-1272.
[14] Chen Minqiong, Tian Ting, Zhu Jin, et al. Paired-sample tests for homogeneity with/without confounding variables[J]. Statistics and its Interface, 2022, 15: 335-348.
[15] Lee A J. U-Statistics: Theory and Practice[M]. New York: Marcel Dekker, Inc., 1990.

K-sample test for paired design based on energy distance

CHEN Min-qiong[1], TAN He-li[2]

(1. Department of Artificial Intelligence and Data Science, Guangzhou Xinhua University, Guangzhou 510520, China;
2. School of Financial Mathematics and Statistics, Guangdong University of Finance, Guangzhou 510521, China)

Abstract: In this paper, a new method based on the concept of energy distance is proposed to test the equality of k paired distributions. First, a measure of the distribution difference of k correlated variables and its sample estimator are introduced. The estimator has the form of a V-statistic. Then, using the theory of V-statistics, the asymptotic properties of the estimator are discussed. A bootstrap resample procedure for the test is also provided and the rationality of the procedure is verified. Both numerical simulations and real data analysis show that, compared with the classical Hotelling's T^2 test and Friedman test, the new method can detect the differences of k correlated variables in other characteristics except location more accurately, and is applicable to multivariate variables, and thus can be applied to a wider range of data types.

Keywords: energy distance; k paired samples; test for distributional difference; Hotelling's T^2 test; Friedman test; V-statistic; bootstrap

MR Subject Classification: 62G99