Stats 300C: Theory of Statistics Spring 2018

Lecture 10 — April 23, 2018

Prof. Emmanuel Candes

Scribe: Emmanuel Candes, Pragya Sur, Jing Miao

1 Outline

Agenda: FDR Control under Dependence

1. The PRDS property
2. Examples of PRDS distributions
3. FDR control under PRDS property

Recall that last time we discussed FDR control via the BH(q) procedure under dependence. In particular, we discussed a result due to Benjamini-Yekutieli (2001) saying that under dependence this procedure controls FDR at level $q \cdot S(n)$. For large n, this is close to $q \cdot (\log n + 0.577)$. In this lecture, we show that the orginal BHq procedure controls the FDR under a notion of positive correlation between test statistics or p-values.

2 The PRDS property

We begin first by the definition of increasing/decreasing sets. (Below $x \geq y$ means that $x_i \geq y_i$ for all coordinates.)

Definition 1: *A set $D \in \mathbb{R}^n$ is called increasing if $x \in D$ and $y \geq x$ implies $y \in D$. (These sets have no boundaries in the North-East directions).*

Now we define PRDS.

Definition 2: *A family of random variable $X = (X_1, ..., X_n)$ is PRDS [positive regression dependence on each of a subset] on I_0, if for any increasing set and each $i \in I_0$, $\mathbb{P}((X_1, ..., X_n) \in D | X_i = x)$ is increasing in x.*

We make a few observations concerning this definition.

- The PRDS property is invariant by co-monotone transformations. If $Y_i = f_i(X)$, where all the f_i's are either increasing or decreasing, then X is PRDS implies that Y is also PRDS.

- D is increasing if and only if D^c is decreasing. As a consequence, we have that a random vector X is PRDS if and only if for any decreasing C, $\mathbb{P}(X \in C | X_i = x_i)$ is decreasing in x.

- If $\{X_i\}$ is PRDS on I_0 (true nulls), then $p_i = \overline{F}_{H_i}(X_i)$ [right-sided p-value] and $p_i = F_{H_i}(X_i)$ [left-sided p-value] are both PRDS. But for two-sided test, $p_i = 2\overline{F}_{H_i}(|X_i|)$. Since $|X_i|$ is not a monotone transformation, p_i may not be PRDS.

3 Examples of PRDS distributions

Claim: Take a multivariate normal Gaussian distribution $X = (X_1, ., , , X_n) \sim \mathcal{N}(\mu, \Sigma)$. If $\Sigma_{ij} \geq 0$ for all $i \in I_0$ and all j, then $X = (X_1, ., , , , X_n)$ is PRDS over I_0. (The converse also holds).

Remark: With Gaussian data, PRDS is equivalent to positive correlations.

Proof.

$$X = \begin{pmatrix} X_1 \\ X_{(-1)} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_{(-1)} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{1(-1)} \\ \Sigma_{(-1)1} & \Sigma_{(-1)(-1)} \end{pmatrix}$$

In this setup, the distribution of $X_{(1)}$ given $X_1 = x$ is given by

$$\mathcal{L}(X_{(1)} \mid X_1 = x) = N\left(\mu_{(-1)} + \Sigma_{(-1)1}\Sigma_{11}^{-1}(x - \mu_1), \Sigma_{(-1)(-1)} - \Sigma_{(-1)1}\Sigma_{11}^{-1}\Sigma_{1(-1)}\right)$$

If $\Sigma_{(-1),1}$ is positive, then the conditional means increase with x. The covariance does not depend on x, so the conditional distribution is stochastically increasing in x. Thus for any nondecreasing f, $x \leq x'$ implies

$$\mathbb{E}(f(X|X_1 = x) \leq \mathbb{E}(f(X)|X_1 = x')$$

Taking f to be the indicator function of an increasing set verifies the PRDS property. □

4 FDR control under PRDS property

Theorem 3 (Benjamini & Yekutieli (2001))**:** *If the joint distribution of the statistics (or joint dist. of the p-values) is PRDS on the set of true nulls $\mathcal{H}_0$, then the Benjamini-Hochberg procedure* BH(q) *controls the FDR at level $q\frac{n_0}{n}$. (BH(q) may become conservative under positive dependence since FDR $= qn_0/n$ under independence).*

Remark: An important feature of this theorem is that it does not make explicit assumptions on the dependence structure among the non-null hypotheses. This is good from the point of view of applications, where we typically have some knowledge of the phenomenon under the null, but know (and are willing to assume) very little of the phenomenon under the alternative. Unfortunately, we typically don't know much about how the non-nulls depend on the nulls, so it's generally not known whether the statistics arising in a particular application are PRDS.

A consequence of the PRDS property is that for $t \leq t'$,

$$\mathbb{P}(D|p_i \leq t) \leq \mathbb{P}(D|p_i \leq t')$$

if i is a null and the set D is increasing.

Proof. **[E. Candès, R. Foygel Barber]**
We know that

$$\text{FDR} = \mathbb{E}\Big(\sum_{i\in\mathcal{H}_0}\frac{V_i}{1\vee R}\Big), \qquad V_i = \mathbf{1}\{\text{ reject } H_i\}.$$

Recall that in the independent case: $\mathbb{E}V_i/(1\vee R) = q/n$ (no matter the null the expectation is always q/n). Now we want to show that $\mathbb{E}V_i/(1\vee R) \le q/n$. This will then imply that FDR is at most qn_0/n. Set $q_k = qk/n$ and note that

$$\begin{aligned}
\frac{V_i}{1\vee R} &= \sum_{k\ge 1}\frac{\mathbf{1}\{p_i\le q_k\}\mathbf{1}\{R=k\}}{k}\\
&= \sum_{k\ge 1}\frac{\mathbf{1}\{p_i\le q_k\}(\mathbf{1}\{R\le k\}-\mathbf{1}\{R\le k-1\})}{k}\\
&= \sum_{k=1}^{n-1}\left[\frac{\mathbf{1}\{p_i\le q_k\}}{k}-\frac{\mathbf{1}\{p_i\le q_{k+1}\}}{k+1}\right]\mathbf{1}\{R\le k\}+\frac{\mathbf{1}\{R\le n\}\mathbf{1}\{p_i\le q\}}{n}\ \text{(Integration by parts)}\ .
\end{aligned}$$

Note that

$$\mathbb{E}\Big(\frac{\mathbf{1}\{R\le n\}\mathbf{1}\{p_i\le q\}}{n}\Big) = \frac{q}{n}$$

always holds, since $R\le n$ all the time and $p_i\sim U(0,1)$ under the null. If we can prove

$$\mathbb{E}\Big(\sum_{k=1}^{n-1}\left[\frac{\mathbf{1}\{p_i\le q_k\}}{k}-\frac{\mathbf{1}\{p_i\le q_{k+1}\}}{k+1}\right]\mathbf{1}\{R\le k\}\Big)\le 0,$$

we are done. For each k, we have

$$\begin{aligned}
&\mathbb{E}\Big(\left[\frac{\mathbf{1}\{p_i\le q_k\}}{k}-\frac{\mathbf{1}\{p_i\le q_{k+1}\}}{k+1}\right]\mathbf{1}\{R\le k\}\Big)\\
&\quad= \frac{\mathbb{P}(p_i\le q_k, R\le k)}{k}-\frac{\mathbb{P}(p_i\le q_{k+1}, R\le k)}{k+1}\\
&\quad= \frac{\mathbb{P}(R\le k|p_i\le q_k)\mathbb{P}(p_i\le q_k)}{k}-\frac{\mathbb{P}(R\le k|p_i\le q_{k+1})\mathbb{P}(p_i\le q_{k+1})}{k+1}\\
&\quad\le \frac{\mathbb{P}(R\le k|p_i\le q_{k+1})\mathbb{P}(p_i\le q_k)}{k}-\frac{\mathbb{P}(R\le k|p_i\le q_{k+1})\mathbb{P}(p_i\le q_{k+1})}{k+1}\\
&\quad= 0
\end{aligned}$$

where the inequality follows from the PRDS property. Indeed $\{R\le k\}$ is an increasing set since when the p_i's increase for each i, R decreases (we make fewer rejections).

□

Example: Suppose $X\sim N(\mu,\Sigma)$. If we wish to test $H_{0i}:\mu_i=0$ vs $H_{1i}:\mu_i>0$, in the setup of the Claim above, our test statistics X_i are PRDS. The same conclusion holds if we wish to test $H_{0i}:\mu_i=0$ vs $H_{1i}:\mu_i<0$. However, for testing against the alternative $\mu_1\neq 0$, the test statistics $|X_i|$ no longer have the PRDS property.