文章编号：1674-8085(2016)03-0035-06

多重相关检验中错误发现率的控制算法

刘遵雄，*陈 昊

（华东交通大学信息工程学院，江西，南昌 330013）

摘 要：假设检验问题是通过比较 p 值和置信水平 α 的值来决定是否拒绝对应的假设，$p<\alpha$ 时我们拒绝原假设。随着试验次数的增加，在所有满足零假设为真的 p 值集合中，数值较小的 p 值存在的可能性会增加，从而使得做出错误的判断，p 值调整算法可以针对多重检验有效地缓解这类问题。本文讨论多重检验的 p 值调整算法的功效，模拟基因序列进行多重检验分析，产生 2000 个模拟量（即基因数量），重复实验 1000 次，得到 1000 组 p 值。对每组 p 值使用相应的调整算法得到新的 p 值，比较每个算法功效的优劣。模拟结果显示在我们所选用的 5 种 p 值调整算法中，q 值方法(Storey 2003)能很好地控制错误发现率(FDR)的大小，同时具有更高的功效值。

关键词：基因序列；零假设；备选假设；错误发现率；p 值调整；检验功效

中图分类号：TP391.9 **文献标识码：**A DOI:10.3969/j.issn.1674-8085.2016.03.008

RESEARCH ON ALGORITHM OF FALSE DISCOVERY RATE CONTROL IN MULTIPLE DEPENDENT TESTS

LIU Zun-xiong, *CHEN Hao

（School of Information Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, China）

Abstract: Hypothesis testing problem determines whether to reject the corresponding hypothesis through comparing p values with confidence level. A smaller p-value than the α -level of the test signifies a statistically significant test. As the number of tests increases, the chance of observing some small p-values is very high even when all null hypotheses are true. Consequently, we make wrong conclusions on the hypotheses. Adjustment of p-values can effectively alleviate this problem. A simulation study with several methods was carried out in multiple hypothesis testing. Simulations generate 2000 analogue. The experiment was repeated 1000 times, 1000 sets of p value are obtained. Use the adjustment methods to access the new p-value, and then compare the advantages and disadvantages of each algorithm. The correlation between genes expression is in the covariance matrix. Simulation results show that among five algorithms selected, q value method can be efficient to control the false discovery rate size, also with higher power.

Key words: gene sequence; null hypothesis; alternative hypothesis; power of test; FDR; adjustment of p-values; power of test

0 引言

多重检验是指成千上万的统计检验同时进行，首先将多个单重检验作为一个整体，然后对这个整体的假设同时进行检验的问题。在众多的统计应用领域，尤其是在生物信息学方面，许多研究要求同时进行数目很大的假设检验。在生物信息学中，如何对微阵列数据的差异表达进行检验，

收稿日期：2016-02-29；修改日期：2016-04-14
基金项目：国家自然科学基金项目(71361009)
作者简介：刘遵雄(1967-)，男，江西瑞昌人，教授，博士，主要从事机器学习、数据挖掘研究(E-mail:Darrent.liu@gmail.com);
*陈 昊(1991-)，男，江苏兴化人，硕士生，主要从事数据挖掘、金融分析研究(E-mail:chenhao806@126.com).

就是一个多重假设检验问题。大量高维数据的出现使得我们面临同时进行多个假设检验的挑战。当面临同时进行成百上千甚至是上万个假设检验时，作为分析工具的多重检验变得越来越重要。

在多重假设检验中最重要的问题是要同时控制多个错误率并保持足够的检验功效。传统的方法是控制总 I 型错误率(FWER)，然而采用这个错误测度的严重后果就是整个多重检验的功效很低。Benjamini 和 Hochberg 提出了一种新的测度，即假发现错误率(FDR)，也就是多重检验中被错误拒绝的检验比例。如今，FDR 已经作为处理多重检验问题的一个非常流行的方法，并且已经凸显出不断增加的重要性和应用价值。

与单个假设检验的思想类似，对于多重检验，需要同时检验一簇假设时，首先需要考虑的问题是如何提出一种合理的错误测度来衡量总体所犯的第 I 类错误，进而寻找一种检验法则将该错误控制在某个合理的范围内，并使得检验的功效尽可能大，即尽可能多地发现显著性假设。

基于上述问题，统计学家提出新的标准以达到控制错误率的目的，其中比较流行的有 FWER(Family-wise Error Rate) 与 FDR(False Discovery Rate)[1]。本文主要通过对得到的 p 值进行调整，用来控制参数 FDR 取值，从而提升检验的功效。

1 多重假设检验基本原理

(1) 假设检验概述

假设检验是统计推断的一个主要方面。单个假设检验的理论已经较为完善。其基本思想是在控制检验所犯第一类错误的基础上，寻找一种检验法则，使得检验所犯的第二类错误尽可能小[2]。

同是考虑 m 个假设检验 H_i，i=1,2,…,m，如果零假设 H_0 为真，那么记 H_i=0，否则 H_i=1。记 M_0={i: H_i=0}，M_1={i: H_i=1}，其中 m_0，m_1 分别为 M_0,M_1 中元素的个数，即 m_0=#M_0,m_1=#M_1。显然可知 m=m_0+m_1[3]。对于这 m 个检验结果的分类如表 1 所示：

表 1 m 次假设检验结果

Table 1 Results of m hypothesis

假设检验	接受 H_0	拒绝 H_0	合计
H_0 为真	U	V	m_0
H_0 非真	T	S	m-m_0
	m-R	R	m

(2) 检验的 p 值

假设检验统计量为 T，其观测样本值为 t，对于一簇拒绝域{Γ}，检验统计量 T=t 的 p 值定义为：

$$p(x) = \min_{\{\Gamma: t\in\Gamma\}} \{\Pr(T \in \Gamma \mid H_0)\}$$

统计学根据显著性方法得到的 p 值，一般以 $p < 0.05$ 为显著，$p < 0.01$ 为非常显著，其含义是样本间的差异由抽样误差所致的概率小于 0.05 或 0.01。

(3) FWER **和** FDR

多重假设检验中，广泛使用的错误控制指标是总体错误率(Family-Wise Error Rate, FWER)，即至少出现一次错误地拒绝真实 H_0 的可能性，且 $\mathrm{FWER} \le \alpha$。而研究者更关心的是能否尽量多地识别出差异表达的基因，并且能够容忍和允许总的拒绝中发生少量的错误识别，称为错误发现(false discovery)。即需要在错误发现和总的拒绝次数 R 之间寻找一种平衡，即在检验中尽可能多的候选变量的同时将错误发现率控制在一个可以接受的范围。

FDR(False Discovery Rate)，是统计学中常见的一个名词，翻译为错误发现率，其意义为是错误拒绝（拒绝真的（原）假设）的个数占所有被拒绝的原假设个数的比例的期望值。FDR 具有以下优点：(1)可以灵活调整其取值，作为假设检验错误率的控制指标，其控制值可以根据需要灵活选取，而传统的假设检验(FWER)的取值则较为固定，通常定为 0.05；(2)FDR 的意义明确。可以作为筛选出的差异变量的评价指标，而 FWER 则主要是用来控制 I 类错误的[4]。

按照表 1 所示，FWER 定义如下：

$$\mathrm{FWER} = \Pr(V \ge 0)$$

我们定义 FDP 为一类错误个数与拒绝原假设的检验个数的比值，即：

$$FDP：\ Q=\frac{V}{R}$$

FDR 是 FDP 的期望值，表示在所有 R 次拒绝中错误发现的期望比例。即：

$$FDR=E(Q)=E(\frac{V}{R})$$

2 错误控制率方法的参数估计

(1) Bonferroni **算法**

采用单步法，选取截断α/m，给定水平α，如果第 i 个检验的 p 值$p\le\alpha/m$，则拒绝 H_{i0}，i=1,2,⋯,m。对于第 i 个检验而言，其对应的错误拒绝数为 V_i (V_i=1 或 0)，则：

$\Pr(V_i=1)\le\frac{\alpha}{m},i=1,2,\cdots,m$（每个检验犯第一类错误的概率不大于$\alpha/m$）

从而我们有

$FWER=\Pr(V_1+V_2+\cdots+V_m\ge1)=$

$\Pr(V_1=1\cup V_2=1\cup\cdots\cup V_m=1)\le$

$\Pr(V_1=1)+\cdots+\Pr(V_m=1)\le\alpha$

所以可以以α严格控制 FWER。

(2) Benjamini and Hochberg **算法流程**

在置信水平α下控制 FDR 的算法步骤：

第一步：将原有的 p 值进行排序：$p(1)\leqslant p(2)\leqslant\cdots\leqslant p(m)$；

第二步：将 BH 方法表示为

$$\hat{k}=\max\{j:p(j)\le\frac{j}{m}\alpha\}$$

第三步：如果这样的$\hat{k}$存在，那么拒绝最小的$\hat{k}$个 p 值对应的假设；如果找不到这样的$\hat{k}$，那么不拒绝任何假设[5]。

BH 方法控制$\mathrm{FDR}\le\frac{m_0}{m}\alpha\le\alpha$，其中$\alpha$为提前给定的检验水平。

杨柳[6]在其论文中第四章节详述了 Storey 的方法过程，降低了 BH 方法的保守性，方便与我们的理解。

(3) Adaptive Benjamini and Hochberg 算法流程

控制 BH 法的 FDR 水平为$m_0\alpha/m$，如果m_0已知，则可令$\alpha'=\alpha m/m_0$取代 BH 法中的检验水平α，进而更精确地控制 FDR 在水平α内从而具有更大的检验功效(即相同条件下拒绝更多原假设)。实际上由于m_0未知，需要在检验之前先给出它的估计值。出于这样一种考虑，Benjamini and Hoehberg(2000)提出调整后的 BH 法，记为 ABH 法。

第一步：在α水平进行 BH 法检验, 如果不存在满足条件的 k，则不拒绝任何零假设，停止；否则继续。

第二步：计算斜率值$m_0(i)=\frac{m+1-i}{1-p(i)}$，i=1,2,⋯,$m$;

第三步：令$J=\min\{j:m_o(j)>m_o(j-1)\}$，计算$m_0$的估计值

$$\hat{m}_0=\min\{m_0(J),m\}$$

第四步：再次做 BH 法，$k=\max\{i:p(i)\le\frac{i}{\hat{m}_0}\alpha\}$，拒绝 $H(1),\cdots,H(k)$对应的零假设[7]。

(4) Benjamini and Yekutieli 算法流程

根据 BH 算法，修改 FDR 的上界值为$m_o\alpha/(m\sum_{i=1}^{m}\frac{1}{i})$。

第一步：将原有的 p 值进行排序：

p(1)≤p(2)≤⋯≤p(m)；

第二步：将 BY 方法表示为

$$\hat{k}=\max\{j:p(j)\le m_0\alpha/m\sum_{i=1}^{m}\frac{1}{i}\}$$

第三步：如果这样的$\hat{k}$存在，那么拒绝最小的$\hat{k}$个 p 值对应的假设；如果找不到这样的$\hat{k}$，那么不拒绝任何假设[8]。

BY 算法得出的上界值过于保守，相对于 BH 算法而言，上界值减小，j 的值也减小，拒绝原假设的个数减少。在表 1 中表示为假设 R 值一定时，V 值减小，S 值增大。

(5) q **值方法**

定义 pFDR 的表达式如下：

$$pFDR=E(V/R\mid R>0)$$

即

$$pFDR=\frac{\pi_0 P(T\in\Gamma|H_i=0)}{P(T\in\Gamma)}=P(H_i=0|T\in\Gamma)$$

其中，

$$P(T\in\Gamma)=\pi_0 P(T\in\Gamma|H_i=0)+\pi_1 P(T\in\Gamma|H_i=1)$$

Storey [9]给 q 值定义如下：

$$q(t)=\min_{\{\Gamma:t\in\Gamma\}} pFDR(\Gamma)=\min_{\{\Gamma:t\in\Gamma\}} P(H_i=0|T\in\Gamma)$$

也就是说 q 值是相对于 pFDR 参数而言的一种衡量数据量强度的一种方法：具体定义为拒绝检验量的最小的 pFDR 值的大小。

另一方面，p 值的定义如下：

$$p(t)=\min_{\{\Gamma:t\in\Gamma\}} P(T\in\Gamma|H_0=0)$$

从以上的表达式可以看出，q 值是在 pFDR 的基础上计算错误率，而 p 值是在第一类错误的基础上计算错误率，因此，q 值和 p 值在某种程度上是相似的。

3 模拟实验

首先我们定义假设检验如下：

$$H_i^0:\mu=0 \quad VS \quad H_i^1:\mu=\mathrm{A}\text{，}$$

i=1,2,…, 1000，A={0.5,1.5,3,4.7}

选取2000大小的统计量Z(对应于微阵列数据中的基因数)，样本服从下列分布：$Z\sim N(\mu,\Sigma)$。使用蒙特卡罗的方法调整不同情况下p值的大小,以此来判断多重检验参数估计方法的性能，很显然调整后的p值的数量也等于2000。基因之间的方差-协方差矩阵结构如下：

$$\sum_m=\begin{pmatrix}\sum_0 & 0\\ 0 & \sum_1\end{pmatrix}\begin{matrix}H0\\ H1\end{matrix}$$

其中，$\sum_0$ 为m_o*m_o的矩阵，对角线元素为1，非对角线元素为0；$\sum_1$ 为m_1*m_1的矩阵，对角线元素为1，非对角线元素为r。结构如下：

$$\sum_0=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&\ldots&0\\0&0&0&1\end{pmatrix}m_0*m_0$$

$$\sum_1=\begin{pmatrix}1&r&r&r\\r&1&r&r\\r&r&\ldots&r\\r&r&r&1\end{pmatrix}m_1*m_1$$

$\sum_1$ 中 r 的取值为{0,0.20,0.50,0.99}，代表选取的 4 个任意非负相关系数，使得基因之间服从正回归依赖关系[10]，用来表示基因之间的相关度大小。可以得到$\sum_{2000*2000}$ 的表达式：

$$\sum=\begin{bmatrix}\begin{pmatrix}1&r\\r&1\end{pmatrix}&&&0\\&\begin{pmatrix}1&r\\r&1\end{pmatrix}&&\\&&\begin{pmatrix}1&r\\r&1\end{pmatrix}&\\0&&&\ddots\\&&&&1\end{bmatrix}$$

所有的 p 值都保存在协方差矩阵Σ之中，该协方差矩阵包含两个基本信息：(1)代表基因数量的相关矩阵数目的值；(2)独立检验量之间的相关值 r。协方差矩阵Σ对角线上的元素都为 1，非对角线上的元素大小代表统计量之间的相关度。因此试验中存在两组包含不同数目相关统计量的样本。第一类样本包含 120 个具有相关性的统计样本，而第二类样本中设置包含 360 个具有相关性的统计样本。

μ_{1*2000} 向量的一般形式如下：

$$\mu=[AA\cdots A000\cdots A]$$

在每个 A 值确定的条件下，重复 2000 次检验量,一共进行1000次模拟仿真。因此对于每个μ_{1*2000}和$\Sigma_{2000*2000}$来说得到 1000 组大小为 2000 的 p 值的集合。然后使用第三节中的 p 值调整算法对原始 p 值进行处理。最后通过计算每个算法的功效值的大小来确定几种错误控制率参数方法的优劣。

4 实验结果分析

以上几种参数估计方法 Bonferroni，BH，BY，ABH 以及 q 值方法得到的检验功效结果图如下表所示，横坐标代表检验的功效大小，纵坐标代表所

使用的参数估计方法，每张图中由 4 个独立的单元拼接而成，代表不同的 *r* 值对检验功效的影响。

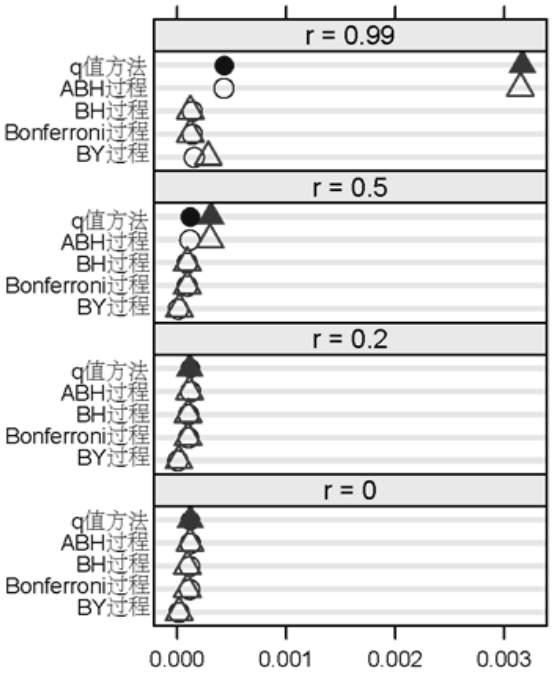

图 1 A=0.5 下 120 次相关检验(圆形)/360 次相关检验(三角)功效值比较

Fig.1 Average power of A=0.5 [120 dependent tests (Circle)/360 dependent tests(Triangle)]

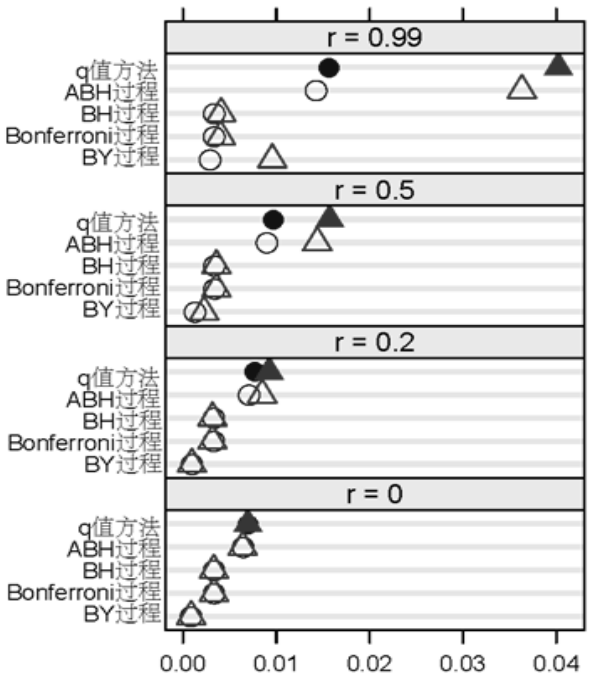

图 2 A=1.5 下 120 次相关检验(圆形)/360 次相关检验(三角)功效值比较

Fig.2 Average power of A=1.5 [120 dependent tests (Circle)/360 dependent tests(Triangle)]

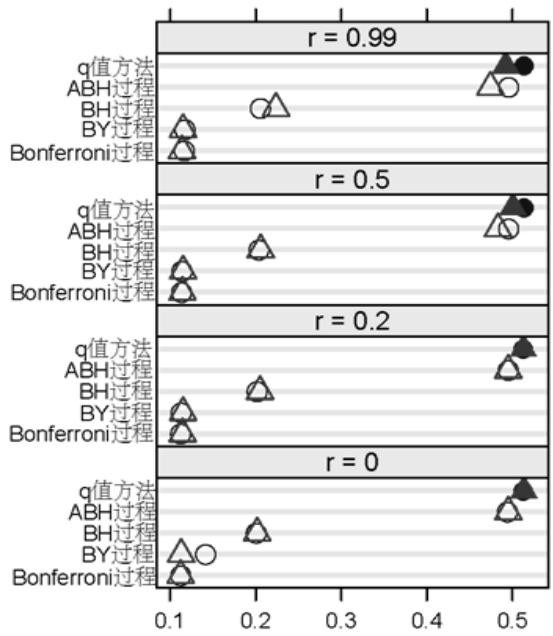

图 3 A=3 下 120 次相关检验(圆形)/360 次相关检验(三角)功效值比较

Fig.3 Average power of A=3 [120 dependent tests (Circle)/360 dependent tests(Triangle)]

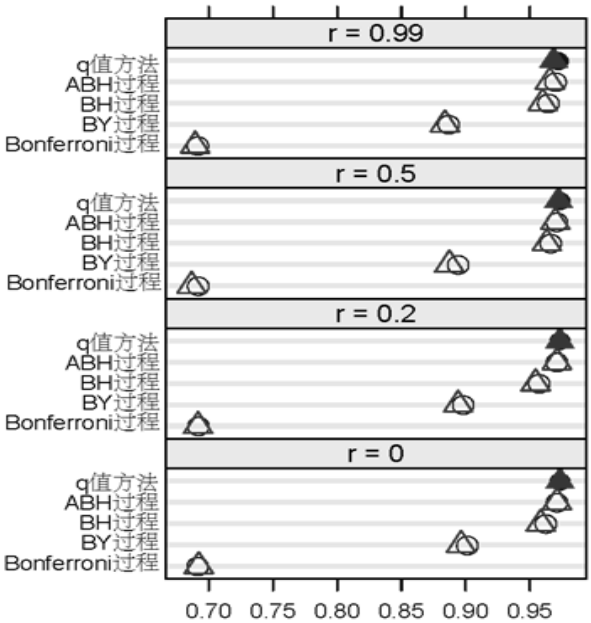

图 4 A=4.7 下 120 次相关检验(圆形)/360 次相关检验(三角)功效值比较

Fig.4 Average power of A=4.7 [120 dependent tests (Circle)/360 dependent tests(Triangle)]

A={0.5,1.5,3,4.7}，以上 4 张图分别对应 A 的四个值所对应的两组相关检验下使用 5 种不同的 p 值调整算法得到的算法平均功效的大小。从每张图中可以看出，在相同 A 值的情况下，r 值越大，检验的功效值也越大。而当把 4 张图进行纵向比较时，增大 A 值的大小，算法功效的提升也是一个有效的方法。

5 结语和存在的问题

当 A<2 时，可以看到所有的算法都没有很好的功效值(见图 1 与图 2)。从图 1 与图 2 中最上端的两个板块(即 $r = 0.99$)可以看出，通过增大 A 值，BH，ABH 以及 q 值方法的功效大小能有一定程度的提升。图 4 清楚地显示当 A 取值为 4.7 时所有的算法检验功效都增大了。同时在 A = 4.7 时，只有 ABH，BH 以及 q 值方法的功效值达到了 0.95 以上，相对而言 q 值方法的检验能力更加突出一点。然而控制 FWER 的 Bonferroni 算法的功效值只是在 0.7 左右徘徊。对比 BY 方法在 A = 3 以及 A = 4.7 两种情况下的功效值，BY 算法的功效大小有了大幅的增长。

基因微阵列研究无法始终在基因序列遵循正回归依赖的条件进行。如果基因呈负相关，那么 Yekutieli(2008)提出另一个方法将很有效。然而该方法将再次给一个保守的 FDR 值。为了消除保守的错误率，可以使用 gate keeping technique 方法(Dmitrienko 和 Tamhane,2007；Dmitrienko, 2008)，也可以考虑 p 值加权的方法(Genovese)，这些方法都是下一步努力的方向。本算法存在的关键问题是不管在在基因存在正相关还是负相关的关系下，仍然不能确定那个方法将提供一个更高的功效值。

参考文献：

[1] 吴小霞. 多重检验中FDR方法及其参数估计问题的研究[D]. 武汉:武汉大学,2010.

[2] 胡建庭,徐沥泉. 统计假设检验的基本思想方法—澄清假设检验中几个容易混淆的概念[J].江苏教育学院学报:自然科学版,2007, 24(3): 52-55.

[3] Conover W J, 崔恒建. 实用非参数统计[M]. 3版.北京:人民邮电出版社,2006.

[4] 李伟. 多重检验相关研究及其在生物数据上的应用[D]. 济南:山东大学,2014.

[5] Jiang A X, Lu P C. Multiple Testing Corrections and FDR Adjustments for Microarray Data Analysis[R]. Vanderbilt Cancer Biostatistics Division Workshop, 2010.

[6] 杨柳. 多重假设检验中错误率控制过程的分析[D]. 武汉:武汉大学, 2010.

[7] Benjamini Y, Krieger A M, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate[J]. Biometrika, 2006, 93(3): 491-507.

[8] Ruth Heller. False Discovery Rate Control in multiple testing problem[R]. Tel-Aviv university, 2013.

[9] Storey J D. The positive false discovery rate: a Bayesian interpretation and the q-value[J]. Annals of statistics, 2003: 2013-2035.

[10] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency[J]. Annals of statistics, 2001: 1165-1188.