Lecture 11. False Discovery Rate Control

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. False Discovery Rate

The False Discovery Rate (FDR) is an error control criterion developed in the 90's as an alternative to the FWER. The FWER makes sense when we are testing a small number of hypotheses. For example, in comparing six or ten different treatments, it is very reasonable to control the probability of returning even one ineffective treatment. The consequences of a false discovery may be severe, and even so strong a criterion as FWER control can leave us with high power. The state of data analysis, however, has changed. Now we are testing millions of hypotheses at once, for example in genome-wide association studies, and making a false discovery is not the end of the world. Moreover, FWER is so stringent a method that individual departures from the null in this setting have little chance of being detected; FWER control often returns nothing. While we do not want to waste time on too many null genes, we need a compromise: the burden of proof for FWER is too high. We prefer to return some false positives along with many potentially interesting genes, because this enables scientists to follow these leads and to distinguish the important genes from the false discoveries.

A new point of view advanced by Benjamini & Hochberg (1995) proposes controlling the expected proportion of errors among the rejected hypotheses. Recall that, generally, we have four types of outcomes in multiple testing:

TABLE 1
Formal definition of FDR

	accept	rejected	total
True	U	V	n_0
False	T	S	$n - n_0$
total	$n - R$	R	n

where U, V, S, T are unobserved random variables but R is observed. Using the notation from the above table, we define the following criteria.

Definition 1.1 (♣ **False Discovery Rate (FDR)**). In comparison with the FWER, which is defined as

$$\text{FWER} = \mathbb{P}(V \geq 1).$$

The false discovery rate FDR is defined as the expectation of the false discovery proportion FDP, where

$$\text{FDP} = \frac{V}{R} \cdot \mathbb{1}(R \geq 1), \text{ and } \text{FDR} = \mathbb{E}\big[\text{FDP}\big]$$

Notice that although we observe R, we do not observe V, and so FDP is an unobserved random variable. The criterion False Discovery Rate (FDP) we propose to control is its expectation.

Consider what it means to control FDR: if we repeat our experiment many times, on average we control the FDP. This is not a statement about our individual experiment, and does not say much about the chance of having our FDP exceed a certain threshold —— other than the very weak bound we can obtain via Chebyshev's inequality. FWER, on the other hand, does control for an individualexperiment. That is, with FWER control, we have managed our false discoveries unless we are very unlucky; with FDR control, on average our test will control FDP, but this time we may not have done a very good job. Moreover, although controlling FWER certainly gives FDR control, because

$$\text{FDR} = \mathbb{E}\big[\text{FDP}\big] = \mathbb{E}\Big[\frac{V}{R} \cdot \mathbb{1}(R \geq 1)\Big]$$

$$\leq \mathbb{E}\Big[\frac{V}{R} \cdot \mathbb{1}(V \geq 1)\Big] \leq \mathbb{E}\big[\mathbb{1}(V \geq 1)\big] = \text{FWER} \leq \alpha. \tag{1.1}$$

the converse does not hold. Overall, FDR is a relatively weak notion of control. Alternatives to FDR in the literature include the false exceedence rate that aims at controlling $\mathbb{P}(\text{FDP} \geq q)$ where q is a number between 0 and 1.

2. Properties of the FDR

Here we mention two key properties. The first relates the FDR to the FWER, and the second is a statement about the relative power and level of conservativeness.

Proposition 2.1. *Under the global null, the* FDR *is equivalent to the* FWER.

Proof. To see that this is the case, note that in this setting, all rejections are false rejections under the null, so $V = R$, hence

$$\text{FDR} = \mathbb{E}\big[\text{FDP}\big] = \mathbb{E}\Big[\frac{V}{R} \cdot \mathbb{1}(R \geq 1)\Big]$$

$$= \mathbb{E}\Big[\mathbb{1}(R \geq 1)\Big] = \mathbb{E}\Big[\mathbb{1}(V \geq 1)\Big] = \mathbb{P}(V \geq 1) = \text{FWER}$$

So control of FDR implies weak control of FWER. □

Proposition 2.2. *Controlling* FWER *insures* FDR *control. That is, procedures that control* FDR *can only be less stringent and thus more powerful.*

Proof. According to equation (1.1), we have FWER≥FDR. □

3. FDR Controlling Procedure, Benjamini & Hochberg Procedure

Consider the multiple hypothesis testing problem

$$H_{0i} \;\; v.s. \;\; H_{1i}, \;\; i = 1, \cdots, n$$

for each of the sub-test we have p-values p_i, $1 \leq i \leq n$, and the corresponding ordered p-values

$$p_{(1)} \leq \cdots \leq p_{(n)}$$

with a fixed level $\alpha \in [0, 1]$.

Benjamini-Hochberg Procedure: Let i_0 be the largest i for which $p_{(i)} \leq i\alpha/n$, i.e.,

$$i_0 = \arg\max_i \left\{ \mathbb{1}\left(p_{(i)} \leq \frac{i\alpha}{n} \right) \prod_{j=i+1}^{n} \mathbb{1}\left(p_{(j)} > \frac{j\alpha}{n} \right) \right\}$$

then the BH procedure rejects all $H_{(i)}$ with $i \leq i_0$, where $H_{(i)}$ is the hypothesis corresponding to $p_{(i)}$.

Theorem 3.1. *For independent test statistics, i.e., $p_1, \cdots, p_n$ are mutually independent with each other, the Benjamini-Hochberg procedure stated above controls the* FDR *at level α. More precisely,* FDR$= n_0\alpha/n \leq \alpha$.

Remark 3.2. An interesting note is that the BH(α) threshold is adaptive; it depends on the specific valuesof $p_{(1)}, \cdots, p_{(n)}$. So two different sets of p-values, both tested at level α, could generate quite different BH(α) thresholds (rejecting different sub-hypothesis), depending on their last point of crossing the line $i\alpha/n$.

Proof. As the conclusion is obvious when $n_0 = 0$, so we assume $n_0 \geq 1$. By defining $V_i = \mathbb{1}(H_i$ is rejected$)$ for each null $i \in \mathcal{H}_0$, we can express the FDP as

$$\text{FDP} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R} \cdot \mathbb{1}(R \geq 1) = \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \frac{V_i}{k} \cdot \mathbb{1}(R = k).$$

First, notice that,

(i) when there are exact k rejections, and if H_i is rejected. Then based on BH procedure, we have there exist some $j \leq k$, such that

$$p_i = p_{(j)} \leq p_{(k)} \leq \frac{k\alpha}{n}$$

(ii) when there are exact k rejections, and if we have

$$p_i \leq \frac{k\alpha}{n}$$

then there exist some $j \in \{1, \cdots, n\}$ such that $p_i = p_{(j)}$.

   (a) If $1 \leq j \leq k$, then based on the BH procedure, since we have exact k rejections, so we must have rejected $H_{(1)}, \cdots, H_{(k)}$, then we have rejected $H_i = H_{(j)}$.

   (b) If $j > k$, then

$$p_i = p_{(j)} \leq \frac{k\alpha}{n} < \frac{j\alpha}{n}.$$

   which means the largest i for which $p_{(i)} \leq i\alpha/n$ should at least be j, and must be strictly larger than k, hence we must have rejected more than k hypotheses and that contradicts with our assumption.

combine these two cases, it suggests that under the case, we would reject H_i.

Hence, when there are exactly k rejections, then H_i is rejected if and only if $p_i \leq k\alpha/n$. In other words, conditional on we have exactly k rejections, then

$$V_i = \mathbb{1}(H_i \text{ is rejected}) = \mathbb{1}(p_i \leq k\alpha/n)$$

Second, notice that,

   (i) when there are exact k rejections, and H_i is rejected, if $H_i = H_{(j)}$ for some $1 \leq j < k$. Then defined the new set of p-values $\{\tilde{p}_s\}_{1 \leq s \leq n}$, such that

$$\tilde{p}_s = p_s, \text{ for } s \neq i, \text{ and } \tilde{p}_i = 0.$$

correspondingly, we have the order statistics,

$$\begin{cases} \tilde{p}_{(1)} = \tilde{p}_i = 0, \\ \tilde{p}_{(s)} = p_{(s-1)}, \text{ for } 2 \leq s \leq j, \\ \tilde{p}_{(s)} = p_{(s)}, \text{ for } s \geq j+1, \end{cases}$$

If we denote $R(p_i \to 0)$ as the number of rejections we gonna make for the multiple hypothesis testing problem based on $\{\tilde{p}_s\}_{1 \leq s \leq n}$ instead of $\{p_s\}_{1 \leq s \leq n}$, then appearently, since $j < k$, so we have

$$\tilde{p}_{(s)} = p_{(s)}, \text{ for } s = k, k+1, \cdots, n$$

and since we have exactly $R = k$ rejections based on the BH procedure, so

$$k = \arg\max_t \left\{ \mathbb{1}\left(p_{(t)} \leq \frac{t\alpha}{n} \right) \prod_{r=t+1}^{n} \mathbb{1}\left(p_{(r)} > \frac{r\alpha}{n} \right) \right\}$$

$$= \arg\max_t \left\{ \mathbb{1}\left(\tilde{p}_{(t)} \leq \frac{t\alpha}{n}\right) \prod_{r=t+1}^{n} \mathbb{1}\left(\tilde{p}_{(r)} > \frac{r\alpha}{n}\right) \right\}$$

meaning that

$$R(p_i \to 0) = R = k.$$

(ii) when there are exact k rejections, and H_i is rejected, if $H_i = H_{(k)}$. Then based on our early argument, we have

$$p_i = p_{(k)} \leq k\alpha/n.$$

Again, if we denote $R(p_i \to 0)$ as the number of rejections we gonna make for the multiple hypothesis testing problem based on $\{\tilde{p}_s\}_{1 \leq s \leq n}$ instead of $\{p_s\}_{1 \leq s \leq n}$, where

$$\tilde{p}_s = p_s, \text{ for } s \neq k, \text{ and } \tilde{p}_k = 0.$$

correspondingly, we have the order statistics,

$$\begin{cases} \tilde{p}_{(1)} = \tilde{p}_k = 0, \\ \tilde{p}_{(s)} = p_{(s-1)}, \text{ for } 2 \leq s \leq k, \\ \tilde{p}_{(s)} = p_{(s)}, \text{ for } s \geq k+1, \end{cases}$$

then appearently, we have

$$\tilde{p}_{(k)} = p_{(k-1)} \leq p_{(k)} \leq k\alpha/n,$$
$$\text{and } \tilde{p}_{(s)} = p_{(s)} > s\alpha/n, \quad \text{for } s = k+1, \cdots, n$$

so

$$\mathbb{1}\left(\tilde{p}_{(k)} \leq \frac{k\alpha}{n}\right) \prod_{s=k+1}^{n} \mathbb{1}\left(\tilde{p}_{(s)} > \frac{r\alpha}{n}\right) = 1$$

which is

$$R(p_i \to 0) = \arg\max_t \left\{ \mathbb{1}\left(\tilde{p}_{(t)} \leq \frac{t\alpha}{n}\right) \prod_{r=t+1}^{n} \mathbb{1}\left(\tilde{p}_{(r)} > \frac{r\alpha}{n}\right) \right\} = k = R$$

Combine two cases, when there are exact k rejections, and H_i is rejected, we have $R = R(p_i \to 0)$. Therefore, we have

$$V_i \cdot \mathbb{1}(R = k) = V_i \cdot \mathbb{1}(R(p_i \to 0) = k) = \mathbb{1}(p_i \leq k\alpha/n) \cdot \mathbb{1}(R(p_i \to 0) = k)$$

Third, if we define

$$\mathcal{F}_i = \sigma(p_1, \cdots, p_{i-1}, p_{i+1}, \cdots, p_n),$$

then

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{E}\left[\frac{V_i}{k} \cdot \mathbb{1}(R = k)\right]$$

$$= \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{E} \left\{ \mathbb{E} \left[\frac{V_i}{k} \cdot \mathbb{1}(R = k) \Big| \mathcal{F}_i \right] \right\}$$

$$= \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{E} \left\{ \mathbb{E} \left[\frac{\mathbb{1}(p_i \le k\alpha/n)}{k} \cdot \mathbb{1}(R(p_i \to 0) = k) \Big| \mathcal{F}_i \right] \right\}$$

$$= \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{E} \left\{ \frac{1}{k} \cdot \mathbb{1}(R(p_i \to 0) = k) \cdot \mathbb{E} \left[\mathbb{1}(p_i \le k\alpha/n) \Big| \mathcal{F}_i \right] \right\}$$

$$= \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{E} \left\{ \frac{1}{k} \cdot \mathbb{1}(R(p_i \to 0) = k) \cdot \mathbb{P}(p_i \le k\alpha/n) \right\}$$

$$= \frac{\alpha}{n} \cdot \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{n} \mathbb{P} \Big(R(p_i \to 0) = k \Big)$$

Since $\tilde{p}_i = 0$, so we have $R(p_i \to 0) \ge 1$ and therefore,

$$\sum_{k=1}^{n} \mathbb{P} \Big(R(p_i \to 0) = k \Big) = \mathbb{P} \Big(R(p_i \to 0) \ge 1 \Big) = 1,$$

which leads to

$$\text{FDR} = \frac{\alpha}{n} \cdot \Big(\sum_{i \in \mathcal{H}_0} 1 \Big) = \frac{n_0 \cdot \alpha}{n}$$

$\square$
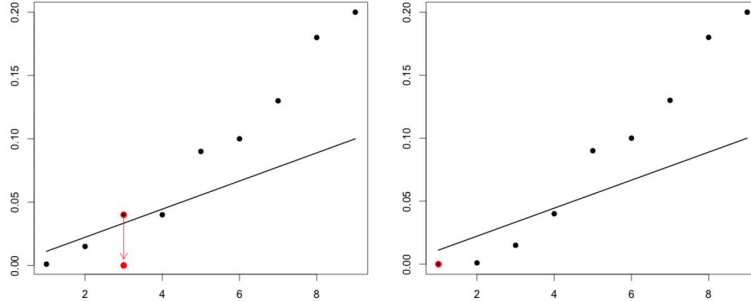

Figure 1: Demonstration of $R(p_i \to 0)$. Here we have plotted ordered p-values and let $i = 3$.

Notice that the assumptions we need here are only the mutual independence of the null p-values $\{p_i : i \in \mathcal{H}_0\}$ among themselves and from the non-nulls. But we do not require the independence between the non-null p-values $\{p_i : i \in \mathcal{H}_1\}$. As a matter of fact, Benjamini-Hochberg's procedure finds the smallest j such that that $p_{(n-j)}$ falls below $(n-j)\alpha/n$. Such definition is a "stopping time" with respect to some filtration.

4. Comparison Benjamini-Hochberg Procedure to other FWER control metrods

4.1. Comparison to Simes' procedure

Under the global null, we have FDR=FWER, in which case, $n_0 = n$ and FDR=FWER=α. Therefore,

$$\alpha = \text{FDR} = \text{FWER} = \mathbb{P}\big(V \geq 1\big) = \mathbb{P}\Big(\min_i\Big\{p_{(i)} \leq \frac{i \cdot \alpha}{n}\Big\}\Big)$$

which is the type-I error for Simes' procedure under the global null.

As we mentioned before, Simes' procedure can't control FWER. As a matter of fact, BH procedure fails to control FWER as well, which turns out to be less stringent than the FWER controling procedure and thus gives us more power at the cost of controlling a different error criterion (FDR).

4.2. Comparison to Hochberg' procedure

Benjamini-Hochberg is a Step-Up procedure. It is interesting to compare it with Hochberg's procedure, which controls the FWER under independence.

Recall that Hochberg's procedure considers the p-values in descending order, and rejects all hypothesis after the first time that

$$p_{(i)} \leq \frac{\alpha}{n - i + 1}.$$

Therefore, the ratio of the two thresholds is

$$\Big(\frac{i\alpha}{n} \Big/ \frac{\alpha}{n - i + 1}\Big) = i\Big(1 - \frac{i - 1}{n}\Big).$$

So roughly speaking, BH(α) is approximately i times more liberal than Hochberg's procedure for small values of i. When i is around $n/2$, the ratio of the thresholds is about $n/4$, which is a very big difference from the viewpoint of power.

5. The Empirical Process Viewpoint of BH(α)

5.1. Empirical Process of the p-values

As seen earlier, the BH(α) is a step-up procedure that relates to Simes' procedure. We plot the sorted p-values and the critical line as shown in Figure.1 and look for the first time (going from large to small p-values) a p-value falls below the critical line. An equivalent representation can be obtained by swapping the x and the y-axes, as in Figure.2. The latter best allows to describe the BH(α) procedure in terms of an empirical process. In particular, the coordinated on
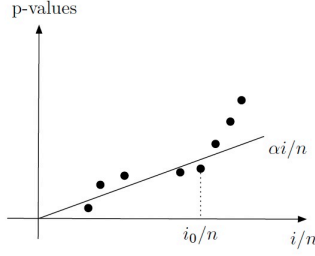
Figure 1: y/x axis: $p-$values/indices
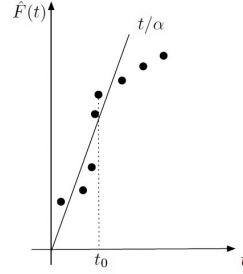
Figure 2: y/x axis: indices/$p-$values

the y-axis of Figure.2 are the values of the empirical CDF $\hat{F}_n(t)$ of the p-values, which is defined as

$$\hat{F}_n(t) = \frac{\#\{i : p_i \le t\}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(p_i \le t).$$

Assuming that the p-values and the hypothesis are ordered accordingly, i.e.,

$$p_{(1)} \le \cdots \le p_{(n)}, \quad H_{(1)} \le \cdots \le H_{(n)}.$$

Recall that we reject $H_{(1)}, \cdots, H_{(i_0)}$ according to BH(α) procedure if

$$i_0 = \arg\max_i \left\{ \mathbb{1}\left(p_{(i)} \le \frac{i\alpha}{n}\right) \prod_{j=i+1}^{n} \mathbb{1}\left(p_{(j)} > \frac{j\alpha}{n}\right) \right\}$$

or in other words,

$$i_0 = \max\left\{ i : p_{(i)} \le \alpha \cdot \frac{i}{n} \right\}$$

thus by the definition of the empirical distribution function, the critical p-value $p* = p_{(i_0)}$ can be rewritten as

$$p* = p_{(i_0)} = \max\left\{ i : p_{(i)} \le \alpha \cdot \frac{i}{n} \right\} = \max\left\{ p_{(i)} : p_{(i)} \le \alpha \hat{F}_n(p_{(i)}) \right\}$$

with the convention that $p* = \alpha/n$ if the above set is empty. Hence,

$$p* = \max\left\{ t \in \{p_1, \cdots, p_n\} : t \le \alpha \hat{F}_n(t) \right\}$$

with the same convention as above if the set is empty. Now set

$$\tau_{\mathrm{BH}} = \max\left\{ t : \frac{t}{\hat{F}_n(t) \vee 1/n} \le \alpha \right\}$$

and notice that $\tau_{\mathrm{BH}} \geq \alpha/n$. By construction, the BH procedure rejects all hypotheses with $p_i \leq \tau_{\mathrm{BH}}$. That is, when k rejections are made, all p-values less or equal to $k\alpha/n$ are rejected.

Again, let's justify the BH(α) procedure using the following simple argument. Take a fixed value of $t \in (0, 1)$ and consider the rule that rejects $H_{0,i}$ if and only if $p_i \leq t$. The entries in the table of outcomes (shown below) will depend on the value of t.

<div align="center">

TABLE 2
Outcome of FDR

	accept	rejected	total
True	$U(t)$	$V(t)$	n_0
False	$T(t)$	$S(t)$	$n - n_0$
total	$n - R(t)$	$R(t)$	n

</div>

The false discovery proportion and false discovery rate, defined as before, will also depend on t, i.e.,

$$\mathrm{FDP}(t) = \frac{V(t)}{R(t)} \cdot \mathbb{1}(R(t) \geq 1).$$

The threshold t should then be chosen as large as possible, while controlling the FDR at level α. Observe that by definition, we have

$$\mathrm{FDR}(t) = \mathbb{E}\big[\mathrm{FDP(t)}\big] = \mathbb{E}\left[\frac{V(t)}{R(t)} \cdot \mathbb{1}(R(t) \geq 1)\right] \leq \alpha. \tag{5.1}$$

Therefore, estimates of FDR process can be inverted to give a FDR-controlling thresholding procedure. If we have an estimate $\widehat{\mathrm{FDR}}(t)$ of $\mathrm{FDR}(t)$, then we can take the threshold

$$\tau = \sup\left\{t \leq 1 : \widehat{\mathrm{FDR}}(t) \leq \alpha\right\}.$$

This defines the largest (most liberal) thresholding cut-off, s.t. our estimated FDR is controlled. Note that, since we choose τ by looking at all the p-values, this is a data-dependent thresholding procedure.

5.2. Estimating FDR

But, how can we estimate the $\mathrm{FDR}(t)$ in (5.1)? Recall that, the number of rejections $R(t)$ is known while the number of false rejections $V(t)$ is unknown.

A solution starts by noting that $\mathbb{E}V(t) = n_0 t$. But still, n_0 is not known. However, one can make a "conservative estimate" and using the upbound nt of $\mathbb{E}V(t)$. This leads to

$$\widehat{\mathrm{FDR}}(t) = \frac{nt}{\max(R(t), 1)} = \frac{t}{\hat{F}_n(t) \vee 1/n}.$$

This choice leads us to exactly the BH(α) procedure since

$$\tau_{\text{BH}} = \sup\left\{ t \leq 1 : \frac{nt}{R(t) \vee 1} \leq \alpha \right\}.$$

As a matter of fact, we further have

Theorem 5.1. *Under independence,* $\mathbb{E}[\widehat{\text{FDR}}(t)] \geq \text{FDR}(t)$.

6. Improving on BH(α)

Consider the intepretation of BH(α) introduced earlier. Can the distributions of p-values be used to improve the simple conservative estimate of $\pi_0 = \frac{n_0}{n}$? Fix $\lambda \in [0, 1)$ and define

$$\hat{\pi}_0^\lambda = \frac{n - R(\lambda)}{(1 - \lambda)n}.$$

Based on this estimate for $\pi_0 = \frac{n_0}{n}$, one obtains the following estimate for FDR:

$$\widehat{\text{FDR}}_0^\lambda = \frac{\hat{\pi}_0^\lambda \cdot nt}{\max(R(t), 1)} \tag{6.1}$$

For $\lambda = 0$, BH(α) is recovered. For a general λ.

$$\hat{\pi}_0^\lambda = \frac{n_0 - V(\lambda) + n_1 - S(\lambda)}{(1 - \lambda)n}$$

$$\geq \frac{n_0 - V(\lambda)}{(1 - \lambda)n}.$$

and hence

$$\mathbb{E}\left[\hat{\pi}_0^\lambda\right] \geq \frac{n_0}{n} = \pi_0.$$

The idea is that if non null p-values are small, then $n_1 - S(\lambda) \approx 0$ and $\hat{\pi}_0^\lambda$ gives an accurate estimate of π_0. The goal would be to show that the threshold

$$\tau = \sup\left\{ t \leq 1 : \widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0^\lambda \cdot nt}{\max(R(t), 1)} \leq q \right\}$$

provides FDR control. While this may actually not be the case, Storey, J., Taylor, J., & Siegmund, D. (2004) proves such a thing for a modified version of (6.1).