STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 7

1 PRDS property

In the previous lecture, we have shown that the BH procedure provides FDR control at the level $n_0 S_n \alpha/n$, where $S_n \approx \log(n)$. To ensure the FDR control at level α, one has to run the BH procedure at level α/S_n, which can be very conservative as $\alpha/S_n \to 0$ as $n \to +\infty$.

In today's lecture, we consider the special type of dependence structure named positive regression dependency on each one from a subset (PRDS) and show that the BH procedure controls the FDR at the desired level when the p-values exhibit the PRDS property.

1.1 Definitions

For two vectors $x, y \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we write $y \geq x$ if $y_i \geq x_i$ for all i. A subset D of $\mathbb{R}^n$ is said to be increasing if for all $x \in D$, $y \geq x$ implies that $y \in D$.

Definition. A family of random variables $(X_1, \ldots, X_n)$ is said to be PRDS on a subset $I_0 \subset \{1, 2, \ldots, n\}$ if for all $i \in I_0$, the function $P((X_1, \ldots, X_n) \in D | X_i = x)$ is an increasing function of x for any increasing subset D.

We have the following two observations.

- If $(X_1, \ldots, X_n)$ is PRDS on I_0 and if $Y_i := f_i(X_i)$ for all $1 \leq i \leq n$ where each f_i is strictly increasing or decreasing, then $(Y_1, \ldots, Y_n)$ is PRDS on I_0 as well. Transformation of this form is called co-monotone transformation. Thus, the PRDS property is preserved under co-monotone transformations.
- If $(X_1, \ldots, X_n)$ is PRDS on I_0 (the set of true nulls), then both $p_i = F(X_i)$ (the right-sided p-values) and $p_i = 1 - F(X_i)$ (the left-sided p-values) are PRDS as well. Here F is the CDF of X_i under the null. This follows from the fact that the CDF and survival functions are co-monotone transforms, and hence, the p-values are PRDS by the preceding observation.

Exercise 7.1: Prove the two observations above.

1.2 An example

Theorem. Let $X = (X_1, \ldots, X_n)$ be a multivariate Gaussian random vector with mean μ and covariance $\Sigma = (\sigma_{ij})$. X is PRDS on I_0 if and only if $\sigma_{ij} \geq 0$ for any $i \in I_0$ and $1 \leq j \leq n$.

Proof. According to the definition, we need to show that $\mathbb{P}(X \in D | X_i = x)$ is an increasing function of x for any increasing subset D for any $i \in I_0$. Without loss of generality, we assume that $1 \in I_0$ and $i = 1$. Write

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_{-1} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}.$$

Then we have

$$X_{-1} | X_1 = x \sim N\left(\mu_{-1} + \Sigma_{-1,1}\Sigma_{1,1}^{-1}(x - \mu_1), \Sigma_{-1,-1} - \Sigma_{-1,1}\Sigma_{1,1}^{-1}\Sigma_{1,-1}\right).$$

As $\sigma_{ij} \geq 0$, we have $\Sigma_{-1,1} \geq 0$ entrywise, which implies that the conditional mean $\mu_x := \mu_{-1} + \Sigma_{-1,1}\Sigma_{1,1}^{-1}(x - \mu_1)$ is a non-decreasing function in x. Hence, if $y \geq x$, $\mu_y \geq \mu_x$ entrywise.

Let $U \sim N\left(0, \Sigma_{-1,-1} - \Sigma_{-1,1}\Sigma_{1,1}^{-1}\Sigma_{1,-1}\right)$. We have

$$\begin{aligned}
&\mathbb{P}(X \in D|X_1 = x) \\
=&\mathbb{P}((x, X_{-1}) \in D|X_1 = x) \\
=&\mathbb{P}((x, U + \mu_x) \in D) \\
\leq&\mathbb{P}((y, U + \mu_y) \in D) \\
=&\mathbb{P}((y, X_{-1}) \in D|X_1 = y) \\
=&\mathbb{P}(X \in D|X_1 = y)
\end{aligned}$$

which implies that $\mathbb{P}(X \in D|X_1 = x)$ is non-decreasing in x.

Conversely, if we want to prove that PRDS implies that all correlations are non-negative, we can proceed by contradiction. Assume that there is some $\sigma_{1j} < 0$ for some $j \neq 1$. We have

$$X_j|X_1 = x \sim N\left(\mu_j + \sigma_{j1}\sigma_{11}^{-1}(x - \mu_1), \sigma_j^2\right).$$

where σ_j^2 does not rely on x. The conditional mean is seen to be a strictly decreasing function of x (as $\sigma_{j1} < 0$), which gives that the conditional probability of the event $\{X_j \geq \mu_j\}$ is strictly decreasing in x. Since the set $\{X_j \geq \mu_j\}$ is increasing, we have a contradiction to the PRDS property.

2 FDR control under PRDS

Benjamini and Yekutieli (2001) proved the following theorem.

Theorem. The BH procedure controls the FDR at the level $n_0\alpha/n$ when the p-values $\{p_1, \ldots, p_n\}$ are PRDS on the set of true nulls.

As noted before, PRDS property translates from statistics to one-sided p-values. Hence, to apply the above theorem, we can simply check the PRDS property on the statistics itself.

This theorem asserts FDR control without assuming any dependence structure on the non-null p-values. This is desirable since we usually do not know about the structure of the non-null p-values. However, it does assume the PRDS property, which involves knowing how the non-nulls relate to the true nulls, which is generally not well known. Thus, the theorem is difficult to apply in practice.

Proof. Without loss of generality, let us assume that $H_{0,1}, \ldots, H_{0,n_0}$ are the true nulls. We know that

$$\text{FDR} = \sum_{i=1}^{n_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq T\}}{R \vee 1}\right],$$

where $T = \alpha R/n$ with R being the number of rejections. We only need to show that

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq T\}}{R \vee 1}\right] \leq \frac{\alpha}{n}$$

for all $i = 1, 2, \ldots, n_0$. Note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq T\}}{R \vee 1}\right] =& \sum_{k=1}^{n} \mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq k\alpha/n, R = k\}}{k}\right] \\
=& \sum_{k=1}^{n} \frac{\mathbb{P}(R = k|p_i \leq k\alpha/n)\mathbb{P}(p_i \leq k\alpha/n)}{k} \\
=& \sum_{k=1}^{n} \frac{k\alpha}{n} \frac{\mathbb{P}(R = k|p_i \leq k\alpha/n)}{k} \\
=& \frac{\alpha}{n} \sum_{k=1}^{n} \mathbb{P}(R = k|p_i \leq k\alpha/n).
\end{aligned}$$

We see that $\{R \leq k\}$ can be written as $\{(p_1, \ldots, p_n) \in D\}$ for some increasing set D. This is because increasing all p-values increases the p-value at each rank. Hence, any ranked p-value above the threshold remains above its threshold, i.e., we accept at least as many as before and, hence, do not reject more hypotheses. Using this fact, we have

$$\begin{aligned}\sum_{k=1}^{n} \mathbb{P}(R = k | p_i \leq k\alpha/n) =& \sum_{k=1}^{n} \{\mathbb{P}(R \leq k | p_i \leq k\alpha/n) - \mathbb{P}(R \leq k-1 | p_i \leq k\alpha/n)\} \\ =& \mathbb{P}(R \leq n | p_i \leq \alpha) - \mathbb{P}(R \leq 0 | p_i \leq \alpha/n) \\ &+ \sum_{k=1}^{n-1} \{\mathbb{P}(R \leq k | p_i \leq k\alpha/n) - \mathbb{P}(R \leq k | p_i \leq (k+1)\alpha/n)\}.\end{aligned}$$

As $\mathbb{P}(R \leq k | p_i \leq x)$ is increasing in x by the Lemma below, each summand in the summation in the second line above is non-positive. Thus, we must have

$$\sum_{k=1}^{n} \mathbb{P}(R = k | p_i \leq k\alpha/n) \leq \mathbb{P}(R \leq n | p_i \leq \alpha) \leq 1,$$

which completes the proof.

Lemma. If the p-values are PRDS on the set of true nulls, then the function $\mathbb{P}((p_1, \ldots, p_n) \in D | p_i \leq t)$ is non-decreasing in t for an increasing set D and true null i.

Proof. Write $\mathbf{p} = (p_1, \ldots, p_n)$. We first observe that

$$\mathbb{P}(\mathbf{p} \in D | p_i \leq t) = \frac{\mathbb{P}(\mathbf{p} \in D, p_i \leq t)}{\mathbb{P}(p_i \leq t)}.$$

For $t' > t$, we get

$$\mathbb{P}(\mathbf{p} \in D | p_i \leq t') = \frac{\mathbb{P}(\mathbf{p} \in D, p_i \leq t) + \mathbb{P}(\mathbf{p} \in D, p_i \in (t, t'])}{\mathbb{P}(p_i \leq t) + \mathbb{P}(p_i \in (t, t'])}.$$

It suffices to show that

$$\frac{\mathbb{P}(\mathbf{p} \in D, p_i \leq t)}{\mathbb{P}(p_i \leq t)} \leq \frac{\mathbb{P}(\mathbf{p} \in D, p_i \in (t, t'])}{\mathbb{P}(p_i \in (t, t'])}.$$

Letting F_i be the CDF of p_i, we have

$$\begin{aligned}\mathbb{P}(\mathbf{p} \in D, p_i \leq t) =& \int_0^t \mathbb{P}(\mathbf{p} \in D | p_i = s) F_i(ds) \\ \leq& \int_0^t \mathbb{P}(\mathbf{p} \in D | p_i = t) F_i(ds) \\ =& \mathbb{P}(\mathbf{p} \in D | p_i = t) F_i(t),\end{aligned}$$

which implies that

$$\frac{\mathbb{P}(\mathbf{p} \in D, p_i \leq t)}{\mathbb{P}(p_i \leq t)} \leq \mathbb{P}(\mathbf{p} \in D | p_i = t).$$

On the other hand,

$$\begin{aligned}\mathbb{P}(\mathbf{p} \in D, p_i \in (t, t']) =& \int_t^{t'} \mathbb{P}(\mathbf{p} \in D | p_i = s) F_i(ds) \\ \geq& \int_t^{t'} \mathbb{P}(\mathbf{p} \in D | p_i = t) F_i(ds) \\ =& \mathbb{P}(p_i \in (t', t]) \mathbb{P}(\mathbf{p} \in D | p_i = t),\end{aligned}$$

which suggests that

$$\frac{\mathbb{P}(\mathbf{p} \in D, p_i \leq t)}{\mathbb{P}(p_i \leq t)} \leq \mathbb{P}(\mathbf{p} \in D | p_i = t) \leq \frac{\mathbb{P}(\mathbf{p} \in D, p_i \in (t, t'])}{\mathbb{P}(p_i \in (t, t'])}.$$

3 The FDR conjecture

Let $X = (X_1, \ldots, X_n)$ be a set of Z-statistics following the multivariate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_n)$ and covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 1$. We are interested in testing the two-sided hypothesis:

$$H_{0,i} : \mu_i = 0 \text{ versus } H_{a,i} : \mu_i \neq 0, \quad i = 1, 2, \ldots, n.$$

The two-sided p-value in this case is defined as $p_i = 2(1 - \Phi(|x_i|))$, where Φ is the CDF of the standard normal distribution.

Conjecture. The BH procedure applied to the p-values $\{p_i\}_{i=1}^n$ controls the FDR at level α regardless of the form of Σ.