基于偏最小二乘分析的 FDR 估计研究*

张 帆[1] 刘 晋[2] 侯 艳[1] 李 康[1△]

【提 要】 目的 基于偏最小二乘模型(PLS)提出一种新的 *FDR* 估计方法,并对其准确性进行验证。方法 利用偏最小二乘的 *vip* 评分筛选变量,结合 *permutation* 方法和后退法对筛选结果进行 *FDR* 估计。结果 模拟实验表明,在变量之间独立时,*PLS-FDR* 方法和三种单变量估计方法都能准确估计 *FDR*;在变量之间存在线性关系时,*PLS-FDR* 方法估计 *FDR* 仍然具有无偏性,而三种单变量分析方法则无法准确地进行估计。实例分析表明,*PLS-FDR* 方法对高维数据分析能够提供重要信息。结论 在线性数据结构下,使用本文给出的 *PLS-FDR* 方法能够得出多变量 *FDR* 估计结果。

【关键词】 偏最小二乘 阳性错误发现率 代谢组学

The Research of False Discovery Rate Estimation of Statistical Analysis Based on Partial Least Squares

Zhang Fan, Liu Jin, Hou Yan, et al. (*Department of Health Statistics, School of Public Heath, Harbin Medical University* (150081), *Harbin*)

【**Abstract**】 **Objective** To provide a new FDR estimation method based on Partial Least Squares(PLS) and to validate its correction as well. **Methods** We estimated the FDR of feature selection results based on the *vip* scores obtained by the Partial Least Squares with the permutation and Step-back technique. **Results** Simulation experiment proved that the PLS-FDR method and three univariate FDR estimation methods have exact estimation results under the independent structure data. But PLS-FDR method had higher accuracy than three univariate FDR estimation methods in dealing with data having liner relationships. Case study proved that PLS-FDR method can provide important information for high dimensional data analysis. **Conclusion** PLS-FDR method can estimate the multivariate FDR accurately in the data having liner relationships.

【**Key words**】 Partial least squares; FDR; Metabonomics

阳性错误发现率(*FDR*)的概念由 Benjamini 和 Hochberg 提出,这一概念的提出,有效地解决了高维组学数据多重比较中假阳性错误的控制问题,并且能够显著提高假设检验的效能[1]。目前,*FDR* 的估计方法很多,其中最具代表性的有 *LBE*[2]、*qvalue*[3] 和 *fdrtool*[4] 等,这些方法都是在贝叶斯公式的框架下进行的,使用两成分模型构建 p 值的分布函数,进而求得 *FDR* 的估计值。然而,这些方法主要基于单变量分析方法,要求变量间独立或弱相关,如果变量高度相关,将会导致上述方法失效。实际中,高维组学数据结构复杂,噪声变量多且变量间存在复杂相关,无法满足上述 *FDR* 估计方法的应用条件;而且单变量分析无法发现变量间的联合作用和交互作用,不能满足研究需要。本文拟提出一种基于偏最小二乘(PLS)多变量模型变量筛选结果的 *FDR* 估计方法(PLS-FDR 法)[5],通过模拟实验探讨 *PLS-FDR* 法的优势,并通过实例分析说明其在实际研究中的意义。

* 基金资助:本研究获高等学校博士学科专项基金(20122307110004);国家自然科学基金资助(81172767)

1. 哈尔滨医科大学卫生统计学教研室(150081)
2. 南京医科大学生物统计学教研室(211166)
△通信作者:李康,E-mail:likang@ems.hrbmu.edu.cn

FDR 控制与估计的基本方法

1. *FDR* 的定义

对于 m 次多重假设检验,表 1 中列出了四种不同检验结果的频数。

表 1 多重假设检验四种结果的频数

真实情况	不拒绝 H_0	拒绝 H_0	合计
H_0 为真	U	V	m_0
H_1 为真	T	S	m_1
合计	W	R	m

FDR 的定义如下:

$$FDR=\begin{cases}E(V/R), & R\neq 0\\ 0, & R=0\end{cases} \quad (1)$$

其中 $E(\cdot)$ 为数学期望。*FDR* 的含义为在规定的检验水准下被判定为阳性的结果中假阳性结果的比例。

2. *FDR* 控制方法

控制是指给定一个显著性水平的界值,从而使 *FDR* 被限制在某一固定水平,对此可以采用线性向上的控制方法,分两步进行:首先将所有检验 p 的值进行排序,即 $p_{(1)}\leqslant p_{(2)}\leqslant p_{(3)}\leqslant\cdots\leqslant P_{(m)}$;然后逐步后退比较 $p_{(i)}\leqslant\frac{i}{m}q(i=m,m-1,m-2,\cdots,1)$,取第一个满足条件的 $p_{(k)}(k\geqslant 1)$,理论上可以证明在此情况下可以

将 FDR 控制在 $q(0 \leqslant q \leqslant 1)$ 水平下[4]。

3. *FDR* 估计方法

FDR 估计指在设定检验拒绝域下，判定为阳性的结果中假阳性结果所占比例的估计值。如果使用假设检验计算出的 p 值进行 *FDR* 估计，其计算公式为:

$$FDR = Pr\{\text{null} \mid p_i\} = \frac{p_0 F_0(p_{(i)})}{F(p_{(i)})} \quad (2)$$

$$F(p) = p_0 F_0(p) + p_1 F_1(p) \quad (3)$$

其中，p_0 为真实无效假设所占总检验次数的比例，$F_0(p)$ 为无效假设下 p 的右侧分布函数; p_1 为实际有差异变量在所有变量中所占的比例，$F_1(p)$ 为备择假设成立下 p 值右侧的分布函数[7]。

偏最小二乘模型 *FDR* 估计原理

偏最小二乘(PLS)是一种将主成分分析、典型相关分析和回归分析结合在一起的方法，可以在建模的同时通过各变量的重要性评分进行变量筛选。算法的基本思想是，以 *PLS* 变量重要性评分值(vip)作为统计量计算 *FDR*，通过估计 $F_0(vip)$ 和 $F(vip)$ 计算 *FDR* 的估计值。本研究利用经验分布对 $F_0(vip)$ 和 $F(vip)$ 进行估计，对于 $F_0(vip)$ 的估计，通过多次打乱数据的分类标签的方法，充分利用样本经验信息估计无效假设下右侧累积概率分布 $F_0(vip)$。$F(vip)$ 的估计，可以直接利用样本数据的经验分布进行估计。由于 PLS 模型各变量评分 vip 不独立、差异变量之间互相影响，为此在估计 $F(vip)$ 时采用逐步后退的方式，在检验水准 α 上，根据 $F_0(vip)$ 的分布进行检验，记录一定数量的差异显著变量的 vip 评分。为保持变量数目不变，需要将这些变量的数值随机置换。上述过程不断循环，直至进行到第 s 步，当 $F^{(s)}(vip) \to F_0(vip)$ 时，停止继续循环。若记每一步选择的差异变量数目为 t，则最后差异变量的个数为 $t \times s$，无效假设变量在所有变量中所占的比例估计值为

$$p_0 = \frac{m - t \times s}{m} \quad (4)$$

最后计算 *FDR*:

$$FDR = Pr\{\text{null} \mid vip_{(i)}\} = \frac{p_0 F_0(vip_{(i)})}{F(vip_{(i)})} \quad (5)$$

其中，m 为数据中变量总数，对上述记录的 vip 评分排序得 $vip_{(1)} \leqslant vip_{(2)} \leqslant vip_{(3)} \leqslant \cdots \leqslant vip_{(i)} \leqslant vip_{(i+1)} \cdots \leqslant vip_{(m)}$，$F_0(vip_{(i)})$ 为无差异变量假设下的右侧分布概率，即

$$F_0(vip_{(i)}) = p_0(vip \geqslant vip_{(i)}) \quad (6)$$

$F(vip_{(i)})$ 为具有差异变量情况下的右侧分布概率，即

$$F(vip_{(i)}) = p(vip \geqslant vip_{(i)}) \quad (7)$$

上述估计 *FDR* 过程称为 *PLS-FDR* 方法。

模拟实验

1. 实验目的

考核在高维数据中 *PLS-FDR* 方法估计 *FDR* 的准确性，并与目前已有的 *LBE*、*fdrtool*、*qvalue* 单变量估计方法进行比较。

2. 实验条件设置

设“疾病组”和“正常组”两组数据样本含量分别为 50 例，组间差异变量 20 个，“疾病组”的差异变量为 $X_i \sim N(1.5, 1)\ (i = 1, 2, \cdots, 20)$，“正常组”的差异变量为 $X_i \sim N(0, 1)\ (i = 1, 2, \cdots, 20)$，同时设定 2000 个噪声变量为 $X_i \sim N(0, 1)\ (i = 1, 2, \cdots, 2000)$。实验分为三种情况: ①差异变量间独立，非差异变量间独立; ②差异变量间独立，非差异变量分为 100 组，每组 20 个变量的相关系数均等于 0.8; ③差异变量的相关系数均等于 0.3，非差异变量分为 100 组，每组 20 个变量的相关系数均等于 0.8。

3. 实验结果

三种实验设置条件下四种方法对 p_0 的估计，真实的 p_0 为 0.990099，由此可见 *PLS-FDR* 法在三种实验条件下估计得非常准确。*LBE*、*qvalue* 和 *fdrtool* 在变量独立情况下比较准确，但当变量间存在相关时，*LBE* 和 *qvalue* 则完全失效，*fdrtool* 虽没有完全失效，但仍没 *PLS-FDR* 法估计准确。三种实验设置条件下四种方法对 *FDR* 的估计结果见图 1 ~ 图 3，结果显示，在差异变量和噪声变量均独立的数据结构下(图 1)，*PLS-FDR* 估计方法与经典单变量 *FDR* 估计方法得到的结果均较为准确; 在另外两种相关数据结构情况下(图 2 ~ 图 3)，经典单变量 *FDR* 估计方法的 *FDR* 估计值与 *FDR* 的真实值有较大的偏差，而 *PLS-FDR* 法的估计值依然准确。

上述过程通过 R 语言编程实现。

实例分析

数据来源: 收集经冠状动脉造影诊断的 43 例动脉粥样硬化患者和 49 例社区人群组的血液样本，使用超高效液相色谱-质谱联用仪分别在正离子和负离子模式下检测其代谢组成分。对检测后得到的血液代谢组指纹图谱数据利用 R 软件包(XCMS，CAMERA)进行数据预处理，正离子模式下得到 1936 个变量，负离子模式下得到 1515 个变量。现利用 *PLS-FDR* 算法估计其中可能具有意义的生物标志物数目。

(1) 对 p_0 的估计: 正离子模式数据 $p_0 \approx 0.8254$，说明 1936 个变量中约有 338 个生物标志物; 负离子模式数据 $p_0 \approx 0.8455$，说明 1515 个变量中约有 234 个生物标志物。

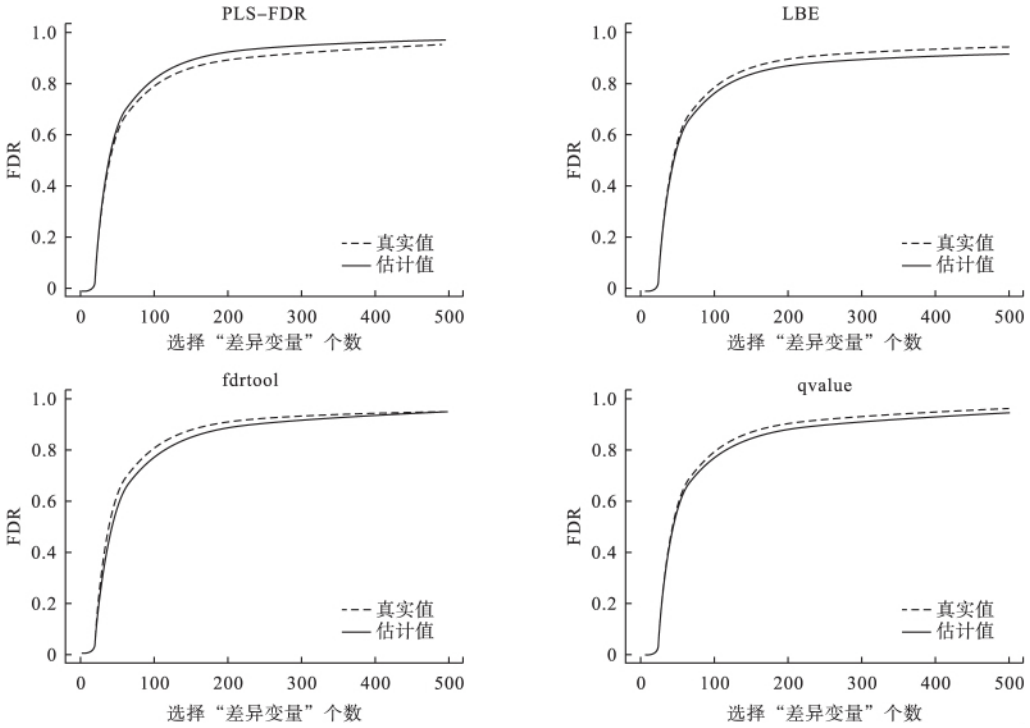

图 1　差异变量与噪声均独立条件下估计 *FDR* 与真实 *FDR* 变化趋势图

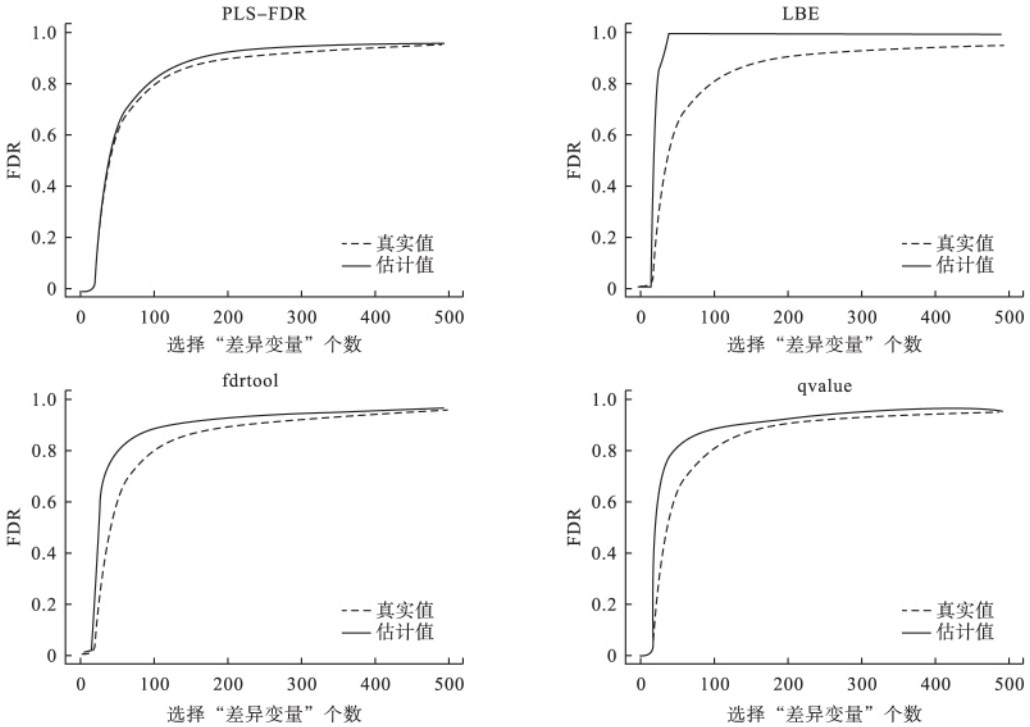

图 2　差异变量独立且噪声相关条件下估计 *FDR* 与真实 *FDR* 变化趋势图

(2) *FDR* 的估计: 结果如图 4 所示，对于正离子模式数据，如果我们选取 *vip* 值排序靠前的 300 个变量作为“差异变量”，其 *FDR* 值约为 0.02，说明其中可能有 294 个生物标志物; 对于负离子模式数据，如果我们选取 *vip* 值排序靠前的 200 个变量作为“差异变量”，其 *FDR* 值约为 0.06，说明其中可能有 188 个生物标志物。

讨　论

1. 三种单变量 *FDR* 估计方法在变量独立的条件下估计值是无偏的，但在变量存在强相关的条件下，其结果与真实值偏离较大，已不具有实用性。本文提出的多变量 *FDR* 估计方法(PLS-FDR) 可以解决单变量分析中出现的问题。

2. 模拟实验结果表明，使用本文提出的 PLS-FDR 方法，在变量独立和相关两种情况下，都能够准确地估计非差异变量占总变量的比例 p_0，同时估计出的 *FDR* 值具有无偏性。

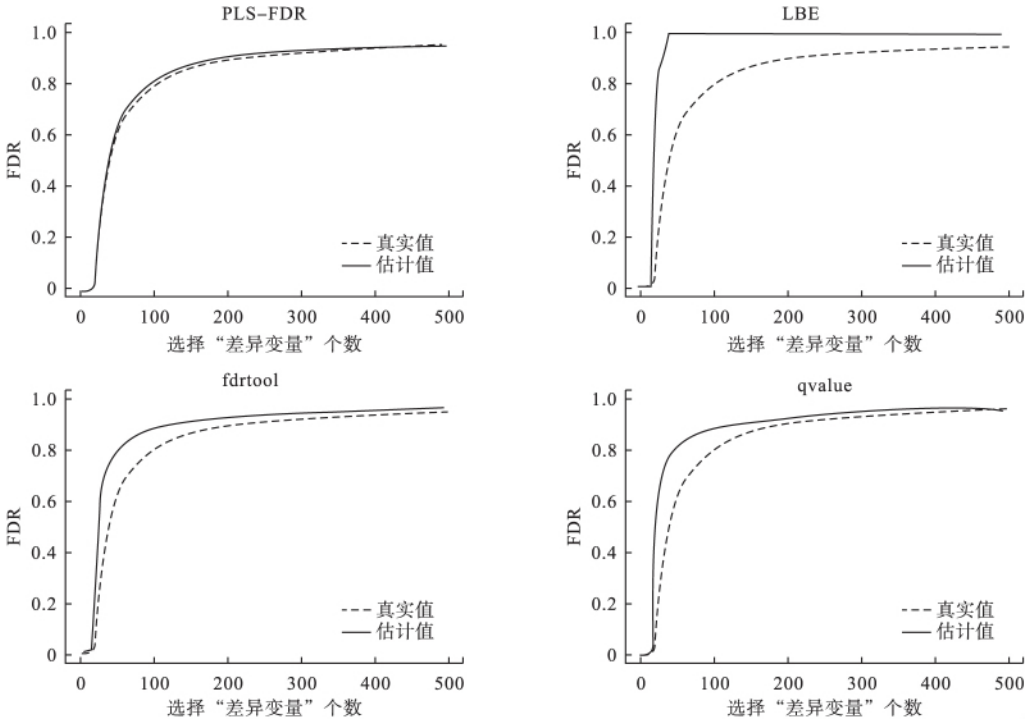

图 3　差异变量与噪声均相关条件下估计 *FDR* 与真实 *FDR* 变化趋势图

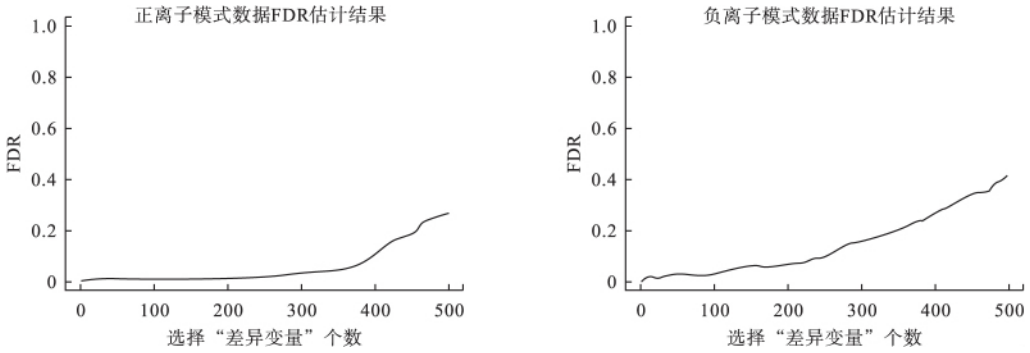

图 4　使用 PLS-FDR 方法对动脉粥样硬化实际数据的 *FDR* 估计结果

3. 偏最小二乘模型主要针对的是线性关系的数据，因此当数据结构中存在大量的非线性关系时，会使估计结果存在一定的偏差，趋于保守。尽管如此，PLS-FDR 算法得到的 *FDR* 估计结果仍然具有一定的实际意义。

4. 本研究使用 PLS-FDR 算法对动脉粥样硬化实际数据进行了分析，分析结果表明，其中含有大量的潜在生物标志物。由于 PLS-FDR 方法使用了多个主成分进行回归，因此可以揭示多变量的联合作用，同时也能够在一定程度上对交互作用的变量进行筛选。

5. 对于多变量分析，PLS 算法中变量的重要性评分 *vip* 是一个相对的量，各变量之间互相影响，因此在 PLS-FDR 算法中使用了后退法，即把有显著作用的变量逐步地进行数据置换，移除其对分类的作用，使其他变量的作用显现出来。本文在每一步中移除的变量数目为 $t=2$，这一参数的最优取值尚需进一步的研究。

参 考 文 献

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B(Methodological), 1995: 289-300.
2. Dalmasso C, Bröet P, Moreau T. A simple procedure for estimating the false discovery rate. Bioinformatics, 2005, 21(5): 660-668.
3. Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat, 2003, 31(6): 2013-2035.
4. Korbinian S. A unified approach to false discovery rate estimation. BMC bioinformati cs, 2008, 9: 303.
5. Boulesteix AL. PLS dimension reduction for classification with high dimensional microarray data . Statistical Applications in Genetics and Molecular Biology, 2004, 3: article 33.
6. 刘晋，张涛，李康. 多重假设检验中 FDR 的控制与估计方法. 中国卫生统计, 2012, 29(2): 305-308.
7. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2002, 64(3): 479-498.

（责任编辑：郭海强）