Positive False Discovery Rate

Jinxi Liu

November 1, 2017

Introduction

Whereas a sequential p-value method fixes the error rate and estimates its corresponding rejection region, Storey proposed the opposite approach. They fixed the rejection region and then estimate its corresponding error rate.

It seems that the method does not offer control of FDR. Actually, control is offered in the same sense as the BH procedure- this methodology provides a conservative bias in expectation.

Definition

FDR:

$$FDR = E\left(\frac{V}{R}|R > 0\right) Pr(R > 0)$$

Positive FDR:

$$FDR = E\left(\frac{V}{R}|R > 0\right)$$

When controlling FDR at level α, and positive findings have occurred, then FDR has really only been controlled at level $\alpha/Pr(R > 0)$.

pFDR is identically 1 when all null hypotheses are true ($m = m_0$).

When m0 = m, one may want the false discovery rate to be 1, and that one is not interested in cases where no test is significant.

Estimation and inference for pFDR and FDR

Theorem

Suppose that m identical hypothesis tests are performed with the independent statistics $T_1, ..., T_m$ and rejection region Γ. Also suppose that a null hypothesis is true with a priori probability $_0$. Then

$$pFDR(\gamma) = \frac{\pi_0 Pr(T \in \Gamma | H = 0)}{Pr(T \in \Gamma)}$$
$$= Pr(H = 0 | T \in \Gamma),$$

Where $Pr(T \in \Gamma) = \pi_0 Pr(T \in \Gamma | H = 0) + \pi_1 Pr(T \in \Gamma | H = 1)$

Estimation and inference for pFDR and FDR

In terms of p-values the above theorem can be written as:

$$pFDR(\gamma) = \frac{\pi_0 Pr(P \leq \gamma | H = 0)}{Pr(P \leq \gamma)} = \frac{\pi_0 \gamma}{Pr(P \leq \gamma)}$$

Where γ refers to the rejection region $[0, \gamma]$.

Estimation and inference for pFDR and FDR

Since $\pi_0 m$ of the p-values are expected to be null, then the largest p-values are most likely to come from the null, uniformly distributed p-values. Hence, a conservative estimate of π_0 is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{(1-\lambda)m} = \frac{W(\lambda)}{(1-\lambda)m}$$

where $p_1, ..., p_m$ are the observed p-values and $W(\lambda) = \#pi > \lambda$. (Optimal λ can be chosen by bootstrap, for now assume that λ is fixed.)

Estimation and inference for pFDR and FDR

A natural estimate of $Pr(P \leq \gamma)$ is

$$\hat{Pr}(P \leq \gamma) = \frac{\#p_i \leq \gamma}{m} = \frac{R(\gamma)}{m}$$

,

Therefore, a good estimate of $pFDR(\gamma)$ for fixed λ is

$$\hat{Q}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{Pr}(P \leq \gamma)} = \frac{W(\lambda)\gamma}{(1-\lambda)R(\gamma)}$$

Estimation and inference for pFDR and FDR

When $R(\gamma) = 0$, the estimate would be undefined. Therefore we replace $R(\gamma)$ with $R(\gamma) \vee 1$. Also, $1 - (1 - \gamma)^m$ is a lower bound for $Pr\{R(\gamma)\} > 0$. Since *pFDR* is conditioned on $R(\gamma) > 0$,we divide by $1 - (1 - \gamma)^m$. In other words $\frac{\gamma}{1-(1-\gamma)^m}$ is a conservative estimate of the type I error, conditional that $R(\gamma) > 0$.

Therefore,we estimate pFDR as

$$\begin{aligned} \hat{pFDR}_\lambda(\gamma) &= \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{Pr}(P \leq \gamma)\{1 - (1 - \gamma)^m\}} \\ &= \frac{W(\lambda)\gamma}{(1 - \lambda)\{R(\gamma) \vee 1\}\{1 - (1 - \gamma)^m\}} \end{aligned}$$

Estimation and inference for pFDR and FDR

Since FDR is not conditioned on at least one rejection occurring, we can set

$$\begin{aligned} \hat{FDR}_\lambda(\gamma) &= \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{Pr}(P \leq \gamma)} \\ &= \frac{W(\lambda)\gamma}{(1-\lambda)\{R(\gamma) \vee 1\}} \end{aligned}$$

Estimation and inference for pFDR and FDR

The expected value of a multiple-hypothesis testing procedure is not a sufficiently broad picture. Since the p-values are independent, we can sample them with replacement to obtain standard bootstrap samples.From these we can form bootstrap versions of our estimate and provide upper confidence limits for pFDR and FDR.

Estimation and inference for pFDR and FDR

Algorithm: estimation and inference for $pFDR(\gamma)$ and $FDR(\gamma)$

1. For the m hypothesis tests, calculate their respective p-values $p_1, ..., p_m$.
2. Estimate π_0 and $Pr(p \leq \gamma)$ by

$$\hat{\pi}_0 = \frac{W(\lambda)}{(1-\lambda)m}$$

and

$$\hat{Pr}(p \leq \gamma) = \frac{R(\gamma) \vee 1}{m},$$

where $R(\lambda) = \#\{p_i \leq \gamma\}$ and $W(\lambda) = \#\{p_i > \gamma\}$.

Estimation and inference for pFDR and FDR

3. For any rejection region of interest $[0, \gamma]$, estimate $pFDR(\gamma)$ by

$$\hat{pFDR}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\hat{Pr}(P \leq \gamma)\{1 - (1-\lambda)^m\}}$$

   for some well-chosen λ.

4. For B bootstrap samples of $p_1, ..., p_m$, calculate the bootstrap estimates $\hat{pFDR}^{*b}_\lambda(\lambda)$ $(b = 1, ..., B)$ similarly to the method above.

5. Form a $1 - \alpha$ upper confidence interval for $pFDR(\gamma)$ by taking the $1 - \alpha$ quantile of the $\hat{pFDR}^{*b}_\lambda(\lambda)$ as the upper confidence bound.

Calculating the optimal λ

An automatic way to estimate

$$\lambda_{best} = \arg \min_{\lambda \in [0,1]} (E[\{p\hat{FDR}_{\lambda}(\gamma) - pFDR(\gamma)\}^2])$$

We use the bootstrap method to estimate λ_{best} and calculate an estimate of $MSE(\lambda)$over a range of . (Call this range R; for example, we may take $R = \{0, 0.05, 0.10, ..., 0.95\}$). We can produce bootstrap versions $p\hat{FDR}_{\lambda}^{*b}(\gamma)$ (for $b = 1, ..., B$) for any fixed λ.

A connection between two procedures

Using the Benjamini and Hochberg (1995) method to control FDR at level $\alpha = \pi_0$ is equivalent to (i.e. rejects the same p-values as) using this method to control FDR at level α.

Let $p_{(1)}, ..., p_{(m)}$ be the ordered, observed p-values for the m hypothesis tests. BH method finds $\hat{k}$ such that

$$\hat{k} = \max\{k : p_{(k)} \leq (k/m)\alpha\}$$

Rejecting $p_{(1)}, ..., p_{(\hat{k})}$ provides $FDR \leq \alpha$.

A connection between two procedures

Now suppose that we use our method and take the most conservative estimate $\hat{\pi}_0 = 1$. Then the estimate $\hat{FDR}(\gamma) \leq \alpha$ if we reject $p_{(1)}, ..., p_{(\hat{l})}$

$$\hat{l} = \max\{l : \hat{FDR}_{(p_{(l)})} \leq \alpha\}.$$

Since

$$\hat{FDR}_{(p_{(l)})} = \frac{\hat{\pi}_0 p_{(l)}}{l/m}$$

this is equivalent to (with $\hat{\pi} = 0$)

$$\hat{l} = \max\{l : p_{(l)} \leq (l/m)\alpha\}$$

Therefore, $\hat{k} = \hat{l}$ when $\hat{\pi} = 0$.

A connection between two procedures

Moreover, if we take the better estimate

$$\hat{\pi}_0(\lambda) = \frac{\#p_i > \lambda}{(1-\lambda)m}$$

then $\hat{l} \geq \hat{k}$, which leads to greater power.

The q-value

The q-value gives the scientist a hypothesis testing error measure for each observed statistic with respect to pFDR.

For an observed statistic $T = t$, the q-value of t is defined to be

$$q(t) = \inf_{\{\Gamma : t \in \Gamma\}} pFDR(\Gamma).$$

For a set of hypothesis tests conducted with independent p-values, the q-value of the observed p-value p is

$$q(p) = \inf_{\gamma \geq p} \{pFDR(\gamma)\} = \inf_{\gamma \geq p} \{\frac{\pi_0 \gamma}{Pr(P \leq \gamma)}\}$$

The q-value is a measure of the strength of an observed statistic with respect to pFDR.

The q-value

Algorithm: calculating the q-value

1. For the m hypothesis tests, calculate the p-values $p_1, ..., p_m$.
2. Let $p_{(1)} \leq ... \leq p_{(m)}$ be the ordered p-values.
3. set $\hat{q}(p_{(m)}) = p\hat{FDR}(p_{(m)})$.
4. Set $\hat{q}(p_{(i)}) = \min\{p\hat{FDR}(p_{(i)}), \hat{q}(p_{(i+1)})\}$ for $i = m-1, m-2, ..., 1$.

Whereas it can be inconvenient to have to fix the rejection region or the error rate beforehand, the q-value requires us to do neither.

A Bayesian interpretation

Suppose we wish to perform m identical tests of a null hypothesis versus an alternative hypothesis based on the statistics $T_1, T_2, ..., T_m$. Let π_0 be the a priori probability that a hypothesis is true: that is, we assume that the H_i are i.i.d. Bernoulli random variables with $Pr(H_i = 0) = \pi_0$ and $Pr(H_i = 1) = 1 - \pi_0 =: \pi_1$.

A Bayesian interpretation

THEOREM

Suppose m identical hypothesis tests are performed with the statistics $T_1, ..., T_m$ and significance region Γ. Assume that (T_i, H_i) are i.i.d random variables,
$T_i|H_i \sim (1 - H - i)F_0 + (1 - H_i)F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_i \sim Bernoulli(\pi_1)$ for $i = 1, ..., m$. Then

$$pFDR(\Gamma) = Pr(H = 0|T \in \Gamma),$$

where $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

Estimation of pFDR and FDR under dependence

We assume we are testing m hypotheses using statistics $T_1, ..., T_m$. We also assume that the null hypothesis is simple, and it is the same for all tests. The alternative hypothesis can be simple or it can be composite in the sense that the alternative is different for each test, but comes from random family of alternatives. The dependence between the T_i can be arbitrary, regardless of whether they follow the null or alternative distributions.

Estimation of pFDR and FDR under dependence

We denote the rejection regions by the set $\{\Gamma\}$. We provide an estimate for both the pFDR and FDR over the fixed rejection region Γ. We make the important assumptions that null versions of the statistics can be simulated; denote these simulated null statistics by $T_1^0, ..., T_m^0$.

Estimation of pFDR and FDR under dependence

Algorithm

1. Let Γ be the rejection region of interest and Γ' be a well chosen rejection region so that its complement is likely to contain mostly null statistics. (An automatic method for choosing Γ' was developed)
2. Simulate the null statistics for B iterations to obtain sets $T_1^{0b}, ..., T_m^{0b}$ for b =11,...,B.
3. Calculate

$$E[R^0(\Gamma)] = \frac{1}{B}\sum_{b=1}^{B} R^{0b}(\Gamma)$$

$$Pr(R^0(\Gamma) > 0) = \frac{1}{B}\sum_{b=1}^{B} 1(R^{0b}(\Gamma) > 0),$$

where $R^{0b}(\Gamma) = \#\{T_i^{0b} \in \Gamma\}$

Estimation of pFDR and FDR under dependence

4. Estimate π_0 by

$$\hat{\pi}_0 = \frac{W(\Gamma')}{E[W^0(\Gamma')]},$$

where $E[W^0(\Gamma')] = m - E[R^0(\Gamma')]$ is calculated similarly to the previous step but with Γ'.

5. Estimate $pFDR(\Gamma)$ by

$$\hat{pFDR}_{\Gamma'}(\Gamma) = \frac{\hat{\pi}_0 E[R^0(\Gamma)]}{Pr(R^0(\Gamma) > 0) \cdot (R(\Gamma) \vee 1)}.$$

6. Estimate $pFDR(\Gamma)$ by

$$\hat{FDR}_{\Gamma'}(\Gamma) = \frac{\hat{\pi}_0 E[R^0(\Gamma)]}{(R(\Gamma) \vee 1)}$$

Estimation of pFDR and FDR under dependence

Ideally, the statistics can be formed so that they are exchangeable in the sense that the $T_i|H_i = 0$ are identically distributed. That way, all the statistics can be used in gathering information about the null distribution, and the same rejection region can be used for each test. If this is not possible, then a p-value can be calculated for each statistics by simulating the null distribution individually.