Lecture Note: FDR Control

1 Multiple Testing and FDR

This note may contain factual and/or typographic errors. They are partially based on the lecture notes of Emmanuel Candès's course from 2018 to 2023 and paper "Controlling the false discovery rate via knockoffs" (2015), written by R. Foygel Barber and E. Candès, which appeared in the Annals of Statistics. This note is written by Panxu Yuan, see more details on `https://candes.su.domains/teaching/stats300c/lectures.html`.

Outline

- False discovery rate (FDR).
- BH method
- FDR control for linear regression: Knockoff filter

1.1 Multiple testing

Motivation: DNA microarrays measure expression levels of tens of thousands of genes. The data consist of levels of mRNA, which are thought to measure how much of a protein the gene produces. A larger number implies a more active gene.

$$X_{ij} = \text{expression level of gene } i \text{ in patient } j.$$

For example, prostate cancer dataset: a total of 6033 genes were measured on 102 patients (50 controls and 52 with prostate cancer). Here we have $n = 6033$ tests of hypotheses:

$$\begin{aligned} H_{0,i}: &\quad \text{gene } i \text{ is "null"} \\ H_{1,i}: &\quad \text{gene } i \text{ is "non-null"} \end{aligned}$$

and we are interested in testing these hypotheses simultaneously.

In general, we can collect p-values $p_1, \ldots, p_n$ for each hypothsis testing. In the prostate cancer example, the testing statistic is

$$T_i = \frac{\text{Ave[Cancer]} - \text{Ave[Control]}}{\text{Est. Std. Error}}, \quad i = 1, \ldots, 6033.$$

Under $H_{0,i}, T_i \sim t_{100}$ and the corresponding p-values are $p_i = \mathbb{P}(|t_{100}| > |T_i|)$.

We have four types of outcomes in multiple testing:

	accepted	rejected	total
true	U	V	n_0
false	T	S	$n - n_0$
total	$n - R$	R	n

Table 1: Potential outcomes for testing multiple null hypotheses.

where U, V, S, T are unobserved random variables. R is an observed random variable. The two quantities of primary interest to us are the size of V (the number of false discoveries) and the size of R (the total number of rejected hypotheses).

1.2 FWER and FDR

Using the notation from Table 1, we define the following criteria.

▶ **Familywise Error Rate (FWER):** Classical multiple testing procedures aim to control

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

in a strong sense; that is, under all configurations of true and false hypotheses.

♣ **Remark:** *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 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.*

▶ **False Discovery Proportion (FDP):**

$$\text{FDP} = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

If we made no rejections, then our false discovery proportion is 0.

► **False Discovery Rate (FDR):** Notice that although we observe R, we do not observe V, and so FDP is an unobserved random variable. The criterion we propose to control is its expectation, which we refer to as the False Discovery Rate:

$$\text{FDR} = \mathbb{E}[\text{FDP}].$$

♣ **Remark:** The FDR is an error control criterion developed in the 1990s as an alternative to the FWER. 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 Markov's inequality. FWER, on the other hand, does control for an individual experiment. 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, the converse does not hold. Overall, FDR is a relatively weak notion of control. Alternatives to FDR in the literature include the false exceedance rate that aims at controlling $\mathbb{P}(\text{FDP} \geq q)$ where q is a number between 0 and 1.

♣ **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.

- *Under the global null (that is all H_{0i} are true), the FDR is equivalent to the FWER.*

To see that this is the case, note that in this setting, all rejections are false rejections, so $V = R$. Therefore $V/(R \vee 1) = 0$ for $V = 0$, and otherwise $V/(R \vee 1) = V/R = 1$. That is,

$$\text{FDP} = V/(R \vee 1) = \mathbb{1}_{\{V \geq 1\}}.$$

Taking expectations gives

$$\text{FDR} = \mathbb{E}(\text{FDP}) = \mathbb{P}(V \geq 1).$$

this means that FDR control implies weak FWER control.

- *FWER $\geq$ FDR. Why?*

Note that $\mathbb{1}_{\{V \geq 1\}} \geq \text{FDP}$, since $V \leq R$, and taking expectations gives $\mathbb{P}(V \geq 1) \geq \text{FDR}$. Therefore, any procedure that controls the FWER must also control the FDR. That is, procedures that control FDR can only be less stringent and thus more powerful.

Invoking the above properties, this note mainly focuses on controlling FDR. We first present the celebrated Benjamini-Hochberg (BH) method, then introduce the knockoff procedure.

2 BH Method for FDR Control

2.1 Benjamini-Hochberg (BH) procedure

The Benjamini-Hochberg procedure is a step-up procedure with adaptive p-values. The BH procedure at level q (BH(q)) is formally defined as follows.

- Begin by ordering the p-values in ascending order

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

- Fixed a level $q \in (0, 1)$, let i_0 be the the largest i for which

$$p_{(i)} \leq \frac{i}{n} q.$$

- Finally, reject all $H_{(i)}$ with $i \leq i_0$.

The BH threshold is adaptive in the sense the final significant threshold depends on the empirical distribution of p-values of $\{p_1, \ldots, p_n\}$.
The BH procedure means that if we know the number of rejections R, then we can choose the threshold of p-values as $\max\{R, 1\}q/n$.

Consider a simple example. We have collected p-values: 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019. Given FDR level $q = 0.05$, the BH procedure as below.

Rank(k)	Sorted p-values	$\frac{kq}{n}$	Reject?
1	0.005	0.0062	YES
2	0.009	0.0125	YES
3	0.019	0.0187	YES
4	0.022	0.0250	YES
5	0.051	0.0312	NO
6	0.101	0.0375	NO
7	0.361	0.0437	NO
8	0.387	0.0500	NO

Not only does the BH procedure control the FDR under independence, we can even get an explicit expression for the FDR:

Theorem 2.1 *If $\{p_1, \ldots, p_n\}$ are mutually independent, the Benjamini-Hochberg procedure at level q satisfies*

$$\text{FDR} = \frac{n_0}{n} q \leq q$$

where n_0 is the number of true nulls.

Proof. The theorem is trivial when $n_0 = 0$, since both sides are zero, so we assume $n_0 \geq 1$. Let $V_i = \mathbb{1}\{H_i \text{ is rejected}\}$. Then

$$\text{FDP} = \sum_{i \in H_0} \frac{V_i}{\max\{R, 1\}}.$$

Note that $\dfrac{V_i}{\max\{R,1\}}$ are identically distributed for all $i \in H_0$, since the null p-values have the same (uniform) distribution. Thus, it suffices to show that each $\dfrac{V_i}{\max\{R,1\}}$ in this sum has expectation $\dfrac{q}{n}$, based on which we immediately have

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \sum_{i \in H_0} \mathbb{E}\left[\frac{V_i}{\max\{R,1\}}\right] = \sum_{i \in H_0} \frac{q}{n} = \frac{n_0}{n} q \leq q.$$

What remains for the proof is to show that this claim is true. It is helpful to re-write

$$\frac{V_i}{\max\{R,1\}} = \sum_{k=1}^{n} \frac{V_i \mathbb{1}\{R = k\}}{k}$$

by summing over the possible values of the number of rejections $R = 1, \ldots, n$. (Note when $R = 0$, we get $V_i/(\max\{R,1\}) = 0$, and the equation holds). Here, we use two key observations that are based on how the BH procedure is defined:

① When there are k rejections, then H_i is rejected if and only if $p_i \leq qk/n$, and therefore, we have

$$V_i = \mathbb{1}_{\{H_i \text{ rejected}\}} = \mathbb{1}_{\{p_i \leq qk/n\}}.$$

② Suppose $p_i \leq qk/n$ (or H_i is rejected), let us take p_i and set its value to 0 , and denote the new number of rejections by $R(p_i \to 0)$. Then this new number of rejections is exactly R, because we are only reordering the first $k$$p$-values, all of which remain below the threshold qk/n. On the other hand, if $p_i > qk/n$, then we do not reject H_i, and so $V_i = 0$. Therefore, we have

$$V_i \mathbb{1}_{\{R=k\}} = V_i \mathbb{1}_{\{R(p_i \to 0) = k\}}.$$

Define $R(p_i \to 0)$ to be the number of rejections we would get from BH(q) if we changed the i-th p-value to zero, keeping all the rest the same. We claim that

$$\sum_{k=1}^{n} \frac{V_i \mathbb{1}\{R = k\}}{k} = \sum_{k=1}^{n} \frac{V_i \mathbb{1}\{R(p_i \to 0) = k\}}{k}.$$

Now let $\mathcal{F}_i$ be the σ-algebra generated by $\{p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n\}$, i.e. every p-value besides p_i. We have

$$\begin{aligned}
\mathbb{E}\left(\frac{V_i}{\max\{R,1\}} \mid \mathcal{F}_i\right) &= \sum_{k=1}^{n} \mathbb{E}\left(\frac{V_i \mathbb{1}\left\{R\left(p_i \to 0\right) = k\right\}}{k} \mid \mathcal{F}_i\right) \\
&= \sum_{k=1}^{n} \mathbb{E}\left(\frac{\mathbb{1}\left\{p_i \leq \frac{kq}{n}\right\} \mathbb{1}\left\{R\left(p_i \to 0\right) = k\right\}}{k} \mid \mathcal{F}_i\right) \\
&= \sum_{k=1}^{n} \frac{\mathbb{1}\left\{R\left(p_i \to 0\right) = k\right\}}{k} \mathbb{P}\left(p_i \leq \frac{kq}{n}\right) \\
&= \frac{q}{n} \sum_{k=1}^{n} \mathbb{1}\left\{R\left(p_i \to 0\right) = k\right\} \\
&= \frac{q}{n}.
\end{aligned}$$

Here, we use that $R\left(p_i \to 0\right)$ is $\mathcal{F}_i$-measurable to pull it out of the expectation, then use that V_i is independent of $\mathcal{F}_i$ (because the p-values are all independent by assumption) to get that

$$\mathbb{P}\left(p_i \leq \frac{kq}{n} \mid \mathcal{F}_i\right) = \mathbb{P}\left(p_i \leq \frac{kq}{n}\right) = \frac{kq}{n}.$$

Finally, note that $R\left(p_i \to 0\right) \in \{1, \ldots, n\}$ can only take on one particular number $k \in \{1, \ldots, n\}$, so $\mathbb{1}\left\{R\left(p_i \to 0\right) = k\right\} = 1$ for one and only one $k \in \{1, \ldots, n\}$. Thus, the Iterated Expectation Theorem verifies that

$$\text{FDR} = \sum_{i \in H_0} \mathbb{E}\left[\frac{V_i}{\max\{R,1\}}\right] = \sum_{i \in H_0} \mathbb{E}\left[\mathbb{E}\left[\frac{V_i}{\max\{R,1\}} \mid \mathcal{F}_i\right]\right] = \sum_{i \in H_0} \frac{q}{n} = \frac{n_0}{n} q,$$

which completes the proof.

♣ **Remark:** In the above proof, we only used that the null p-values are mutually independent, and that the null p-values are independent of the non-null p-values.

2.2 The empirical process viewpoint of BH

Given $\{p_1, \ldots, p_n\}$, the empirical CDF of p-values is

$$\widehat{F}_n(t) = \frac{\#\{i : p_i \leq t\}}{n}.$$

This allows us to describe the BH procedure in terms of an empirical process.
Denote $p_{(i)}$ as the i th smallest p-value, so that

$$p_{(1)} \leq \ldots \leq p_{(n)},$$

and let $H_{(i)}$ be the corresponding hypothesis for each $p_{(i)}$. BH is defined to reject $H_{(1)}, \ldots, H_{(i_0)}$ where

$$i_0 = \max\left\{i : p_{(i)} \leq \frac{qi}{n}\right\}.$$

The critical p-value is $p^* = p_{(i_0)}$ and can be written as

$$\begin{aligned} p^* &= \max\left\{p_{(i)} : p_{(i)} \leq \frac{qi}{n}\right\} \\ &= \max\left\{p_{(i)} : p_{(i)} \leq q\widehat{F}_n\left(p_{(i)}\right)\right\} \\ &= \max\left\{t \in \{p_1, \ldots, p_n\} : t \leq q\widehat{F}_n(t)\right\}. \end{aligned}$$

If the set is empty, the convention is $p^* = q/n$. Therefore, the BH procedure is equivalent to rejecting all hypotheses with $p_i \leq \tau_{BH}$, where

$$\tau_{BH} = \max\left\{t : \frac{t}{\widehat{F}_n(t) \vee 1/n} \leq q\right\}, \quad \widehat{F}_n(t) \vee 1/n := \max\left\{\widehat{F}_n(t), 1/n\right\}.$$

Equivalently,

$$\tau_{BH} = \max\left\{t : \widehat{F}_n(t) \vee \frac{1}{n} \geq \frac{t}{q}\right\}.$$

Notice that $\tau_{BH} \geq q/n$. This formulation has a simple interpretation. Let $t \in (0, 1)$ be fixed and consider rejecting H_i iff $p_i \leq t$. We can construct the rejection/acceptance table for the hypotheses whose values depend on t.

	H_0 accepted	H_0 rejected	Total
H_0 true	$U(t)$	$V(t)$	n_0
H_0 false	$T(t)$	$S(t)$	$n - n_0 = n_1$
	$n - R(t)$	$R(t)$	n

We define

$$\mathrm{FDP}(t) = \frac{V(t)}{R(t) \vee 1}, \quad \mathrm{FDR}(t) = \mathbb{E}[\mathrm{FDP}(t)].$$

Then, the problem can be restated as follows: choose the largest possible threshold t while controlling the FDR level. But the FDR and FDP are unknown. If we had an estimate $\widehat{\mathrm{FDR}}$ of FDR, we can take the threshold τ to be

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

and define the rejection rule to reject H_i if and only if $p_i \leq \tau$, where τ is a data-dependent threshold.

This method defines the most liberal thresholding cutoff that controls $\widehat{\mathrm{FDR}}(t)$, and we hope that it will also control the true $\mathrm{FDR}(t)$ at level q. The first question is how to estimate $\mathrm{FDR}(t)$.

Assuming $p_i \overset{iid}{\sim} \text{Unif}(0,1)$ under H_0, then $\mathbb{E}[V(t)] = n_0 t$, but n_0 is not known. Therefore, a conservative estimate of $n_0 t$ is nt, which leads to our first estimate

$$\widehat{\text{FDR}}(t) = \frac{nt}{R(t) \vee 1} = \frac{t}{\widehat{F}_n(t) \vee 1/n}$$

This leads us to exactly the BH procedure since

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

The following theorem shows that $\widehat{\text{FDR}}(t)$ is a conservatively biased estimate. This means that the procedure that controls $\widehat{\text{FDR}}(t)$ also controls $\text{FDR}(t)$ at level q.

Theorem 2.2 *Under independence of p-values, this FDR estimate is biased upwards:*

$$\mathbb{E}[\widehat{\text{FDR}}(t)] \geq \text{FDR}(t).$$

Proof. We complete the proof by considering two cases: $S(t) \geq 1$ and $S(t) = 0$. Note that $S(t)$ denotes the number of true discoveries.

- First suppose $S(t) \geq 1$. Then,

$$\begin{aligned}
\mathbb{E}[\widehat{\text{FDR}}(t) - \text{FDR(t)}] &= \mathbb{E}\left[\frac{nt}{S(t)+V(t)} - \frac{V(t)}{S(t)+V(t)} \right] \\
&= \mathbb{E}\left[\frac{nt - (S(t)+V(t)-S(t))}{S(t)+V(t)} \right] \\
&= \mathbb{E}\left[\frac{nt+S(t)}{S(t)+V(t)} \right] - 1 \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{nt+S(t)}{S(t)+V(t)} \mid S(t) \right] \right] - 1.
\end{aligned}$$

  Note that $\dfrac{nt+S(t)}{S(t)+V(t)}$ is convex in $V(t)$, so by Jensen's inequality,

$$\mathbb{E}\left[\mathbb{E}\left[\frac{nt+S(t)}{S(t)+V(t)} \mid S(t) \right] \right] \geq \mathbb{E}\left[\frac{nt+S(t)}{S(t)+\mathbb{E}[V(t) \mid S(t)]} \right].$$

  $V(t)$ and $S(t)$ are independent by assumption, so

$$\mathbb{E}[V(t) \mid S(t)] = \mathbb{E}[V(t)] = n_0 t,$$

  which implies that

$$\mathbb{E}[\widehat{\text{FDR}}(t) - \text{FDR(t)}] \geq \mathbb{E}\left[\frac{nt+S(t)}{S(t)+n_0 t} \right] - 1 \geq 0.$$

- Now, we consider $S(t) = 0$. Then,

$$\mathbb{E}[\widehat{\mathrm{FDR}}(t) - \mathrm{FDR}(t)] = \mathbb{E}\left[\frac{nt - V(t)}{1 \vee V(t)}\right]$$

We know that $V(t) \sim \mathrm{Bin}\,(n_0, t)$. This implies that $\mathbb{E}[1 \vee V(t)] = \mathbb{P}(V(t) = 0) + n_0 t$. Applying Jensen's inequality with this identity yields

$$\begin{aligned}
\mathbb{E}\left[\frac{nt}{1 \vee V(t)}\right] &\geq \mathbb{E}\left[\frac{n_0 t}{1 \vee V(t)}\right] \\
&\geq \frac{n_0 t}{\mathbb{E}[1 \vee V(t)]} \\
&= 1 + \frac{n_0 t - \mathbb{E}[1 \vee V(t)]}{\mathbb{E}[1 \vee V(t)]} \\
&= 1 - \frac{\mathbb{P}(V(t) = 0)}{\mathbb{E}[1 \vee V(t)]} \\
&\geq 1 - \mathbb{P}(V(t) = 0) \\
&= \mathbb{P}(V(t) \geq 1)
\end{aligned}$$

In addition, note that

$$\mathbb{E}\left[\frac{V(t)}{1 \vee V(t)}\right] = \mathbb{P}(V(t) \geq 1).$$

Therefore, we have

$$\mathbb{E}[\widehat{\mathrm{FDR}}(t) - \mathrm{FDR}(t)] = \mathbb{E}\left[\frac{nt - V(t)}{1 \vee V(t)}\right] \geq \mathbb{P}(V(t) \geq 1) - \mathbb{P}(V(t) \geq 1) = 0.$$

The proof is completed.

2.3 Martingale theory for BH method (NOT REQUIRED)

Now, we consider the BH method from the viewpoint of martingale theory. We first present a famous theorem in martingale theory.

♣ Doob's Optional Stopping Time Theorem

Doob's optional stopping time theorem is contained in many basic texts on *Probability and Martingales*. The essential content of the theorem is that you can't make money (in expectation) by buying and selling an asset whose price is a martingale. Precisely, the theorem states that if you buy the asset at some time and adopt any strategy at all for deciding when to sell it, then the expected price at the time you sell is the price you originally paid. Thus — if the market price is a martingale — you cannot make money in expectation by "timing the market".

A discrete-time version of the theorem is given below: Let $X = (X_t)_{t\in\mathbb{N}_0}$ be a discrete-time martingale and τ a stopping time with values in $\mathbb{N}_0 \cup \{\infty\}$, both with respect to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$.
Assume that one of the following three conditions holds:

(a) The stopping time τ is almost surely bounded, i.e., there exists a constant $c \in \mathbb{N}$ such that $\tau \leq c$ a.s.

(b) The stopping time τ has finite expectation and the conditional expectations of the absolute value of the martingale increments are almost surely bounded, more precisely, $\mathbb{E}[\tau] < \infty$ and there exists a constant c such that $\mathbb{E}\left[\left|X_{T_+1} - X_t\right| \mid \mathcal{F}_t\right] \leq c$ almost surely on the event $\{\tau > t\}$ for all $t \in \mathbb{N}_0$.

(c) There exists a constant c such that $|X_{t\wedge\tau}| \leq c$ a.s. for all $t \in \mathbb{N}_0$ where $\wedge$ denotes the minimum operator.

Then X_τ is an almost surely well defined random variable and

$$\mathbb{E}\left[X_\tau\right] = \mathbb{E}\left[X_0\right].$$

Similarly, if the stochastic process $X = (X_t)_{t\in\mathbb{N}_0}$ is a *submartingale* or a *supermartingale* and one of the above conditions holds, then

$$\mathbb{E}\left[X_\tau\right] \geq \mathbb{E}\left[X_0\right],$$

for a submartingale, and

$$\mathbb{E}\left[X_\tau\right] \leq \mathbb{E}\left[X_0\right],$$

for a supermartingale.

Now, we present the results of BH method via Martingale Theory.

Theorem 2.3 *The procedure rejecting all hypotheses with $p_i \leq \tau_{BH}$ controls the FDR:*

$$\mathbb{E}\left[\mathrm{FDR}\left(\tau_{\mathrm{BH}}\right)\right] = qn_0/n, \quad \text{for} \quad \tau_{\mathrm{BH}} = \sup\left\{t \leq 1 : \frac{nt}{R(t)\vee 1} \leq q\right\}.$$

Proof. For simplicity, we let $\tau = \tau_{\mathrm{BH}}$. Define the filtration

$$\mathcal{F}_t = \sigma(V(s), R(s) : t \leq s \leq 1).$$

Notice this is a backwards filtration: for $t_1 < t_2, \mathcal{F}_{t_2} \subset \mathcal{F}_{t_1}$. Define the reverse martingale $\{V(t)/t, 0 \leq t \leq 1\}$. We prove this is indeed a martingale: Let $k \leq t$,

$$\begin{aligned}
\mathbb{E}\left[\frac{V(k)}{k} \mid \mathcal{F}_t\right] &= \frac{1}{k}\mathbb{E}\left[V(k) \mid \mathcal{F}_t\right] \\
&= \frac{1}{k}\cdot\frac{k}{t}V(t) \qquad (*) \\
&= \frac{V(t)}{t},
\end{aligned}$$

where in $(*)$ we used the fact that under $\mathcal{F}_t, V(t) = \#\{p_i : p_i \le t, H_i \text{ null}\}$ and these $p_i \sim U[0,t]$ and are independent. This proves $\{V(t)/t, 0 \le t \le 1\}$ is a martingale.

Next, notice that τ is a stopping time with respect to $\{\mathcal{F}_t\}$. We are ready to apply Doob's Optional Stopping Theorem. By definition,

$$R(\tau) \vee 1 = n\tau/q. \qquad \left(\tau_{\text{BH}} = \sup\left\{t \le 1 : \frac{nt}{R(t) \vee 1} \le q\right\}\right)$$

Therefore,

$$\begin{aligned}
\text{FDR}(\tau) &= \mathbb{E}\left[\frac{V(\tau)}{R(\tau) \vee 1}\right] \\
&= \frac{q}{n}\mathbb{E}\left[\frac{V(\tau)}{\tau}\right] \\
&= \frac{q}{n}\mathbb{E}\left[\frac{V(1)}{1}\right] \\
&= \frac{q}{n} \cdot n_0.
\end{aligned}$$

3 False discovery rate control via Gaussian mirror

We have known that a desirable property of a feature selection procedure is its capability of controlling the number of false positives, which can be mathematically formulated as controlling the false discovery rate (FDR). Previously, we introduced the BH method and knockoff filter. However, in high-dimensional regression, it can be challenging to either construct valid p-values (even asymptotically) or estimate accurately the joint distribution of features, thus, limiting the applicability of both BH and the knockoff filter. Now, we introduce a newly proposed Gaussian Mirror (GM) method, which constructs the test statistic marginally, the relative reference is as below.

- Xing, X., Zhao, Z. G. and Liu, J.-S. (2023). False discovery rate control via data splitting. *Journal of the American Statistical Association*. 118(541): 222–241.

Suppose the data are generated from the following linear regression model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^{n \times p}$ is the design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of iid draws from $N(0, \sigma^2)$. We begin with the low-dimensional case with $p < n$ and focus on the OLS estimator. Without loss of generality, we assume that $\mathbf{X}$ is normalized so that $\sum_{i=1}^n X_{ij} = 0$ and $n^{-1}\|\boldsymbol{X}_j\|_2 = 1$.

We construct the j-th GM by replacing $\boldsymbol{X}_j$ with a pair of variables $\left(\boldsymbol{X}_j^+, \boldsymbol{X}_j^-\right)$with $\boldsymbol{X}_j^+ = \boldsymbol{X}_j + c_j\boldsymbol{z}_j$ and $\boldsymbol{X}_j^- = \boldsymbol{X}_j - c_j\boldsymbol{z}_j$, where $\boldsymbol{z}_j$ is an independently simulated Gaussian random vector with mean 0 and covariance I_n and c_j is a scalar which depends on $\mathbf{X}$ and $\boldsymbol{z}_j$. This construct can be viewed as a quantifiable way of perturbing the feature, and is equivalent to treating $c_j\boldsymbol{z}_j$ as a "knockoff" variable of $\boldsymbol{X}_j$ if OLS is used for the model fitting.

Let $S_0 = \{j : \beta_j = 0\}$ and $S_1 = \{j : \beta_j \neq 0\}$. Note that $\beta_j^+ = \beta_j^- = 0$ when $j \in S_0$ and $\beta_j^+ = \beta_j^- = \beta_j/2$ when $j \in S_1$. Let GM_j be the design matrix with the j-th Gaussian mirror, that is,

$$\mathrm{GM}_j := \mathbf{X}^j = \left(\boldsymbol{X}_j^+, \boldsymbol{X}_j^-, \mathbf{X}_{-j}\right) = \left(\boldsymbol{X}_j + c_j\boldsymbol{z}_j, \boldsymbol{X}_j - c_j\boldsymbol{z}_j, \mathbf{X}_{-j}\right).$$

Fit the model $\boldsymbol{y} = \mathbf{X}^j\boldsymbol{\beta}^j + \boldsymbol{\epsilon}$ and estimate the coefficients by the least square

$$\hat{\boldsymbol{\beta}}^j = \arg\min_{\boldsymbol{\beta}^j = \left(\beta_j^+, \beta_j^-, \boldsymbol{\beta}_{-j}\right)} \left\|\boldsymbol{y} - \mathbf{X}^j\boldsymbol{\beta}^j\right\|_2^2.$$

The mirror statistic for the j-th variable is constructed as

$$M_j = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|.$$

The mirror statistic M_j has two parts: the first part reflects the importance of the j-th predictor; and the second part captures the noise cancelation effect. Intuitively, when $j \in S_0$, the coefficients $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ are due to noise and the mirror statistic M_j should center around zero; when $j \in S_1$, $\hat{\beta}_j^+ + \hat{\beta}_j^-$ reflects the signal, whereas $\hat{\beta}_j^+ - \hat{\beta}_j^-$ cancels out the signal and only reflects the noise effect. Based on this observation, we would declare a variable $\boldsymbol{X}_j$ as "significant" if $M_j \geq t$ for some threshold $t > 0$.

The choice of t depends on the number of false positives and the associated FDP. If we can choose GM properly such that M_j follows a distribution symmetric about zero, that is,

$$P\left(M_j \geq s\right) = P\left(M_j \leq -s\right), \quad \forall s.$$

Then, $\#\left\{j \in \mathcal{S}_0 \mid M_j \geq t\right\}$ can be approximated by $\#\left\{j \in \mathcal{S}_0 \mid M_j \leq -t\right\}$. The FDP$(t)$ can be estimated by

$$\widehat{\mathrm{FDP}}(t) \triangleq \frac{\#\left\{j \mid M_j \leq -t\right\}}{\#\left\{j \mid M_j \geq t\right\} \vee 1}.$$

For any designated FDR level q, we could choose τ_q such that

$$\tau_q = \min_t \{\widehat{\mathrm{FDP}}(t) \leq q\}$$

and reject the j th hypothesis if $M_j \geq \tau_q$.

4 False discovery rate control via data splitting

Motivated by the GM method, now, we introduce another method for constructing mirror statistics based on data splitting (DS). DS aims at obtaining two independent regression coefficients for each feature via two potentially different statistical fitting procedures applied to each part of the data. Using the two estimates, DS then constructs a test statistic M_j for each feature X_j, referred to as the "mirror statistic" or "symmetric statistic" in the literature, which should possess the following two key properties

(P1) A feature with a larger positive mirror statistic is more likely to be a relevant feature.

(P2) The sampling distribution of the mirror statistic of any null feature is symmetric about zero.

Property (P1) suggests that we can rank the importance of each feature by its mirror statistic, and select those features with mirror statistics larger than a cutoff (τ). Property (P2) implies that we can estimate (conservatively) the number of false positives, that is, $\#\{j : j \in S_0, M_j > \tau\}$, using the left tail of the distribution, $\#\{j : M_j < -\tau\}$. The relative references are as below.

- Dai, C., Lin, B., Xing, X. and Liu, J.-S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, online.

Feature selection commonly relies on a set of coefficients $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$ to measure the importance of each feature. The larger $|\widehat{\beta}_j|$ is, the more likely feature X_j is useful in predicting y (since features have been normalized). For example, in linear regressions, $\widehat{\boldsymbol{\beta}}$ can be the vector of coefficients estimated via OLS or some shrinkage methods.

In contrast to common practices that select features based on a single set of estimated coefficients, we construct two independent sets of estimates, $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$, potentially with two different statistical procedures, in order to set up an FDR control framework.

The independence between $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ is ensured by employing a data-splitting strategy. More precisely, we split the n observations into two groups, denoted as $\left(\boldsymbol{y}^{(1)}, \boldsymbol{X}^{(1)}\right)$, and $\left(\boldsymbol{y}^{(2)}, \boldsymbol{X}^{(2)}\right)$, and then estimate $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ using each part of the data. The data-splitting procedure is flexible, as long as it is independent of the response vector y. A general form of the mirror statistic M_j is

$$M_j = \text{sign}(\widehat{\beta}_j^{(1)}\widehat{\beta}_j^{(2)}) f(|\widehat{\beta}_j^{(1)}|, |\widehat{\beta}_j^{(2)}|), \tag{4.1}$$

where function $f(u, v)$ is a nonnegative, exchangeable, and monotonically increasing function defined for nonnegative u and v. Three convenient choices of $f(u, v)$ are:

$$f(u, v) = 2\min(u, v), \quad f(u, v) = uv, \quad f(u, v) = u + v. \tag{4.2}$$

The symmetry property of the mirror statistics for the null features gives us an upper bound of the number of false positives:

$$\#\{j \in S_0 : M_j > t\} \approx \#\{j \in S_0 : M_j < -t\} \leq \#\{j : M_j > t\}, \quad \forall t > 0, \tag{4.3}$$

where S_0 is the null set. The $\mathrm{FDP}(t)$ of the selection $\widehat{S}_t = \{j : M_j > t\}$, as well as an "overestimate" of it are thus given by

$$\mathrm{FDP}(t) = \frac{\#\{j : M_j > t, j \in S_0\}}{\#\{j : M_j > t\} \vee 1}, \quad \widehat{\mathrm{FDP}}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1}.$$

For any designated FDR control level $q \in (0, 1)$, we can choose the data-driven cutoff τ_q as follows:

$$\tau_q = \min\{t > 0 : \widehat{\mathrm{FDP}}(t) \leq q\},$$

and the final selection is $\widehat{S}_{\tau_q} = \{j : M_j > \tau_q\}$. To achieve FDR control, the following two assumptions are very important.

- **Assumption 1.** (Symmetry) For $j \in S_0$, the sampling distribution of at least one of $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$ is symmetric about zero.
- **Assumption 2.** (Weak dependence) The mirror statistics M_j's are continuous random variables, and there exist constants $c > 0$ and $\alpha \in (0, 2)$ such that

$$\mathrm{Var}\big(\sum_{j \in S_0} (M_j > t)\big) \leq c p_0^{\alpha}, \quad \forall t \in \mathbb{R}, \quad \text{where} \quad p_0 = |S_0|.$$

Now, we consider the application of DS in Gaussian graphical models. Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a p-dimensional multivariate Normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ as the precision matrix. Without loss of generality, we assume $\boldsymbol{\mu} = 0$. One can define a corresponding Gaussian graphical model (V, E), in which the set of vertices is $V = (X_1, \ldots, X_p)$, and there is an edge between two different vertices X_i and X_j if X_i and X_j are conditionally dependent given $\{X_k, k \neq i, j\}$. The graph estimation can be recast as a node-wise regression problem. Given a nominal level q, we apply DS to each node-wise regression targeting at an FDR control level $q/2$ and then combine the node-wise selection results using the "OR" rule. A heuristic justification of the proposed method is

given below:

$$\begin{aligned}
\mathrm{FDP} &= \frac{\#\{(i,j) \in \widehat{E}_{\mathrm{OR}}, (i,j) \notin E\}}{|\widehat{E}_{\mathrm{OR}}| \vee 1} \\
&\leq \frac{\sum_{j=1}^{p} \#\left\{i \notin ne_j, i \in \widehat{ne}_j\right\}}{\frac{1}{2}\sum_{j=1}^{p} \#\left\{i \in \widehat{ne}_j\right\} \vee 1} \\
&= \frac{\sum_{j=1}^{p} \#\left\{i \notin ne_j, M_{ji} > \tau^j_{q/2}\right\}}{\frac{1}{2}\sum_{j=1}^{p} \#\left\{M_{ji} > \tau^j_{q/2}\right\} \vee 1} \\
&\approx \frac{\sum_{j=1}^{p} \#\left\{i \notin ne_j, M_{ji} < -\tau^j_{q/2}\right\}}{\frac{1}{2}\sum_{j=1}^{p} \#\left\{M_{ji} > \tau^j_{q/2}\right\} \vee 1} \\
&\leq 2 \max_{1\leq j\leq p} \frac{\#\left\{i \notin ne_j, M_{ji} < -\tau^j_{q/2}\right\}}{\#\left\{M_{ji} > \tau^j_{q/2}\right\} \vee 1} \leq q,
\end{aligned} \tag{4.4}$$

where ne_j is the neighborhood of vertex X_j. For the j-th nodewise regression, M_{ji} is the mirror statistic of X_i, $i \neq j$, and $\tau^j_{q/2}$ is the selection cutoff of the mirror statistics. The first inequality in (4.4) is based on the fact that each edge can be selected at most twice. The approximation in the middle uses the symmetry property of the mirror statistics. The second to last inequality follows from the elementary inequality that $\left(\sum_n a_n\right) / \left(\sum_n b_n\right) \leq \max_n a_n/b_n$ for $a_n \geq 0, b_n > 0$.

Let $s = \max_{j\in\{1,\ldots,p\}} |ne_j|$. To theoretically justify the methods, we first show that with probability approaching one, the symmetry assumption is simultaneously satisfied in all nodewise regressions under the following assumptions.

Assumption 3.

1. (Regularity condition) $c \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq 1/c$ for some $c > 0$.
2. (Sparsity condition) $s = o(n/\log p)$.
3. (Signal strength condition)

$$\min\left\{|\Omega_{ij}| : \Omega_{ij} \neq 0\right\} \gg \sqrt{s \log p/n}.$$

Under Assumption 3, we have the following proposition.

Proposition 1 *Under Assumption 3, as $n, p \to \infty$ satisfying $\log p = o(n)$, the symmetry assumption (Assumption 1) is simultaneously satisfied in all nodewise regressions with probability approaching one. Also, the weak dependence assumption (Assumption 2) is implied by the regularity condition in Assumption 3.*

The following proposition shows the asymptotic FDR control of DS and MDS.

Theorem 4.1 *Assume that Assumption 3 holds and that* $\min_{j\in\{1,\ldots,p\}} |ne_j| / \log p \to \infty$. *For any nominal FDR level* $q \in (0,1)$, *the DS method achieves*

$$\limsup_{n,p\to\infty} \ FDR \ \leq q \ \textit{and} \ \liminf_{n,p\to\infty} \ Power \ = 1$$

in the asymptotic regime where $\log p = o\left(n^{\xi}\right)$ *for some* $\xi \in (0,1)$.