Week 12
Spring 2009

Lecture 23. FDR, Model selection and Sharp Asymptotic Minimaxity (I) – Introduction to Benjamini and Hochberg procedure

Why multiple testing?

For example, Tukey notes that carrying out 250 independent tests of significance, each at the 0.05 level, will result on average in 12.5 apparently significant results when the intersection null hypothesis of no effects is true. Thus obtaining (say) 18 significant results is no cause for exultation [TUKEY, J. W. (1953). *The problem of multiple comparisons*, pages 75–76].

Bonferroni correction.

Bonferroni CE (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3-62.

Definition of FDR

Seeger, P. (1968, Technometrics). Benjamini and Hochberg (1995, JRSSB).

Let Θ be the parameter space.

Let $H_1, H_2, \ldots, H_n$ be null-hypotheses.

Let $I_n = \{i : i = 1, 2, \ldots, n\}$, $I_{n,0} = \{i : H_i \text{ is true}\}$, $n_0 = |I_{n,0}|$.

Let $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ be multiple test procedure, and $\varphi_i = 1$ means "reject H_i" and $\varphi_i = 0$ means "remain H_i".

Let $V_n = |\{i : \varphi_i = 1 \text{ but } H_i \text{ is true}\}|$ = number of true hypotheses rejected.

Let $R_n = |\{i : \varphi_i = 1\}|$ = number of hypotheses rejected.

Define

$$FDR_\theta(\varphi) = E_\theta \frac{V_n}{R_n \vee 1}$$

BH procedure*:*Let $P_1, P_2, \ldots, P_n$ be independent p-values.

Reject all H_i with $P_i \le k\alpha/n$, $k = \max\left\{i, P_{(i)} \le i\alpha/n\right\} = \max\left\{i, \frac{P_{(i)}}{\alpha} \le i/n\right\}$.

Result:

$$FDR_\theta(\varphi_{BH}) \le \frac{n_0}{n}\alpha.$$

Proof under the assumption that $P_1, P_2, \ldots, P_n$ are independent:

$$
\begin{aligned}
E\frac{V_n}{R_n \vee 1} &= \sum_{i \in I_{n,0}} \sum_{j=1}^{n} \frac{1}{j} P\left(R_n = j, \varphi_i = 1\right) \\
&= n_0 \sum_{j=1}^{n} \frac{1}{j} P\left(R_n = j, P_1 \leq \frac{j}{n}\alpha\right) \qquad \text{(assume that } 1 \in I_{n,0} \text{ WLOG)} \\
&= n_0 \sum_{j=1}^{n} \frac{1}{j} P\left(R_n = j | P_1 \leq \frac{j}{n}\alpha\right) P\left(P_1 \leq \frac{j}{n}\alpha\right) \\
&\leq \frac{n_0}{n}\alpha \left[\begin{array}{c} \sum_{j=1}^{n-1} \left(P\left(R_n \geq j | P_1 \leq \frac{j}{n}\alpha\right) - P\left(R_n \geq j+1 | P_1 \leq \frac{j}{n}\alpha\right)\right) \\ +P\left(R_n \geq n | P_1 \leq \frac{n}{n}\alpha\right) \end{array} \right] \\
&= \frac{n_0}{n}\alpha \left[\begin{array}{c} \sum_{j=2}^{n} \left(P\left(R_n \geq j | P_1 \leq \frac{j}{n}\alpha\right) - P\left(R_n \geq j | P_1 \leq \frac{j-1}{n}\alpha\right)\right) \\ +P\left(R_n \geq 1 | P_1 \leq \frac{1}{n}\alpha\right) \end{array} \right] = \frac{n_0}{n}\alpha.
\end{aligned}
$$

where the last equality follows from the the assumption of independence which implies

$$
P\left(\exists k \geq j, \sum_i \left\{i : P_i \leq \frac{k}{n}\alpha\right\} \geq k | P_1 \leq \frac{j-1}{n}\alpha\right) = P\left(\exists k \geq j, \sum_i \left\{i : P_i \leq \frac{k}{n}\alpha\right\} \geq k | P_1 \leq \frac{j}{n}\alpha\right).
$$

Remark 1 *The positive regression dependence implies*

$$
P\left(R_n \geq j | P_1 \leq \frac{j}{n}\alpha\right) - P\left(R_n \geq j | P_1 \leq \frac{j-1}{n}\alpha\right) \leq 0.
$$

Bayesian approach.

Model: Assume that H_i be i.i.d. Ber(π_0). Let Y_i be i.i.d. with

$$
Y_i | H_i = 0 \sim F_0,\ Y_i | H_i = 1 \sim F_1
$$

Let f_0 and f_1 be corresponding densities of F_0 and F_1. Write $F(y) = \pi_0 F_0(y) + (1 - \pi_0) F_1(y)$. It is the distribution function of y.

Then

$$
\begin{aligned}
P\left(H_i = 0 | Y_i \leq y\right) &= \frac{\pi_0 F_0(y)}{F(y)}, \text{ “}q-value\text{” in Storey (2001)} \\
P\left(H_i = 1 | Y_i \leq y\right) &= \frac{(1-\pi_0) F_1(y)}{F(y)}
\end{aligned}
$$

Empirical Bayes interpretation of BH procedure: Storey (2001). Let $Y_i = P_i$. An obvious estimator for “$q - value$” is

$$
\widehat{P}\left(H_i = 0 | Y_i \leq y\right) \approx \frac{\pi_0 F_0(y)}{\widehat{F}(y)} = \frac{\pi_0 y}{\widehat{F}(y)}
$$

where $\widehat{F}(y) = \sum_i \{Y_i \le y\}$. We choose y or $\{Y_i \le y\}$ as large as possible subject to that the estimated Bayes proportion of false discoveries

$$\frac{\pi_0 y}{\widehat{F}(y)} \le \alpha,$$

or equivalently the largest i or $P_{(i)}$ such that

$$\frac{\pi_0 P_{(i)}}{i/n} \le \alpha, \text{ or } \frac{n\pi_0 P_{(i)}}{i} \le \alpha,$$

where i is number of rejection and $n\pi_0 P_{(i)}$ is approximately the number of false rejections.

Local FDR: Robbins (1951) and some recent papers of Efron.

$$\begin{aligned} P(H_i = 0 | Y_i = y) &= \frac{\pi_0 f_0(y)}{\pi_0 f_0(y) + (1-\pi_0) f_1(y)} \\ P(H_i = 1 | Y_i = y) &= \frac{(1-\pi_0) f_1(y)}{\pi_0 f_0(y) + (1-\pi_0) f_1(y)} \end{aligned}$$

In Efron (2006), estimate f_0 and f_1 and π_0.

More Topics

1. Asymptotics.
2. Estimating $\pi_0 = \frac{n_0}{n}$.
3. Incorporating additional structure, e.g., group, graph etc.
4. Optimality.
5. Model selection.

Lecture 24. FDR, Model selection and Sharp Asymptotic Minimaxity (II) – Lower and Upper bounds

Consider the multivariate Gaussian mean problem:

$$y_i = \theta_i + \sigma z_i,\ z_i \stackrel{i.i.d.}{\sim} N(0,1),\ i = 1, 2, \ldots, n.$$

Goal: adaptively recover θ to unknown sparsity.
Sparsity assumption:

$$\Theta_{n,p} = \left\{ \theta \in \mathbb{R}^n : \left(\frac{1}{n} \sum_{i=1}^{n} |\theta_i|^p \right)^{1/p} \leq \eta_n \right\}.$$

Penalized estimation: find $\hat{\theta}$ to minimize

$$K(\theta, y) = \|y - \theta\|_2^2 + Pen(\theta)$$

1. l_0 penalty:
$$Pen(\theta) = \lambda \|\theta\|_0 .$$
Includes AIC (Akaike,1973), BIC(Schwarz, 1978), RIC (Foster and George, 1994).

2. l_1 penalty:
$$Pen(\theta) = \lambda \|\theta\|_1 .$$
LASSO or soft thresholding.

3. FDR penalty:
$$Pen(\theta) = \sum_{j=1}^{\|\theta\|_0} t_j^2 \triangleq \sum_{j=1}^{\|\theta\|_0} z^2 \left(\frac{q}{2} \cdot \frac{j}{n} \right)$$
(weighted l_0 penalty?) and denote
$$\hat{\theta}^P = \arg\min_{\theta} \left[\|y - \theta\|_2^2 + Pen(\theta) \right].$$
Two FDR procedures are very closed to the solution $\hat{\theta}^P$ above: (1) Let $\hat{k}^F = \max\left\{k : P_{(k)} \leq kq/n\right\}$, and $\hat{t}^F = t_{\hat{k}^F}$, then estimate θ_i by y_i if $|y_i| \geq \hat{t}^{FDR}$ and 0 otherwise. The corresponding estimator of θ is denoted by θ^F. (2) Let $\hat{k}^G = \max\left\{k : P_{(i)} \leq iq/n \text{ for all } i \leq k\right\}$. The corresponding estimator of θ is denoted by θ^G.

4. Some other penalties
$$\hat{k} = \arg\min_{k} RSS(k) + \sigma^2 \lambda_k.$$

Foster and Stine (1999), $\lambda_k = 2\sigma^2 \sum_{j=1}^{k} \log(n/j)$. Tibshirani and Knight (1999), $\lambda_k = 4\sigma^2 \sum_{j=1}^{k} \log(n/j)$. Birgé and Massart (2001), $\lambda_k = 2\sigma^2 k \log(n/k)$. All of them are close to the FDR penalty which has

$$\lambda_k = \sigma^2 \sum_{j=1}^{k} z^2 \left(\frac{q}{2} \cdot \frac{j}{n} \right) \approx 2\sigma^2 k \log(n/k)$$

because

$$z^2 \left(\frac{q}{2} \cdot \frac{k}{n} \right) /2 \approx \log(n/k) - \frac{1}{2} \log\log(n/k) + c.$$

due to the fact that $\frac{\varphi(x)}{x} \sim \bar{\Phi}(x)$ as $x \to \infty$.

FDR and Sharp Asymptotic Minimaxity

Theorem. Let $q_n = 1/\log n$ and

$$\Theta_n = \left\{ \theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \leq n\eta_n^p \right\}$$

with $\eta_n^p \in \left[n^{-1} \log^5 n,\ n^{-\delta} \right], \delta > 0$. Then as $n \longrightarrow \infty$

$$\sup_{\theta \in \Theta_n} \mathbb{E}_{y|\theta} \sum_{i=1}^{n} \left(\hat{\theta}_i^P - \theta_i \right)^2 = (1 + o(1)) R(\Theta_n).$$

The result holds when $\hat{\theta}^P$ is replaced by $\hat{\theta}^F$ or $\hat{\theta}^G$.

(For a more complete statement of the result, please see Theorem 1.1. of Abramovich, Benjamini, Donoho and Johnstone, 2006, AOS)

Sketch of the proof for the lower bound

Let as $\eta_n \to 0$. In Donoho and Johnstone (1994, PTRF)

$$R(\Theta_n) \sim n\eta_n^p \left(2\log \eta_n^{-p} \right)^{(2-p)/2},\ p \leq 2\ (n\eta_n^2,\ p \geq 2).$$

Let $x \sim N(\mu, 1)$, the minimax Bayes risk

$$\mathcal{B}(\mathcal{P}_1, 1) = \inf_{\widehat{\mu}} \sup_{\mathcal{P}_1} \mathbb{E}_\mu \mathbb{E}_{x|\mu} (\widehat{\mu} - \mu)^2$$

where $\mathcal{P}_1 = \left\{ \pi : (\pi |\mu|^p)^{1/p} \leq \eta_n \right\}$, is

$$\mathcal{B}(\mathcal{P}_1, 1) \sim \eta_n^p \left(2\log \eta_n^{-p} \right)^{(2-p)/2}.$$

The least favorable configuration is

$$\pi = (1 - \beta_n)\delta_0 + \beta_n \delta_{\mu_n}$$

where

$$\beta_n = \eta_n^p \mu_n^{-p}, \mu_n \sim 2\left(\log \beta_n^{-1}\right)^{1/2}.$$

Roughly we use $\widehat{\mu}_{\sqrt{2\log \eta_n^{-p}}}$ to estimate μ, and $\left(2\log \eta_n^{-p}\right)^{1/2} \sim \left(2\log\left(n/k_n\right)\right)^{1/2}$ where k_n is effectively the nonzero number.

Sketch of the proof for the upper bound

Let

$$\theta_0 = \arg\min_{\mu}\left[\|\theta-\mu\|_2^2 + Pen\left(\mu\right)\right].$$

It is valid for all $\theta \in \mathbb{R}^n$,

$$\mathbb{E}_{y|\theta}\left\|\hat{\theta}^P - \theta\right\|^2 \leq K\left(\theta_0,\theta\right) + 2\mathbb{E}_{y|\theta}\left\langle \hat{\theta}^P - \theta_0, z\right\rangle.$$

since

$$\left\|\hat{\theta}^P - \theta\right\|^2 = \left\|y - \hat{\theta}^P\right\|^2 + 2\left\langle \hat{\theta}^P - \theta, z\right\rangle - \|z\|^2$$

and

$$\left\|y - \hat{\theta}^P\right\|^2 + Pen\left(\hat{\theta}^P\right) \leq \|y-\theta_0\|^2 + Pen\left(\theta_0\right).$$

Note that $\mathbb{E}_{y|\theta}\left\langle \hat{\theta}^P - \theta_0, z\right\rangle = \mathbb{E}_{y|\theta}\left\langle \hat{\theta}^P - \theta, z\right\rangle$.

Step 1: show that

$$\sup_{\theta\in\Theta_n} K\left(\theta_0,\theta\right) = \left(1+o\left(1\right)\right) R\left(\Theta_n\right).$$

Let $\theta_{[1]}^2 \geq \theta_{[2]}^2 \geq \cdots \geq \theta_{[n]}^2$. Note that

$$\begin{aligned}
\sup_{\theta\in\Theta_n} K\left(\theta_0,\theta\right) &= \sup_{\theta\in\Theta_n}\inf_{\mu}\left[\|\theta-\mu\|_2^2 + \sum_{j=1}^{\|\mu\|_0} t_j^2\right] \\
&= \sup_{\theta\in\Theta_n}\inf_{k}\left[\sum_{j=k+1}^{n}\theta_{[j]}^2 + \sum_{j=1}^{k} t_j^2\right] \\
&= \sup_{\theta\in\Theta_n}\sum_{j=1}^{n}\min\left\{\theta_{[j]}^2, t_j^2\right\}
\end{aligned}$$

subject to $\sum_{i=1}^{n}\left|\theta_{[i]}\right|^p \leq n\eta_n^p$. Equivalently we write

$$\sup_{x}\sum_{j=1}^{n}\min\left\{x_j^{2/p}, t_j^2\right\}, \text{ subject to} \sum_{j=1}^{n} x_j \leq n\eta_n^p,\ x_1 \geq x_2 \geq \ldots.$$

It is strictly convex on $\prod_j \left[0, t_j^p\right]$. The maximum is obtained at an extreme point of the constrained set. The extreme set has the form $[s_1^p, s_2^p, \ldots, s_l^p, 0, \ldots, 0]$. Let $\tilde{k}_n$ be the largest k for which

$$\sum_{j=1}^{k} t_j^p \leq n\eta_n^p$$

and recall that $t_j^2 = z^2\left(\frac{q}{2} \cdot \frac{j}{n}\right) \sim 2\log\left(n/j\right)$, then

$$\sup_{\theta \in \Theta_n} K\left(\theta_0, \theta\right) \sim \tilde{k}_n t_{\tilde{k}_n}^2 \sim n\eta_n^p \left(2\log \eta_n^{-p}\right)^{(2-p)/2}$$

Step 2: show that

$$\sup_{\theta \in \Theta_n} \mathbb{E}_{y|\theta} \left\langle \hat{\theta}^P - \theta, z \right\rangle \leq \frac{q_n}{1-q_n}\left(1+o\left(1\right)\right) R\left(\Theta_n\right).$$

Write

$$\left\langle \hat{\theta}^P - \theta, z \right\rangle = \sum_{i=1}^{n} z_i \left[\eta_H\left(y_i, t_{\hat{k}^P}\right) - \theta_i\right]$$

It is easy to observe that

$$z_i\left[\eta_H\left(y_i, t_2\right) - \theta_i\right] \leq z_i\left[\eta_H\left(y_i, t_1\right) - \theta_i\right]$$

if $|\theta_i| \leq t_1 \leq t_2$. This inspires to define a quantity $t_{k\_}\left(\theta\right)$ such that $t_{\hat{k}^P} \geq t_{k\_}\left(\theta\right)$ with high probability and

$$\mathbb{E}\sum_{i=1}^{n} z_i\left[\eta_H\left(y_i, t_{k\_}\right) - \theta_i\right] \leq \frac{q_n}{1-q_n}\left(1+o\left(1\right)\right) R\left(\Theta_n\right)$$

and one can define

$$k\_ = \frac{\eta_n^p\left(2\log\eta_n^{-p}\right)^{(2-p)/2}}{1-q_n-1/\log\log n}.$$

Let $S_n\left(\theta\right) = \left\{i : |\theta_i| \leq t_{k\_}\right\}$. Then write

$$\begin{aligned}
\mathbb{E}\sum_{i=1}^{n} z_i\left[\eta_H\left(y_i, t_{k_}\right) - \theta_i\right] &= \mathbb{E}\left\{\sum_{i \in S_n(\theta)} z_i\left[\eta_H\left(y_i, t_{k_}\right) - \theta_i\right] + \sum_{i \in S_n^c(\theta)} z_i\left[\eta_H\left(y_i, t_{k_}\right) - \theta_i\right]\right\} \\
&= T_0 + T_1
\end{aligned}$$

It can be shown the dominating term is

$$T_0 = \sum_{i \in S_n(\mu)} \mathrm{Cov}\left(y_i, \eta_H\left(y_i, t_{k\_}\right)\right) \sim 2nt_{k\_}\phi\left(t_{k\_}\right) \sim q_n R\left(\Theta_n\right),$$

and

$$T_1 = o\left(1\right) R\left(\Theta_n\right).$$