Week 13
Spring 2009

Lecture 25. FDR, Model selection and Sharp Asymptotic Minimaxity (II)

Model. Consider the standard multivariate normal mean problem:

$$y_i = \theta_i + \sigma_n z_i, \;\; z_i \overset{i.i.d.}{\sim} N(0,1), \;\; i = 1, ..., n. \tag{1}$$

The goal is to estimate parameter $\theta = (\theta_1, ..., \theta_n)$, given observations $y = (y_1, ..., y_n)$ and known error variance σ_n.

Assumption:

- l_0 ball:

$$\Theta_{n,0,\eta_n} = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq \eta_n n\}$$

  constrains the percentage of nonzero θ_i, η_n be small.

- l_p ball:

$$\Theta_{n,p,\eta_n} = \left\{\theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \leq \eta_n^p n, 0 < p < 2\right\}$$

  constrains the overall magnitude of θ.

- m_p ball:

$$\Theta_{n,p,\eta_n^*} = \left\{\theta \in \mathbb{R}^n : |\theta|_{[k]} \leq \eta_n \left(\frac{n}{k}\right)^{1/p}, 0 < p < 2, k = 1, ..., n\right\}$$

Penalized Estimation. Find $\hat{\theta}$ to minimize

$$K(\theta, y) = \|y - \theta\|_2^2 + Pen\left(\|\theta\|_0\right).$$

Denote $\|\theta\|_0 = k$.

- AIC, Akaike (1973): $Pen(\theta) = 2k$.
- BIC, Schwarz (1978): $Pen(\theta) = k \log n$.
- CIC, Tibshirani and Knight (1999): $Pen(\theta) = 4\sum_{i=1}^{k} \log \frac{n}{i} \approx 4k \log \frac{n}{k}$.
- RIC, George and Foster (1994): $Pen(\theta) = 2k \log n$.
- Foster and Stine (1999): $Pen(\theta) = 2\sum_{i=1}^{k} \log \frac{n}{i}$.
- George and Foster (2000): $Pen(\theta) = 2\sum_{i=1}^{k} \log\left(\frac{n+1}{i} - 1\right)$.

- Bergé and Massart (2001): $Pen(\theta) = 2k \log \frac{n}{k}$.

Note that

$$\begin{aligned} \min_{\theta} K(\theta, y) &= \min_{k} \min_{\theta:\|\theta\|_0 = k} K(\theta, y) \\ &= \min_{k} \left[\sum_{i=k+1}^{n} y_{[i]}^2 + Pen(k) \right] \end{aligned}$$

where $y_{[1]}^2 \geq ... \geq y_{[n]}^2$. Let $\hat{k}$ be the global minimizer of k for the function:

$$S(k) = \sum_{i=k+1}^{n} y_{[i]}^2 + \sum_{i=1}^{k} u_i^2 \tag{2}$$

where $u_1^2 = Pen(1)$ and $u_i^2 = Pen(i) - Pen(i-1)$ for $i > 1$. The global minimizer of $K(\cdot, y)$, is a hard thresholding procedure,

$$\hat{\theta}_i = y_i I \left\{ |y_i| \geq u_{\hat{k}} \right\}.$$

Conjecture 1.2 in ABDJ (2006). Let

$$R(\Theta_{n,p}) \equiv \inf_{\hat{\theta}} \sup_{\theta \in \Theta_{n,p}} E_\theta \left\| \hat{\theta} - \theta \right\|_2^2.$$

Let $Pen(\theta) = 2k \log \frac{n}{k}$ with $k = \|\theta\|_0$. Define

$$\hat{\theta} = \arg\min_{\theta} \left\{ \|y - \theta\|_2^2 + Pen(\theta) \right\},$$

then

$$\sup_{\theta \in \Theta_{n,p}} E \left\| \hat{\theta} - \theta \right\|_2^2 = (1 + o(1)) R_n(\Theta_{n,p}).$$

Remark. We will search the minimizer over $\|\theta\|_0 \leq n/\log n$. Without any constraint, it is easy to see $\hat{\theta} = y$ since $2n \log \frac{n}{n} = 0$.

Main results.

Condition: η_n in l_0 ball, or η_n^p in weak l_p ball and strong l_p, is in $\left[n^{-1} \log^\gamma n, b_2 n^{-b_3}\right]$, $\gamma > 4.5$.

Theorem 1 *Let* $Pen(\theta) = \sum_{i=1}^{\|\theta\|_0} u_i^2$ *with*

$$c \log \frac{n}{i} - (1 - \varepsilon) \log \log \frac{n}{i} \leq u_i^2 \leq c \log \frac{n}{i} + c' \log \log n$$

for some $\varepsilon > 0$, $c \geq 2$ *and any* $c' > 0$. *Define*

$$\hat{\theta} = \arg \min_{\|\theta\|_0 \leq \frac{n}{\log n}} \left[\|y - \theta\|_2^2 + Pen(\theta) \right]$$

then

$$\sup_{\theta \in \Theta_{n,p}} E \left\| \hat{\theta} - \theta \right\|_2^2 = (1 + o(1)) \left(\frac{c}{2} \right)^{1-p/2} R_n(\Theta_{n,p}).$$

Remark: Let $Pen(\theta) = ck \log\left(\frac{n}{k}\right)$, with $c \geq 2$ and $k = \|\theta\|_0$. Define

$$\hat{\theta} = \arg \min_{\|\theta\|_0 \leq \frac{n}{\log n}} \left[\|y - \theta\|_2^2 + Pen(\theta) \right],$$

then

$$\sup_{\theta \in \Theta_{n,p}} E \left\| \hat{\theta} - \theta \right\|_2^2 = (1 + o(1)) \left(\frac{c}{2}\right)^{1-p/2} R_n(\Theta_{n,p}).$$

Proof of main results:
A brief outline for the upper bound

Let

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

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

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

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

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

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

We will show

$$\sup_{\Theta_{n,p}} K(\theta_0, \theta) \leq (1 + o(1)) c^* R_n(\Theta_{n,p})$$

and

$$\sup_{\Theta_{n,p}} \mathbb{E}_{y|\theta} \left\langle \hat{\theta}^P - \theta_0, z \right\rangle = o(1) R_n(\Theta_{n,p}).$$

Step 1.
We prove

$$\sup_{\theta} K(\theta_0, \theta) = \sup_{\theta \in \Theta_{n,p}} \inf_{\mu} \left[\|\theta - \mu\|^2 + Pen(\mu) \right] \leq (1 + o(1)) c^* R(\Theta_{n,p})$$

In l_0 ball, it is easy to see

$$\sup_{\theta} K(\theta_0, \theta) = \sup_{\theta} \inf_{k} \left[\sum_{i=k+1}^{n} \theta_{[i]}^2 + \sum_{i=1}^{k} u_i^2 \right] \leq \sum_{i=1}^{k_n} u_i^2.$$

where $k_n = n\eta_n \le b_2 n^{1-b_3}$. So

$$\sum_{i=1}^{k_n} u_i^2 \sim c \sum_{i=1}^{k_n} \log \frac{n}{i} \sim ck_n \log \frac{n}{k_n} = cn\eta_n \log \left(\eta_n^{-1}\right) \sim \frac{c}{2} R\left(\Theta_{n,p}\right).$$

In l_p ball, recall

$$\sup_{\theta \in \Theta_{n,p}} K\left(\theta_0, \theta\right) = \sup_{\theta \in \Theta_{n,p}} \inf_k \left[\sum_{i=k+1}^{n} \theta_{[i]}^2 + \sum_{i=1}^{k} u_i^2\right]$$

We will see later that the left hand size is equal to $\sup_{\theta \in \Theta_{n,p}} \sum_{i=1}^n \left[\theta_{[i]}^2 \wedge u_i^2\right]$. Let $x_i = |\theta|_{[i]}^p$, by the definition of the strong l_p ball $\Theta_{n,p}$,

$$\sup_{\theta \in \Theta_{n,p}} K\left(\theta_0, \theta\right) = \sup_{S_1} \sum_{i=1}^{n} \left[x_i^{2/p} \wedge u_i^2\right] = \sup_{S_2} \sum_{i=1}^{n} x_i^{2/p}$$

where set $S_1 = \{\mathbf{x} : \sum_{i=1}^n x_i \le n\eta_n^p,\ x_1 \ge ... \ge x_n\}$ is a convex set so that the set

$$S_2 = \left\{\mathbf{x} : \sum x_i \le n\eta_n^p,\ x_1 \ge ... \ge x_n,\ 0 \le x_i \le u_i^p,\ i = 1, ..., n\right\}$$

is a convex subset. In S_2, $\sum_{i=1}^n x_i^{2/p}$ is a convex function under sparsity condition $p < 2$. So the maximizer $\mathbf{x}^*$ locates at the extreme points of S_2. That is, $\mathbf{x}^* = (s_1^p, ..., s_k^p, 0...0)$, so that $\left(\theta_1^{*2}, ..., \theta_n^{*2}\right) = \left(s_1^2, ..., s_k^2, 0, ..., 0\right)$ for some k, under the constrain of strong l_p ball that limits the total mass of the mean: $\sum_{i=1}^n |\theta^*|_{[i]}^p \le n\eta_n^p$. To get the supremum, solve equation $n\eta_n^p = \sum_{i=1}^k u_i^p$ for k. Assume k' is the solution. So

$$\begin{aligned}\sup_{\theta \in \Theta_{n,p}} \left\{K\left(\theta_0, \theta\right)\right\} &= \sum_{i=1}^{k'} u_i^2 \sim k' u_{k'}^2 \sim n\eta_n^p \left(c \log \frac{n}{k'}\right)^{1-p/2} \\ &\sim n\eta_n^p \left(c \log \eta_n^{-p}\right)^{1-p/2} \sim \left(\frac{c}{2}\right)^{1-p/2} R_n\left(\Theta_{n,p}\right).\end{aligned}$$

Step 2.

Let q_nShow that

$$\sup_{\theta \in \Theta_{n,p}} \mathbb{E}_{y|\theta} \left\langle \hat{\theta}^P - \theta, z \right\rangle \le o(1) R\left(\Theta_{n,p}\right).$$

Write

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

It is easy to observe that

$$z_i \left[\eta_H\left(y_i, u_2\right) - \theta_i\right] \le z_i \left[\eta_H\left(y_i, u_1\right) - \theta_i\right]$$

if $|\theta_i| \le u_1 \le u_2$. This inspires to define a quantity $u_{k\_}(\theta)$ such that $u_{\hat{k}^P} \ge u_{k\_}(\theta)$ with high probability and

$$\mathbb{E}\sum_{i=1}^{n} z_i \left[\eta_H\left(y_i, u_{k\_}\right) - \theta_i\right] \le o(1)\, R(\Theta_{n,p})$$

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

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

It can be shown the dominating term is

$$T_0 = \sum_{i\in S_n(\mu)} \mathrm{Cov}\left(y_i, \eta_H\left(y_i, u_{k\_}\right)\right) \sim 2n u_{k\_}\, \phi\left(u_{k\_}\right) = o(1)\, R(\Theta_{n,p}),$$

and

$$T_1 = o(1)\, R(\Theta_{n,p}).$$

Remark: *Lower bound.* To prove the lower bound that $\sup_\theta \mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \ge c^* R_n(\Theta_{n,p}(\eta_n))$, we find a specific $\theta \in \Theta_{n,p}(\eta_n)$ such that $\mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \sim c^* R_n(\Theta_{n,p}(\eta_n))$. Let $\varepsilon_n = 1/\log\log n$. Let $k^* = \lfloor \eta_n n \rfloor$ and define

$$\theta_k = \begin{cases} \sqrt{c(1-\varepsilon_n)\log \frac{n}{k^*}}, & k \le k^* - 1 \\ 0, & k \ge k^* \end{cases}.$$

It is easy to see $\theta = (\theta_k)_{1\le k\le n}$ is contained in $l_0[\eta_n]$ ball. Similarly for $l_p[\eta_n]$ ball we define

$$\theta_k = \begin{cases} \sqrt{c(1-\varepsilon_n)\log \frac{n}{k^*}}, & k \le k^* - 1 \\ 0, & k \ge k^* \end{cases}$$

where $k^* = n\eta_n^p \left(c\log \eta_n^{-p}\right)^{-p/2}$. Then

$$P\left(\hat{k} = o(k^*)\right) \to 1$$

which implies

$$\mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 = c^*(1 + o(1))\, R_n(\Theta_{n,p}(\eta_n)).$$