系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
39(1) (2019, 1), 106–119

改进的基于图方法对真实原假设比例的估计*

崔甲蓉 刘佳敏 许王莉

(中国人民大学统计学院, 北京 100872)

摘要 在多重假设检验中, 对原假设为真的比例或数量的估计是一个基础性问题. 一类估计方法是基于排序 p 值的图方法, 由 Schweder 和 Spjøtvoll(1982) 首次提出. 但是这类方法只利用了特定切点 p 值的信息, 当原假设和备择假设差异较小时, 现有方法的有效性会降低. 文章通过利用所有原假设 p 值的信息, 对估计进行了改进. 由于利用了更多的信息, 改进后的估计更有效. 数值模拟结果表明, 文章所提出的估计方法比现有方法有效性更好, 最后将所提出的方法应用到实际数据分析中.

关键词 多重假设检验, 真实原假设的比例或数量, 排序 p 值.

MR(2000) 主题分类号 62-09, 62H15

Improved Graphical Based Estimators for the Proportion of True Null Hypothesis

CUI Jiarong LIU Jiamin XU Wangli

(*School of Statistics, Renmin University of China, Beijing* 100872)

Abstract Estimation for proportion or number of true null hypotheses is an essential task in multiple hypothesis test. One type of estimating methods is based on the graphic of the sorted p-values first proposed by Schweder and Spjøtvoll(1982). But these methods only utilize the information of a specific cutpoint p-value, existing methods may lose efficiency when the deviation between null hypothesis and alternative hypothesis is small. In this paper, we improve estimating methods by taking all information of p-values on the null hypotheses. Improved methods are more accurate and efficient than corresponding original methods due to more information can be used. Compared with the existing methods, the accuracy of the proposed methods is competitive through numerical study.

* 国家自然科学基金 (11471335) 及教育部人文社会科学重点研究基地重大项目 (16JJD910002) 资助课题.
收稿日期: 2018-06-15, 收到修改稿日期: 2018-09-01.
编委: 邹国华.

Keywords Multiple hypothesis test, the proportion or number of true null hypotheses, sorted p-values.

1 引 言

在多重假设检验中, 假定有 m 个原假设 $H_{01}, H_{02}, \cdots, H_{0m}$, 其中 m_0 个原假设为真, 通常需要考虑控制第一类错误. 族错误率 (FWER: Familywise error rate) 和错误发现率 (FDR: False discovery rate) 常被用来衡量第一类错误的概率, 但是 Storey[1] 指出, 当大部分原假设不成立时, 它们存在过于保守的问题. Hochberg 和 Benjamini[2] 以及 Benjamini 和 Hochberg[3] 分别提出了改进的 FWER 控制方法和 FDR 控制方法来克服保守性. 这两个改进方法都通过用原假设为真的个数 m_0 代替 m 来控制 FWER 和 FDR. 由于 m_0 未知, 给出 m_0 的估计是很必要的. 此外, 对原假设为真的个数 m_0 或比例 π_0 的估计也被应用于对 DNA 微阵列中不存在差异表达的基因比例的研究 [4], 功能性核磁共振的影像研究 [5], 天体物理图像中的源探测问题 [6] 等. 因此, 估计 m_0 在理论和实践方面都具有重要意义.

已经有许多学者对如何估计 m_0 (或 π_0) 进行了研究. 例如, Pounds 和 Morris[7] 在参数贝塔 - 均匀分布混合模型中估计出 π_0 的上界; Guan 等 [8] 根据统计量在原假设和备择假设下的分布, 提出了基于模型的方法, 得到 π_0 的极大似然估计; Markitsis 和 Lai[9] 通过给定一个切点, 删除小于该切点的 p 值, 改进了贝塔 - 均匀分布混合模型, 得到 π_0 的估计. 除此之外也有一些非参数的估计方法. Nettleton 等 [10] 依赖于观测到的 p 值直方图, 提出用迭代算法估计 m_0; 基于经验 Bayes 方法, Efron 等 [11] 估计出 p 值的经验密度函数 $f(p)$, 通过最小化和平滑密度函数 $f(p)$, 得到 π_0 的估计; Storey[1] 基于 Bayes 原理构造估计量 $\widehat{m}_0^{\mathrm{ST}} = \sum_{j=1}^{m} \frac{I(p_j > \lambda)}{1-\lambda}$, 其中 λ $(0 < \lambda < 1)$ 是调和参数; Storey 和 Tibshirani[12] 提出了三次样条平滑方法估计 π_0; Storey 等 [13] 使用 Bootstrap 方法估计 λ 从而修正了估计量 $\widehat{m}_0^{\mathrm{ST}}$; Langaas 等 [4] 基于递减密度估计和凸减密度估计给出 π_0 的一种新的估计; Jiang 和 Doerge[14] 提出将不同 λ 对应的估计 $\widehat{m}_0^{\mathrm{ST}}$ 结合起来的方法给出参数的估计.

另外一类非参数方法是基于由 Schweder 和 Spjϕtvoll[15] 首次提出的图方法. 主要的想法是基于等式 $E(F(p)) = m_0(1-p)$, 其中 $F(p) = \sum_{j=1}^{m} I(p_j > p)$, p_j 是第 j 个假设检验的 p 值. 根据原假设下 $E(F(p))$ 和 p 之间的线性关系估计参数 m_0. 之后, 通过修正图方法, Benjamini 和 Hochberg[3] 提出, 找到排序后的 p 值 $p_{(1)}, p_{(2)}, \cdots, p_{(m)}$ 的切点 $p_{(J)}$, 如果 $p_j \geq p_{(J)}$ 成立, 则原假设 H_{0j} 为真, 通过 $E(p_{(J)})$ 与 m_0 之间的关系估计 m_0. Benjamini 和 Hochberg[3], Hwang 等 [16] 以及 Qiao 等 [17] 分别提出了不同的标准来找到切点 $p_{(J)}$, 基于切点 $p_{(J)}$ 给出 m_0 的估计. 但是以上估计方法全都忽略 $p_{(j)}$ $(j = J+1, J+2, \cdots, m)$ 所包含原假设成立的信息. 一个很自然的想法是, 充分利用 $p_{(j)}$ $(j = J+1, J+2, \cdots, m)$ 估计参数 m_0 可以提高估计的有效性.

本文主要目的是基于已有的图方法, 对多重假设检验中真实原假设比例的估计进行改进. 在第 2 节, 我们简要回顾已有的 4 种估计方法, 分别记为 $\widehat{m}_0^{\mathrm{BH}}, \widehat{m}_0^{\mathrm{SD}}, \widehat{m}_0^{\mathrm{AC}}, \widehat{m}_0^{\mathrm{DIS}}$, 然后提出改进的方法. 第 3 节基于蒙特卡洛模拟, 比较 4 种 m_0 估计的中位数和均方误差. 最后, 利用实际数据研究来说明改进估计的有效性.

2 已有估计以及提出的方法

首先,我们给出一个简单的例子来直观说明估计 m_0 的图方法. 假定样本数据 $X_1, X_2, \cdots, X_n$ 来自于多元正态分布 $N(\mu, \Sigma)$, 这里 $X_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}$, $\mu = (\mu_1, \mu_2, \cdots, \mu_m)^{\mathrm{T}}$ 是 $m \times 1$ 维向量, μ 的 m_0 项分量为 0, $m - m_0$ 项分量为 0.2. $\Sigma = I$, I 是 $m \times m$ 维单位矩阵. 令 $n = 50$, $m = 11$, $m_0 = 4$, 待检验的 m 个原假设为

$$H_{0j} : \mu_j = 0, \quad j = 1, 2, \cdots, m = 11. \tag{2.1}$$

对于这个假设检验问题, 统计量为 $z_j = \sqrt{n} \cdot |\overline{x}_j|$, 这里 $\overline{x}_j$ 是 $(x_{j1}, x_{j2}, \cdots, x_{jn})$ 的样本均值. 对应的 p 值为 $p_j = 2[1 - \Phi(z_j)]$, $\Phi(\cdot)$ 是标准正态分布的累积分布函数. 进行一次模拟, 得到 $m = 11$ 个检验的排序 p 值为: 0.0250, 0.0279, 0.0386, 0.0473, 0.1170, 0.1232, 0.1621, 0.4732, 0.9008, 0.9396, 0.9597. 图 1 中展示了排序的 p 值.

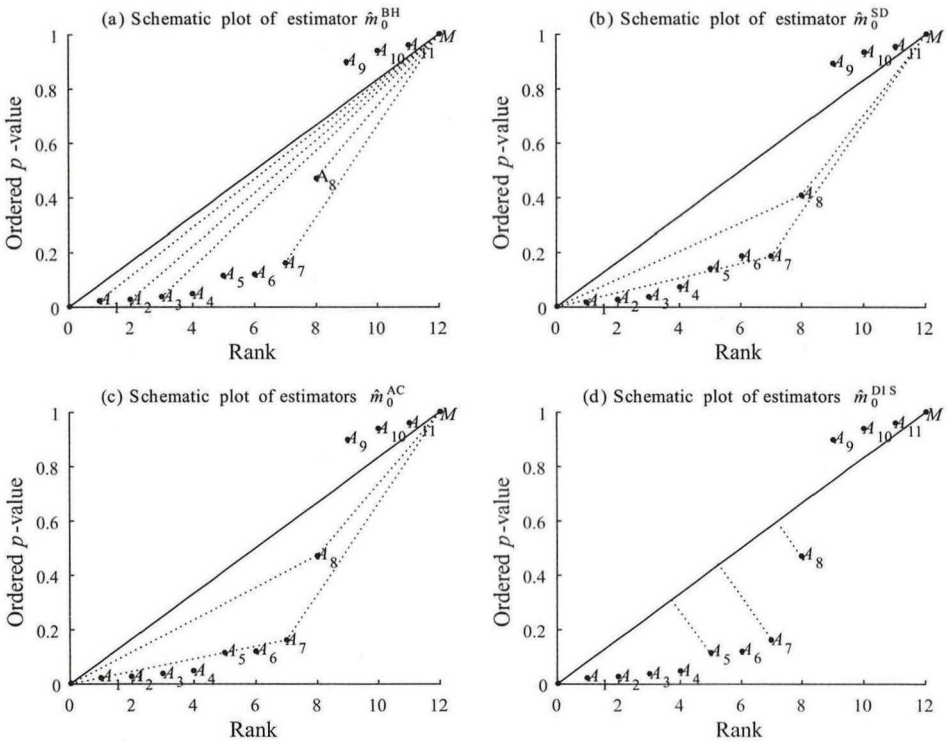

图 1 估计 m_0 的图方法

(Figure 1 The graphical methods for estimating m_0)

在图 1 中, 我们将点 $(0,0)$, $(j, p_{(j)})$, $(m+1, 1)$ 分别记为 O, A_j 和 M. 当全部原假设为真时 ($m_0 = m$), p_j 可以看作 $[0,1]$ 区间上均匀分布的样本. 因此 $p_{(j)}$ 的期望为 $E(p_{(j)}) = \frac{j}{m+1}$. 易知 $E(p_{(j)})$ 和 j 之间存在线性关系, 因此对应于真实原假设计算得到的点 A_j, 应该在直线 OM 附近. 图 1 中的子图 (a), (b), (c), (d) 对应现有的 4 种确定切点 $p_{(J)}$ 的方法. 在下一小节中, 我们将分别基于图 1 中的子图, 对现有估计方法进行介绍.

2.1 现有估计 m_0 方法综述

Benjamini 和 Hochberg[3] 提出的估计方法, 其想法基于当原假设 H_{0j} 为假时, 图 1 (a) 中直线 A_jM 的斜率 S_j^U 会随着 j 的增大而增大; 而当原假设 H_{0j} 为真时, 直线 A_jM 的斜率 S_j^U 会随着 j 的增大而减小. 切点 $p_{(J)}$ 定义为斜率 S_j^U $(j=1,2,\cdots,m)$ 第一次减小时对应的 p 值. 基于 $p_{(J)}$ 得到 m_0 的估计记为 $\widehat{m}_0^{\mathrm{BH}}$. Hwang 等 [16] 提出当原假设为假时, 图 1 (b) 中直线 A_jM 的斜率 S_j^U 和直线 OA_j 的斜率 S_j^D 的差值会随着 j 的增大而增大; 当原假设为真时, 趋势则相反. 基于这一想法, 他们提出估计量 $\widehat{m}_0^{\mathrm{SD}}$. Hwang 等 [16] 提出, 当原假设 H_{0j} 为真时, 图 1 (c) 中 $\angle A_j$ 更大且接近于 π. 因此基于 $\angle A_j$ 的变化提出了估计量 $\widehat{m}_0^{\mathrm{AC}}$. Qiao 等 [17] 根据图 1 (d) 中点 A_j 到直线 OM 的垂直距离 d_j 提出了 m_0 的估计量 $\widehat{m}_0^{\mathrm{DIS}}$. 上述 4 种方法寻找切点 $p_{(J)}$ 的具体步骤如下

估计 $\widehat{m}_0^{\mathrm{BH}}$ 的切点 J 计算步骤

1) 计算 $S_j=\frac{1-p_{(j)}}{m+1-j}$, $j=1,2,\cdots,m$;

2) 从 $j=1$ 开始, 如果 $S_{j-1}\le S_j$, j 继续增大, 直到找到第一个指标 J, 使得 $S_J<S_{J-1}$.

估计 $\widehat{m}_0^{\mathrm{SD}}$ 的切点 J 计算步骤

1) 令 $k=\max\{j:p_{(j)}<\lambda\},\lambda=0.5$;

2) 计算 $S_j^U=\frac{1-p_{(j)}}{m+1-j}$ 和 $S_j^D=\frac{p_{(j)}}{j}, j=1,2,\cdots,k$, J 的定义为 $J=\max\{j:S_j^U-S_j^D, j=1,2,\cdots,k\}$.

估计 $\widehat{m}_0^{\mathrm{AC}}$ 的切点 J 计算步骤

1) 计算 $\angle A_j$, $j=1,2,\cdots,m$;

$$\angle A_j=\arccos\left(\frac{\overline{OA_j}^2+\overline{A_jM}^2-\overline{OM}^2}{2\cdot\overline{OA_j}\cdot\overline{A_jM}}\right).$$

2) 从 $j=1$ 开始, 如果 $\angle A_j\le\angle A_{j-1}$, j 继续增大, 直到找到第一个指标 J, 使得 $\angle A_J>\angle A_{J-1}$.

估计 $\widehat{m}_0^{\mathrm{DIS}}$ 的切点 J 计算步骤

1) 计算 d_j, $j=1,2,\cdots,m$;

$$d_j=\frac{|j-(m+1)p_{(j)}|}{\sqrt{1+(m+1)^2}}.$$

2) 从 $j=1$ 开始, 如果 $d_j>d_{j-1}$, j 继续增大, 直到找到第一个指标 J, 使得 $d_J<d_{J-1}$.

对于上述 4 种不同方法得到的切点 $p_{(J)}$, 参数 m_0 的估计定义为

$$\widehat{m}_0=\min\left\{\left\lceil\frac{1}{S_J}\right\rceil-1,m\right\},\quad S_J=\frac{1-p_{(J)}}{m+1-J}. \tag{2.2}$$

2.2 改进的图方法估计参数 m_0

上述提到的 4 种方法利用公式 (2.2) 估计 m_0, 都只利用了 $(J,p_{(J)})$ 一个点的信息. 直观上, 切点 $p_{(J)}$ 是原假设为真和原假设为假的分界, $p_{(j)}$ $(j=J,J+1,\cdots,m)$ 都是来自原假设为真的 p 值, 接下来我们说明它们都可以用来估计 m_0.

如果原假设 H_{0j} 为真, 则有 $p_j\sim U(0,1)$. 均匀分布概率密度函数为 $f(p)=I\ (0<p<1)$,

I 为示性函数, 累积分布函数 $F(p)$ 为

$$F(p)=\begin{cases}0, & p<0,\\ p, & 0\le p\le 1,\\ 1, & p\ge 1.\end{cases} \tag{2.3}$$

对于假设检验个数为 m 的多重检验问题, 当不同假设之间的样本相互独立时, 对应的 p 值也是相互独立的. 相应地得到排序 p 值 $p_{(j)}, j=m-m_0+1, m-m_0+2, \cdots, m$ 的概率密度函数 $\widetilde{f}_j(p)$ 如下

$$\begin{aligned}\widetilde{f}_j(p)&=\frac{m_0!}{(j-m+m_0-1)!(m-j)!}[F(p)]^{j-m+m_0-1}[1-F(p)]^{m-j}f(p)\\ &=\frac{m_0!}{(j-m+m_0-1)!(m-j)!}p^{j-m+m_0-1}(1-p)^{m-j}\\ &=\frac{\Gamma(m_0+1)}{\Gamma(j-m+m_0)\Gamma(m-j+1)}p^{j-m+m_0-1}(1-p)^{m-j}.\end{aligned} \tag{2.4}$$

因此, 在 m 个假设检验问题中, 如果 m_0 个原假设为真, $p_{(j)}\sim Beta(j-m+m_0, m-j+1)$ $(j=m-m_0+1, m-m_0+2, \cdots, m)$, 且有

$$E(p_{(j)})=\frac{j-m+m_0}{m_0+1}=1-\frac{m+1-j}{m_0+1}. \tag{2.5}$$

根据公式 (2.5), 容易得到 $g_0=\frac{1}{m_0+1}$ 的估计

$$\widehat{g}_{0j}=\frac{1-p_{(j)}}{m+1-j},\quad j=m-m_0+1, m-m_0+2, \cdots, m. \tag{2.6}$$

利用现有方法找到的切点指标 J, 对 $j=J, J+1, \cdots, m$, 可以通过公式 (2.6) 给出 g_0 的估计. 最终 g_0 的估计为 $\widehat{g}_{0j}$ 的平均, 即

$$\widehat{g}_0=\frac{1}{m+1-J}\sum_{j=J}^{m}\frac{1-p_{(j)}}{m+1-j}. \tag{2.7}$$

应用上述公式 (2.7) 得到估计量 $\widehat{m}_0$. 我们将改进的 4 种方法 m_0^{BH}, m_0^{SD}, m_0^{AC}, m_0^{DIS} 的估计分别记作 $\widehat{m}_0^{\mathrm{IBH}}$, $\widehat{m}_0^{\mathrm{ISD}}$, $\widehat{m}_0^{\mathrm{IAC}}$, $\widehat{m}_0^{\mathrm{IDIS}}$. 改进的估计定义为

$$\widehat{m}_0=\min\left\{\left\lceil\frac{1}{\widehat{g}_0}\right\rceil-1, m\right\}. \tag{2.8}$$

这里, $\lceil x\rceil$ 表示对 x 向上取整.

当不同假设之间的样本不独立时, p 值也不满足独立性假定. 由于 p 值次序统计量 $p_{(j)}$ 与协方差矩阵 Σ 有关, 因此很难得出 $p_{(j)}$ 的概率密度函数去计算 $E(p_{(j)})$. 在不独立的情形下, 我们通过模拟来计算 $\widehat{g}_{0j}$, 从而说明 $\widehat{g}_{0j}$ 与真实值 $\frac{1}{m_0+1}$ 很接近. 除了协方差矩阵 Σ 和 $m_0=m$, 其他参数与原假设 (2.1) 中设置相同. 协方差矩阵 Σ 是 $m\times m$ 维矩阵, 我们考虑了 3 种常见的协方差矩阵. 1) AR(1) 相关矩阵, 即对角线元素为 1, 非对角线元素为 $\sigma_{i,j}=0.6^{|i-j|}, 1\le i,j\le m$. 2) 三对角相关矩阵, 即对角线元素为 1, 若 $|i-j|=1$, $\sigma_{i,j}=0.4$, 若 $|i-j|\ne 1$, $\sigma_{i,j}=0$. 3) 复合对称相关矩阵, 即对角线元素为 1, 非对角线元素为 0.4. 令 $m=m_0=11$, 重复模拟 10000 次, 表 1 为 3 种不同协方差结构下 $\widehat{g}_{0j}$ 的值, 模拟结果与真实值 $\frac{1}{m_0+1}=0.0833$ 非常接近. 因此, 改进的 m_0 估计对协方差矩阵不是很敏感, 对于样本不独立的情形仍然适用.

表 1 协方差矩阵 Σ 不同形式下 g_{0j} 的估计

(Table 1 Estimation of g_{0j} under various dependent patterns Σ)

协方差矩阵 Σ	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
AR(1) 相关	0.0819	0.0821	0.0827	0.0834	0.0844	0.0853
三对角相关	0.0831	0.0834	0.0837	0.0838	0.0841	0.0845
复合对称相关	0.0816	0.0813	0.0817	0.0825	0.0835	0.0849
协方差矩阵 Σ	$j=7$	$j=8$	$j=9$	$j=10$	$j=11$	
AR(1) 相关	0.0862	0.0870	0.0883	0.0886	0.0891	
三对角相关	0.0845	0.0844	0.0845	0.0851	0.0852	
复合对称相关	0.0867	0.0888	0.0912	0.0935	0.0961	

3 数值研究

3.1 模拟分析

在这一小节, 我们研究用模拟分析来研究提出估计的有效性. 模拟分析中, 改进的估计分别记作 $\widehat{m}_0^{\mathrm{IBH}}$, $\widehat{m}_0^{\mathrm{ISD}}$, $\widehat{m}_0^{\mathrm{IAC}}$, $\widehat{m}_0^{\mathrm{IDIS}}$, 对应相应的已有方法估计为 $\widehat{m}_0^{\mathrm{BH}}$, $\widehat{m}_0^{\mathrm{SD}}$, $\widehat{m}_0^{\mathrm{AC}}$, $\widehat{m}_0^{\mathrm{DIS}}$. 通过中位数和均方误差来比较提出的估计和已有方法的估计. 均方误差定义为: $\mathrm{RMSE}=(\sum_{i=1}^{K}\frac{(\widehat{m}_{0,i}-m_0)^2}{K})^{\frac{1}{2}}$, 这里 K 是模拟次数. 模拟分析分为 4 部分: 单样本检验, 两样本检验, 非正态数据检验以及数据维数大于样本维数的情形.

例 1 (单样本检验的情形) 数据 $X_1, X_2, \cdots, X_n$ 是来自于多元正态分布总体 $N(\mu, \Sigma)$ 的样本, 这里 $X_i=(x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}$, $i=1,2,\cdots,n$, $\mu=(\mu_1,\mu_2,\cdots,\mu_m)^{\mathrm{T}}$ 是 $m\times 1$ 维向量, Σ 是 $m\times m$ 维的协方差矩阵. 待检验的原假设为

$$H_{0j}:\mu_j=0,\quad j=1,2,\cdots,m.$$

关于参数向量 μ 的设置, 不失一般性, 令 μ 的前 m_0 项分量为 0, 其余 $m-m_0$ 项分量为 a. 协方差矩阵设置为两种类型: $\Sigma=I$ 表示样本独立的情形, $\Sigma=D$ 表示样本不独立的情形. 这里矩阵 I 是单位矩阵, 矩阵 D 的对角线元素都是 1, 非对角线元素为 $\sigma_{i,j}=0.6^{|i-j|}, 1\le i,j\le m$.

首先, 在 $m=100$, $m_0=75$, $n=50$, $\Sigma=I$ 情况下, 图 2 展示了参数 a 和 m_0 不同估计之间的关系. 参数 a 越小, 说明原假设和备择假设之间差异越小. 当参数 a 很小的时候, 已有方法估计 m_0 的精确度很差, 但是改进后的方法可以提高估计的精确度. 从图 2 中可以看出, 改进后的估计比已有方法估计更快地收敛到真实值 m_0. 随着 a 逐渐增大, 所有方法得到的估计都比较好.

在 $m=100$, $m_0=25,50,75$, $n=50,100$ 的情形下, 模拟结果的趋势与图 2 中呈现的趋势完全类似, 即, 当原假设和备择假设之间的差异 a 相对大的时候 ($a\ge 0.5$), 不同方法得到 m_0 的估计一致地收敛到真实值 m_0. 因此, 我们固定 $a=0.15,0.25,0.35,0.45$, 研究不同方法得到 m_0 的估计. 表 2 和表 3 展示了模拟研究的结果.

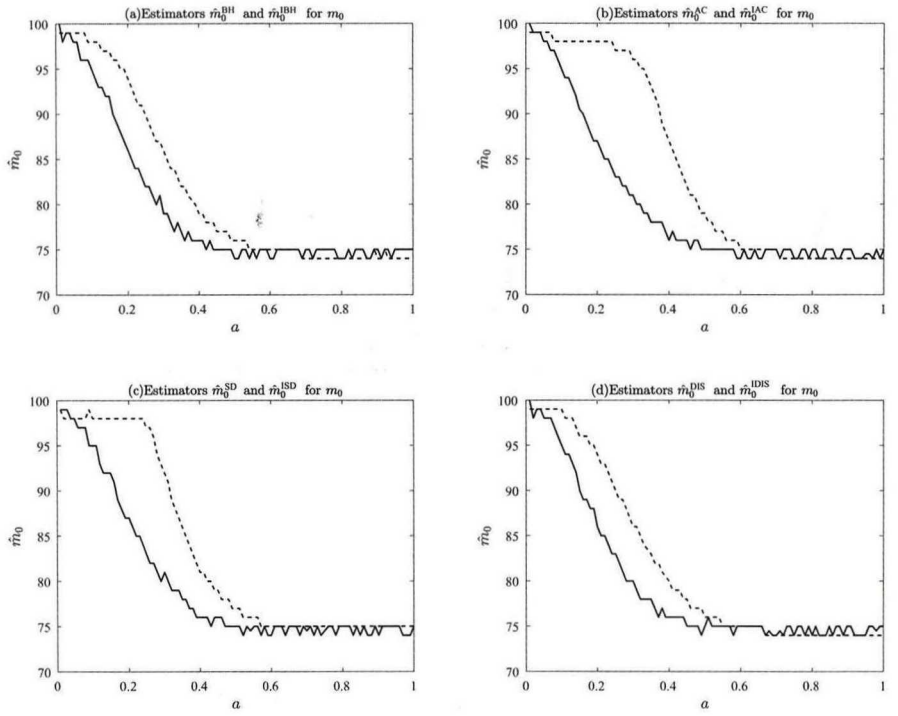

图 2 参数 m_0 的估计 ($m = 100$, $n = 50$, $m_0 = 75$, 协方差矩阵 $\Sigma = I$), 虚线和实线分别表示已有方法的估计和相应改进后的估计

(Figure 2 Estimators for m_0 under $m = 100$, $n = 50$ and $m_0 = 75$ with covariance matrix I, the dashed and solid lines denote the original and proposed estimators respectively)

表 2 单样本检验 ($m = 100, n = 50$) 中 m_0 的估计

(Table 2 Performance of methods under one sample when $m = 100, n = 50$)

Σ	m_0	a	$\widehat{m}_0^{BH}$	$\widehat{m}_0^{IBH}$	$\widehat{m}_0^{DIS}$	$\widehat{m}_0^{IDIS}$	$\widehat{m}_0^{AC}$	$\widehat{m}_0^{IAC}$	$\widehat{m}_0^{SD}$	$\widehat{m}_0^{ISD}$
I	25	0.15	88(63.06)	74(49.70)	89(63.80)	74(49.77)	98(72.14)	76(51.53)	95(63.03)	75(49.70)
		0.25	61(36.74)	47(23.50)	66(41.98)	49(24.83)	92(65.15)	55(30.49)	43(20.11)	44(21.36)
		0.35	38(14.23)	32(9.12)	45(21.45)	33(9.96)	46(23.90)	34(10.52)	28(6.17)	30(9.66)
		0.45	28(4.97)	26(5.89)	33(9.02)	27(5.17)	30(6.49)	27(5.55)	23(4.34)	26(7.11)
	50	0.15	93(42.54)	82(33.56)	93(42.95)	82(33.60)	98(47.41)	83(34.24)	98(45.83)	83(33.99)
		0.25	77(27.52)	65(17.32)	79(29.32)	65(17.35)	96(45.02)	69(20.25)	67(22.98)	64(17.26)
		0.35	61(12.63)	55(8.81)	64(15.37)	55(8.74)	74(26.19)	57(9.74)	54(8.87)	54(10.04)
		0.45	54(5.09)	51(7.42)	56(6.61)	51(7.01)	57(8.78)	51(6.85)	50(6.26)	51(8.37)
	75	0.15	97(21.25)	91(16.97)	97(21.30)	91(16.99)	98(22.89)	91(17.20)	98(22.38)	91(17.25)
		0.25	90(15.18)	82(11.54)	90(15.56)	83(11.40)	97(21.98)	83(11.52)	97(20.98)	83(11.49)
		0.35	82(8.15)	77(8.92)	83(8.58)	77(8.75)	93(17.86)	78(8.38)	86(12.41)	78(8.92)
		0.45	77(3.84)	75(8.76)	78(3.98)	75(8.49)	82(8.51)	76(7.98)	79(5.96)	75(8.45)

续表 2 单样本检验 ($m=100, n=50$) 中 m_0 的估计

(Table 2 Performance of methods under one sample when $m=100, n=50$ (Continued))

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
	25	0.15	89(62.84)	74(50.23)	96(69.28)	86(60.72)	98(72.03)	75(52.08)	92(59.79)	74(49.66)
		0.25	60(37.58)	47(24.70)	85(59.73)	68(45.50)	93(63.95)	54(30.75)	42(22.57)	43(22.55)
		0.35	37(15.37)	32(10.34)	68(45.07)	50(29.28)	45(27.93)	34(12.32)	28(7.61)	30(10.66)
		0.45	28(5.80)	26(6.92)	52(30.26)	38(17.25)	29(7.59)	27(6.57)	23(5.30)	26(7.92)
	50	0.15	93(42.43)	82(34.00)	98(45.83)	90(39.69)	98(47.47)	83(34.70)	97(43.15)	83(34.12)
D		0.25	77(28.14)	65(18.50)	91(40.26)	78(31.31)	96(44.09)	68(20.72)	66(25.18)	65(19.70)
		0.35	61(13.52)	55(10.33)	80(31.52)	66(21.79)	74(28.03)	57(11.09)	53(10.57)	54(11.37)
		0.45	53(5.87)	51(8.84)	69(21.77)	58(15.16)	57(9.88)	52(8.17)	50(7.54)	51(9.84)
	75	0.15	97(21.12)	91(17.25)	99(22.39)	94(18.94)	98(23.01)	91(17.25)	98(21.72)	91(17.45)
		0.25	90(15.54)	82(12.53)	96(20.20)	88(16.78)	97(21.72)	83(12.55)	96(19.67)	83(12.94)
		0.35	82(8.91)	77(10.47)	91(16.76)	83(14.77)	93(17.68)	78(9.69)	85(13.62)	78(10.81)
		0.45	77(4.71)	75(10.17)	85(12.65)	79(13.52)	82(9.75)	75(9.35)	78(7.64)	76(10.31)

注: 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

表 3 单样本检验 ($m=100, n=100$) 中 m_0 的估计

(Table 3 Performance of methods under one sample when $m=100, n=100$)

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
	25	0.15	73(48.05)	57(32.87)	76(51.11)	58(33.63)	97(70.64)	63(38.31)	52(31.43)	53(30.24)
		0.25	37(13.70)	31(8.77)	45(21.04)	33(9.75)	45(23.00)	33(10.28)	28(5.95)	30(9.44)
		0.35	27(3.42)	26(5.76)	30(5.69)	26(4.80)	27(3.87)	26(5.28)	23(4.45)	26(6.79)
		0.45	25(2.08)	25(5.58)	25(1.46)	25(4.92)	25(1.85)	25(5.30)	22(4.60)	25(6.12)
	50	0.15	84(34.17)	71(23.02)	85(35.42)	72(23.40)	97(46.70)	74(25.19)	86(35.17)	72(23.65)
I		0.25	61(12.36)	55(8.93)	64(15.09)	55(8.74)	74(25.80)	57(9.58)	53(8.65)	54(9.96)
		0.35	52(3.57)	51(7.41)	53(4.33)	50(6.89)	54(5.26)	51(6.83)	49(5.81)	51(8.09)
		0.45	50(2.06)	50(7.26)	50(1.61)	50(7.09)	50(1.66)	50(7.07)	49(5.15)	50(7.67)
	75	0.15	93(18.22)	86(13.48)	93(18.35)	86(13.55)	98(22.40)	86(13.79)	98(22.13)	86(13.83)
		0.25	82(8.09)	77(8.99)	83(8.60)	77(8.80)	93(17.92)	78(8.36)	86(12.40)	78(8.98)
		0.35	76(2.92)	75(8.62)	77(2.92)	75(8.52)	79(5.37)	75(8.12)	77(4.42)	75(8.45)
		0.45	75(2.04)	75(8.77)	75(1.79)	75(8.59)	75(1.44)	75(8.47)	75(2.64)	75(8.48)

续表 3 单样本检验 ($m=100, n=100$) 中 m_0 的估计

(Table 3 Performance of methods under one sample when $m=100, n=100$ (Continued))

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
	25	0.15	72(48.27)	56(33.75)	91(64.48)	75(52.10)	97(69.66)	62(38.62)	52(35.08)	53(31.59)
		0.25	37(14.81)	31(10.04)	68(44.88)	50(28.96)	44(26.84)	34(11.96)	28(7.54)	30(10.67)
		0.35	27(4.18)	26(6.64)	46(24.42)	34(13.71)	27(4.68)	26(6.38)	23(5.23)	26(7.49)
		0.45	25(2.77)	25(6.42)	33(11.23)	27(8.41)	25(2.54)	25(6.30)	22(5.29)	25(7.10)
	50	0.15	84(34.55)	71(24.10)	95(43.24)	83(35.21)	97(46.26)	74(25.90)	82(33.74)	72(24.50)
D		0.25	61(13.30)	54(10.24)	80(31.53)	66(21.71)	74(27.70)	57(11.04)	53(10.55)	54(11.28)
		0.35	52(4.38)	51(8.79)	65(18.06)	56(13.65)	54(6.01)	51(8.13)	49(7.31)	50(9.64)
		0.45	50(2.86)	50(8.67)	56(8.75)	51(11.50)	50(2.20)	50(8.36)	49(6.37)	50(9.10)
	75	0.15	93(18.31)	85(14.22)	98(21.37)	91(17.84)	98(22.39)	86(14.32)	97(20.84)	86(14.60)
		0.25	82(8.93)	77(10.45)	91(16.82)	83(14.68)	93(17.72)	79(9.75)	85(13.49)	78(10.89)
		0.35	76(3.77)	75(10.11)	83(10.76)	78(13.50)	79(6.50)	75(9.55)	77(6.14)	76(10.16)
		0.45	75(2.79)	75(10.27)	78(6.23)	75(12.97)	75(1.97)	75(9.97)	75(4.47)	75(10.23)

注: 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

根据表 2 和表 3 的结果, 可以看出, 在相同的样本量下, 改进方法得到 m_0 的估计的均方误差比已有方法更小. 当样本量逐渐增大, 估计量收敛到真实值 m_0 的速度更快. 估计 $\widehat{m}_0^{\mathrm{DIS}}$ 和 $\widehat{m}_0^{\mathrm{IDIS}}$ 的有效性在独立样本情形下略好一点; 同时, 估计 $\widehat{m}_0^{\mathrm{BH}}$, $\widehat{m}_0^{\mathrm{AC}}$, $\widehat{m}_0^{\mathrm{SD}}$ 及其相应的改进估计量几乎不受到协方差矩阵的影响.

例 2 (两样本检验的情形) 令 $X_i=(x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}, i=1,2,\cdots,n_1$, $Y_k=(y_{1k}, y_{2k}, \cdots, y_{mk})^{\mathrm{T}}$, $k=1,2,\cdots,n_2$. 样本 $X_1, X_2, \cdots, X_{n1}$ 和 $Y_1, Y_2, \cdots, Y_{n2}$ 分别来自多元正态分布总体 $N(\mu_1, \Sigma_1)$ 和 $N(\mu_2, \Sigma_2)$. 设置 $\mu_1=0$, μ_2 与例 1 中的 μ 相同. 为了探究协方差矩阵 Σ_1, Σ_2 是否相等对 m_0 估计的影响, 我们讨论样本的协方差矩阵相等和不等两种情形, 即 $\Sigma_1=I, \Sigma_2=D$ 和 $\Sigma_1=\Sigma_2=D$, 这里 I 和 D 的设置与例 1 中相同. 样本量分别为 $n_1=40, n_2=60$ 和 $n_1=120, n_2=80$. 待检验的原假设为

$$H_{0j}: \mu_{1j}=\mu_{2j}, \quad j=1,2,\cdots,m.$$

两样本的模拟结果在表 4 和表 5 中展示.

表 4 两样本检验 ($m=100, n_1=40, n_2=60$) 中 m_0 的估计

(Table 4 Performance of methods for two sample when $m=100, n_1=40, n_2=60$)

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
情形 1	25	0.15	96(69.88)	86(61.28)	96(70.07)	86(61.38)	98(72.81)	86(61.79)	98(70.32)	86(61.52)
		0.25	84(58.41)	68(43.89)	85(59.69)	68(44.01)	97(71.81)	71(46.80)	71(52.37)	67(42.93)
		0.35	63(39.29)	49(25.84)	68(44.00)	50(26.78)	95(66.92)	57(32.43)	44(22.73)	45(23.24)
		0.45	45(21.66)	36(13.21)	52(28.68)	38(14.50)	62(42.05)	41(17.60)	32(10.07)	34(12.89)

续表 4 两样本检验 ($m=100, n_1=40, n_2=60$) 中 m_0 的估计

(Table 4 Performance of methods for two sample when $m=100, n_1=40, n_2=60$ (Continued)

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
情形 1	50	0.15	97(46.45)	90(40.49)	97(46.58)	90(40.54)	98(48.10)	91(40.72)	98(46.49)	91(40.61)
		0.25	90(40.11)	78(29.94)	91(40.51)	78(29.92)	98(47.22)	79(30.92)	97(42.75)	79(30.63)
		0.35	78(29.22)	66(18.62)	80(30.98)	67(19.11)	97(45.44)	70(21.40)	70(26.58)	67(19.19)
		0.45	67(18.01)	58(11.23)	69(20.57)	58(11.42)	87(36.72)	62(13.71)	56(12.48)	57(12.09)
	75	0.15	98(22.86)	95(19.46)	98(22.91)	95(19.40)	98(23.50)	95(19.50)	98(22.22)	95(19.64)
		0.25	96(20.43)	89(15.89)	96(20.54)	89(15.93)	98(22.78)	89(16.08)	98(22.13)	89(16.23)
		0.35	91(16.13)	83(12.03)	91(16.40)	83(12.21)	98(22.06)	84(12.21)	97(20.98)	84(12.48)
		0.45	85(11.03)	79(9.83)	86(11.48)	79(9.78)	96(20.23)	80(9.55)	90(16.42)	80(9.91)
情形 2	25	0.15	96(69.60)	86(61.30)	96(69.92)	86(61.44)	98(72.86)	86(61.86)	98(68.56)	86(61.07)
		0.25	84(58.10)	67(44.05)	85(60.01)	68(44.64)	97(71.53)	70(47.09)	69(51.34)	67(42.91)
		0.35	63(39.63)	49(26.38)	68(44.37)	50(27.48)	95(65.94)	56(32.42)	44(24.76)	45(24.26)
		0.45	44(22.29)	36(14.32)	53(29.17)	38(15.32)	62(43.74)	41(18.63)	32(11.08)	33(13.44)
	50	0.15	97(46.22)	90(40.09)	98(46.39)	90(40.35)	98(48.19)	91(40.49)	98(45.51)	90(40.11)
		0.25	91(40.03)	78(30.43)	91(40.63)	78(30.55)	98(47.16)	79(31.41)	96(40.45)	79(30.56)
		0.35	79(29.67)	66(19.60)	80(31.00)	67(19.78)	96(44.95)	69(21.83)	69(26.88)	66(19.92)
		0.45	66(18.52)	58(12.38)	69(20.95)	58(12.45)	87(36.68)	62(14.47)	56(14.24)	57(13.25)
	75	0.15	98(22.67)	95(19.33)	99(22.74)	95(19.26)	99(23.57)	95(19.38)	98(21.83)	95(19.29)
		0.25	96(20.31)	89(16.30)	96(22.44)	89(16.16)	98(22.85)	89(16.30)	98(21.42)	89(16.31)
		0.35	91(16.30)	83(13.09)	91(16.59)	83(13.02)	98(21.95)	84(12.83)	97(20.10)	84(13.50)
		0.45	85(11.48)	79(10.95)	85(11.83)	79(11.03)	96(19.98)	80(10.46)	90(16.58)	80(11.25)

注: 情形 1: $\Sigma_1=I, \Sigma_2=D$; 情形 2: $\Sigma_1=\Sigma_2=D$. 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

表 5 两样本检验 ($m=100, n_1=120, n_2=80$) 中 m_0 的估计

(Table 5 Performance of methods for two sample when $m=100, n_1=120, n_2=80$)

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
情形 1	25	0.15	90(64.12)	75(51.27)	90(64.76)	75(51.35)	98(72.19)	76(52.83)	94(61.91)	76(50.74)
		0.25	62(38.73)	48(25.43)	68(43.86)	50(26.67)	95(66.51)	56(31.95)	44(22.76)	45(23.23)
		0.35	39(16.16)	32(10.35)	46(22.99)	34(11.19)	48(29.38)	35(12.56)	29(7.21)	30(10.39)
		0.45	29(5.80)	27(6.40)	34(10.11)	28(5.66)	30(7.79)	27(5.95)	24(4.67)	27(7.46)
	50	0.15	94(43.29)	83(34.60)	94(43.61)	83(34.67)	98(47.59)	84(35.23)	98(44.48)	84(34.92)
		0.25	78(29.11)	66(18.61)	80(30.85)	66(19.04)	97(45.33)	69(21.30)	69(26.38)	66(19.15)
		0.35	62(14.19)	55(9.86)	65(16.77)	56(9.75)	77(29.94)	58(11.18)	54(10.27)	55(10.86)
		0.45	54(5.92)	51(7.92)	56(7.53)	52(7.45)	58(10.39)	52(7.38)	50(6.82)	52(8.99)

续表 5 两样本检验 ($m=100, n_1=120, n_2=80$) 中 m_0 的估计

(Table 5 Performance of methods for two sample when $m=100, n_1=120, n_2=80$ (Continued))

Σ	m_0	a	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
情形 2	75	0.15	97(21.63)	91(17.48)	97(21.80)	91(17.50)	98(23.15)	92(17.62)	98(21.98)	92(17.72)
		0.25	91(16.08)	83(12.28)	91(16.52)	83(12.37)	97(22.01)	84(12.33)	97(20.64)	84(12.64)
		0.35	83(9.23)	78(9.77)	83(9.67)	78(9.57)	94(18.56)	79(8.96)	86(13.91)	78(9.66)
		0.45	78(4.60)	76(9.20)	78(4.76)	76(9.06)	83(10.14)	76(8.45)	79(7.07)	76(9.10)
	25	0.15	90(63.98)	75(51.47)	91(64.74)	75(51.49)	98(72.16)	77(53.12)	92(60.06)	75(50.56)
		0.25	62(39.21)	48(25.93)	68(43.94)	50(27.15)	95(65.75)	56(32.12)	43(23.64)	44(23.57)
		0.35	38(16.44)	32(10.94)	46(23.28)	34(11.60)	48(31.22)	35(13.36)	28(8.13)	31(11.08)
		0.45	29(6.30)	27(6.97)	33(10.20)	28(6.28)	30(8.23)	27(6.55)	24(5.33)	26(8.03)
	50	0.15	94(43.11)	83(34.63)	95(43.58)	83(34.78)	98(47.60)	84(35.23)	97(43.07)	84(34.93)
		0.25	78(29.31)	66(19.49)	80(30.88)	66(19.55)	96(44.91)	69(21.77)	68(26.64)	66(19.79)
		0.35	62(14.60)	55(10.74)	65(16.92)	56(10.62)	77(30.47)	59(11.89)	54(11.21)	55(11.71)
		0.45	54(6.49)	51(8.84)	56(7.79)	52(8.31)	58(11.07)	52(8.25)	50(7.70)	51(9.99)
	75	0.15	97(21.59)	91(17.56)	98(21.70)	91(17.54)	98(23.21)	92(17.73)	98(21.49)	92(17.81)
		0.25	91(16.35)	83(13.02)	91(16.55)	83(12.88)	97(21.93)	84(12.81)	96(20.05)	84(13.42)
		0.35	83(9.56)	78(10.57)	83(9.84)	78(10.47)	94(18.46)	79(9.76)	86(14.30)	78(10.95)
		0.45	78(5.03)	76(10.27)	78(5.09)	75(10.15)	83(10.62)	76(9.46)	79(8.05)	76(10.20)

注: 情形 1: $\Sigma_1=I, \Sigma_2=D$; 情形 2: $\Sigma_1=\Sigma_2=D$. 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

根据表 4 和表 5, 我们可以发现, 改进后的估计比相应已有方法的估计有效性更好. 当样本量逐渐增大时, 估计量收敛到真实值 m_0 的速度更快. 两样本检验中, 所有的估计量对协方差矩阵都不是很敏感. 当 a 比较小时, 改进后的估计更为精确. 随着备择假设逐渐远离原假设, 所有估计的精确性和有效性都有所提高.

例 3 (非正态检验的情形) 数据 $X_1, X_2, \cdots, X_n$ 的分量 $X_i=(x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}$, $i=1,2,\cdots,n$ 相互独立. 我们考虑两种类型的非正态数据. 一组假设检验问题是

$$H_{0j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \sim F(20,20), \quad j=1,2,\cdots,m,$$

$$H_{1j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \sim F(20,20)-0.1, \quad j=1,2,\cdots,m.$$

另一组假设检验问题是

$$H_{0j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \sim \chi^2(1), \quad j=1,2,\cdots,m,$$

$$H_{1j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \sim \chi^2(1)-0.3, \quad j=1,2,\cdots,m.$$

对上述两组假设检验问题, 原假设和备择假设之间的差异相对不大. 令 $m=100$, $m_0=25,50,75$, 样本量 $n=50,100$. 模拟结果展示在表 6 中.

表 6 非正态数据检验 ($m=100$) 中 m_0 的估计

(Table 6 Performance of methods for non-normal data when $m=100$)

H_0	n	m_0	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
$F(20,20)$	50	25	76(50.62)	64(39.61)	77(52.64)	64(39.93)	92(66.46)	67(42.64)	65(43.71)	63(38.55)
		50	85(34.58)	75(26.87)	85(35.31)	75(26.96)	96(44.84)	77(28.30)	90(38.47)	76(27.41)
		75	92(17.20)	87(14.38)	92(17.34)	87(14.34)	97(21.86)	87(14.48)	98(21.76)	87(14.39)
	100	25	58(34.03)	47(22.90)	63(38.96)	48(23.88)	78(53.34)	51(27.22)	42(19.71)	43(21.02)
		50	75(25.10)	64(16.56)	76(26.85)	65(16.81)	90(39.70)	67(18.71)	69(21.62)	64(16.85)
		75	88(13.62)	82(11.05)	89(13.93)	82(10.98)	96(20.72)	83(11.06)	95(19.19)	82(11.22)
$\chi^2(1)$	50	25	48(24.20)	39(16.12)	70(45.12)	57(33.62)	80(54.98)	60(35.57)	56(34.03)	56(32.37)
		50	63(14.11)	54(9.24)	79(29.64)	71(22.55)	88(38.10)	72(23.43)	80(29.72)	71(22.79)
		75	76(5.75)	69(10.38)	88(13.56)	84(12.34)	94(19.05)	84(12.28)	93(17.89)	84(12.54)
	100	25	28(7.33)	23(5.91)	56(31.72)	42(18.34)	62(38.48)	44(19.97)	37(14.11)	38(16.36)
		50	48(6.55)	42(10.86)	71(21.47)	60(13.05)	80(30.38)	63(14.42)	63(15.89)	60(13.50)
		75	67(10.08)	59(17.62)	85(10.66)	79(9.64)	92(17.17)	80(9.59)	89(14.91)	80(9.90)

注: 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

根据表 6, 可以得出当样本量 n 和 m_0 相同时, 改进后的估计比已有方法更精确, 均方误差更小. 对于非正态数据, 即使原假设和备择假设之间的差异不大, 改进后的方法仍然可以更准确地估计 m_0. 随着样本量的增大, 估计的精确性和有效性有所提高, 均方误差减小.

例 4 (非正态检验的情形) 本例讨论原假设和备择假设来自不同分布的非正态数据情形. 原假设和备择假设分别来自如下分布

$$H_{0j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \overset{\text{i.i.d}}{\sim} \chi^2(1), \quad j=1,2,\cdots,m,$$

$$H_{1j}: \{x_{j1}, x_{j2}, \cdots, x_{jn}\} \overset{\text{i.i.d}}{\sim} F(20,20), \quad j=1,2,\cdots,m.$$

令 $m=100$, $m_0=25,50,75$, 样本量 $n=50,100$. 模拟结果展示在表 7 中. 模拟结果显示, 改进后的估计比已有方法的估计更精确, 均方误差更小. 随着样本量的增大, 估计量的精确性和有效性有所提高, 均方误差减小. 因此, 当原假设和备择假设来自不同分布时, 改进方法仍然更有效.

表 7 非正态数据检验 ($m=100$) 中 m_0 的估计

(Table 7 Performance of methods for non-normal data when $m=100$)

n	m_0	$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
50	25	74(50.10)	56(31.96)	78(53.65)	57(32.86)	97(71.82)	62(37.62)	50(28.84)	51(28.00)
	50	85(35.13)	70(21.63)	86(36.29)	70(21.78)	98(47.09)	73(23.88)	79(33.35)	70(21.90)
	75	92(17.20)	83(12.00)	92(17.41)	83(11.99)	98(22.33)	84(12.17)	98(22.07)	84(12.37)
100	25	45(22.41)	36(13.32)	54(30.52)	39(15.00)	91(61.01)	47(22.46)	32(9.68)	34(12.51)
	50	67(18.79)	58(10.99)	71(21.55)	58(11.12)	96(44.08)	63(14.80)	56(11.54)	57(11.71)
	75	85(11.06)	78(9.29)	86(11.49)	78(9.18)	97(21.61)	80(9.01)	91(17.13)	79(9.45)

注: 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

例 5 (数据维数远大于样本量的情形) 令 $X_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}$, $i = 1, 2, \cdots, n_1$, $Y_k = (y_{1k}, y_{2k}, \cdots, y_{mk})^{\mathrm{T}}$, $k = 1, 2, \cdots, n_2$. 数据 $X_1, X_2, \cdots, X_{n1}$ 和 $Y_1, Y_2, \cdots, Y_{n2}$ 分别来自多元正态分布总体 $N(\mu_1, \Sigma)$ 和 $N(\mu_2, \Sigma)$. 其中 $\mu_1 = 0$, μ_2 的前 m_0 项为 0, 其余 $m - m_0$ 项为 1. 基于例 2 的模拟结果, 由于估计对协方差矩阵不太敏感, 我们设置 $\Sigma = I$, 即数据之间相互独立. 令 $m = 5000$, $m_0 = 3500$, 样本量为 $n_1 = 30, n_2 = 15$. 待检验的原假设为

$$H_{0j} : \mu_{1j} = \mu_{2j}, \quad j = 1, 2, \cdots, m.$$

模拟结果展示在表 8 中. 模拟结果显示, 在维数大于样本量的情况下, 改进后的估计比已有方法的估计更靠近真实值, 均方误差更小.

表 8 高维小样本下 m_0 的估计 ($m = 5000, m_0 = 3500, n_1 = 30, n_2 = 15$)

(Table 8 Performance of methods for high-dimensional data when $m = 5000, m_0 = 3500, n_1 = 30, n_2 = 15$)

$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
3858(365.23)	3508(54.56)	3903(414.56)	3529(59.51)	4234(744.81)	3563(78.63)	3812(314.45)	3522(60.73)

注: 小括号内数据为估计的均方误差 (RMSE), 小括号外数据为估计的中位数.

3.2 实证研究

我们应用实际数据分析来说明改进方法的优越性. 数据集在网址 http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi 下载得到. 该数据集包含 38 个急性白血病的基因表达谱样本, 其中 27 个样本为急性淋巴细胞白血病 (ALL: Acute lymphoblastic leukemia), 11 个样本为急性髓细胞白血病 (AML: Acute myelogenous leukemia). 每个样本包含 5000 个基因变量. 我们对数据集应用多重假设检验, 确定对两种急性白血病分类无影响的基因数量. 这是一个两样本检验问题, 待检验的原假设为

$$H_{0j} : \mu_{1j} = \mu_{2j}, \quad j = 1, 2, \cdots, 5000.$$

样本量分别为: $n_1 = 27, n_2 = 11$. 假如第 j 个原假设为真, 就意味着第 j 个基因对急性白血病的分类没有影响. 因此我们采用已有方法和改进方法估计无影响的基因数量 m_0. 表 9 展示了估计结果. 通过表 9 可以发现, 采用改进方法得到的估计基本相同, 都在 3600 附近. 估计值 $\widehat{m}_0^{\mathrm{SD}}$ 是最小的, $\widehat{m}_0^{\mathrm{AC}}$ 是最大的, $\widehat{m}_0^{\mathrm{IAC}}$ 在改进的估计中是最大的. 除了 $\widehat{m}_0^{\mathrm{SD}}$ 以外, 所有的已有方法估计都比相应的改进估计大.

表 9 实际数据分析中 m_0 的估计

(Table 9 Estimators of real data analysis by different methods)

$\widehat{m}_0^{\mathrm{BH}}$	$\widehat{m}_0^{\mathrm{IBH}}$	$\widehat{m}_0^{\mathrm{DIS}}$	$\widehat{m}_0^{\mathrm{IDIS}}$	$\widehat{m}_0^{\mathrm{AC}}$	$\widehat{m}_0^{\mathrm{IAC}}$	$\widehat{m}_0^{\mathrm{SD}}$	$\widehat{m}_0^{\mathrm{ISD}}$
4304	3594	4304	3594	4984	3717	2934	3594

4 结 论

本文基于图方法, 对多重假设检验中真实原假设个数估计进行了改进. 与已有方法相比, 改进方法充分利用了原假设下得到 p 值的信息. 并且改进方法不仅能应用于正态数据,

还可以应用于非正态数据, 如 F 分布数据, χ^2 分布数据等. 此外, 即使原假设和备择假设的差异较小, 改进的方法仍然可以更精确地估计出真实原假设的个数. 改进估计的效果采用中位数和均方误差衡量. 数值分析结果证明了提出方法的有效性.

参 考 文 献

[1] Storey J D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2002, **64**(3): 479–498.

[2] Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 1990, **9**(7): 811–818.

[3] Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 2000, **25**(1): 60–83.

[4] Langaas M, Lindqvist B H, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA micro array data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, **67**(4): 555–572.

[5] Turkheimer F E, Smith C B, Schmidt K. Estimation of the number of "true" null hypotheses in multivariate analysis of neuro imaging data. *Neuro Image*, 2001, **13**(5): 920–930.

[6] Miller C J, Genovese C, Nichol R C, et al. Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 2001, **122**(6): 3492–3505.

[7] Pounds S, Morris S W. Estimating the occurrence of false positives and false negatives in micro array studies by approximating and partitioning the empirical distribution of p-values. *Bio Informatics*, 2003, **19**(10): 1236–1242.

[8] Guan Z, Wu B, Zhao H. Model-based approach to FDR estimation. *Research Report*, 2004, **2004-016**.

[9] Markitsis A, Lai Y. A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bio Informatics*, 2010, **26**(5): 640–646.

[10] Nettleton D, Hwang J G, Caldo R A, et al. Estimating the number of true null hypotheses from a histogram of p- values. *Journal of Agricultural, Biological, and Environmental Statistics*, 2006, **11**(3): 337–356.

[11] Efron B, Tibshirani R, Storey J D, et al. Empirical Bayes analysis of a micro array experiment. *Journal of the American Statistical Association*, 2001, **96**(456): 1151–1160.

[12] Storey J D, Tibshirani R. Statistical significance for genome wide studies. *Proceedings of the National Academy of Sciences*, 2003, **100**(16): 9440–9445.

[13] Storey J D, Taylor J E, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2004, **66**(1): 187–205.

[14] Jiang H, Doerge R W. Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Informatics*, 2008, **6**(1): 25–32.

[15] Schweder T, Spjϕtvoll E. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 1982, **69**(3): 493–502.

[16] Hwang Y, Kuo H, Wang C, et al. Estimating the number of true null hypotheses in multiple hypothesis testing. *Statistics and Computing*, 2014, **24**(3): 399–416.

[17] Qiao Y, Yu W, Xu W. Two new estimators for the proportion of true null hypotheses in multiple test. *Journal of Statistical Computation and Simulation*, 2017, **87**(4): 712–723.