STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 10 (April 26): In this lecture we will continue with the results of Greenshtein and Ritov (2004).

Greenshtein and Ritov (2004)

See also Greenshtein (2006). Suppose that we observe

$$Z_i \equiv (Y^i, \underline{X}^i) = (Y^i, X_1^i, \ldots, X_p^i), \qquad i = 1, \ldots, n$$

where Z_i are i.i.d. $P_n \in \mathcal{P}$. We are interested in this triangular array setting with $p = p_n = n^\alpha$ for some $\alpha > 1$. Furthermore we want to "predict"' Y by predictors of the form $\sum_{j=1}^p \beta_j X_j$ where $\beta = (\beta_1, \ldots, \beta_p)' \in B_n \subset \mathbb{R}^p$ for each n.

Natural sets B_n to consider are of the form

$$\begin{aligned} B_{n,k} &\equiv \{\beta \in \mathbb{R}^p : \ \ \#\{j : \ \beta_j \neq 0\} = k\}, \qquad \text{and} \\ B_{n,b} &\equiv \{\beta \in \mathbb{R}^p : \ \ \|\beta\|_1 \leq b\} \end{aligned}$$

where $k = k_n \to \infty$ and $b = b_n \to \infty$.

Suppose that $Z = (Y, \underline{X}) \sim P$ on $(\mathbb{R}^{p+1}, \mathcal{B}_{p+1})$, and define

$$L_P(\beta) = E_P(Y - \sum_{j=1}^p \beta_j X_j)^2.$$

For $P_n \in \mathcal{P}$ and $B_n \subset \mathbb{R}^p$ given, define

$$\beta^*(P_n) \equiv \beta_n^* \equiv \text{argmin}_{\beta \in B_n} L_{P_n}(\beta);$$

Thus β_n^* is a deterministic sequence in $\mathbb{R}^p$ determined by P_n and B_n.

Definition 1. Given a set of possible predictors B_n, a sequence of procedures $\{\hat{\beta}_n\}$ is *persistent* (or persistent relative to $\{B_n\}$ and $\{\mathcal{P}_n\}$) if, for every sequence $P_n \in \mathcal{P}_n$

$$L_{P_n}(\widehat{\beta}_n) - L_{P_n}(\beta^*(P_n)) \to_p 0.$$

Let $\gamma' = (-1, \beta_1, \ldots, \beta_p)' \equiv (\beta_0, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$, and let $Y \equiv X_0$. Thus

$$L_P(\beta) = E_P(Y - \underline{X}'\beta)^2 = \gamma' \Sigma_P \gamma$$

where

$$\Sigma_P \equiv (\sigma_{ij}) = (E_P(X_i X_j))_{0 \leq i,j \leq p}.$$

Let $\mathbb{P}_n$ be the empirical measure of $Z_1, \ldots, Z_n$. Then

$$L_{\mathbb{P}_n}(\beta) = \gamma' \Sigma_{\mathbb{P}_n} \gamma = \gamma'(\hat{\sigma}_{ij})\gamma \equiv \gamma' \widehat{\Sigma} \gamma.$$

Define ϵ_{ij}^n by

$$\hat{\sigma}_{ij} = \sigma_{ij} + \epsilon_{ij}^n,$$

and write

$$\widehat{\Sigma} = \Sigma_P + E, \qquad \text{so} \qquad E = (\epsilon_{ij}^n).$$

Condition 1. Suppose that the random variables $Y_{ij} \equiv X_i X_j$, $0 \le i, j \le p$ satisfy $Var_P(Y_{ij}) \le C$ for all $P \in \mathcal{P}_n$ and all i, j. Moreover, assume that $\phi_{ij}(t) = E_P \exp(tY_{ij})$ exist for t in a neighborhood of 0 and $\sup_{|t| \le \epsilon} |\phi_{ij}^{(3)}(t)| \le C_2$ for all $P \in \mathcal{P}_n$ and all i, j for some small $\epsilon > 0$.

Lemma 0. Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. with $E\xi_1 = 0$, $E\xi_1^2 = 1$. Suppose that $\varphi(t) \equiv E \exp(t\xi_1) < \infty$ for all $0 \le t \le t_0$ with $t_0 > 0$. Then there exists a constant c depending only on the distribution of ξ_1 such that for all $x > 0$ and all $n \ge 1$, $S_n = \xi_1 + \cdots + \xi_n$ satisfies

$$P(S_n > x) \le \begin{cases} \exp\left(-\frac{x^2}{4n}\right), & \text{if } x < cn, \\ \exp\left(-cx/4\right), & \text{if } x \ge cn. \end{cases}$$

Proof. (Breiman and Freedman, 1983). Let $\varphi(t) = E \exp(t\xi_1)$. Then by Markov's inequality

$$P(S_n > x) \le \exp(-tx)\varphi(t)^n \qquad \text{for each} \quad 0 \le t \le t_0.$$

Now since ξ_1 has $E(\xi_1) = 0$ and $Var(\xi_1) = 1$,

$$\varphi(t) = E \exp(t\xi_1) = 1 + \frac{1}{2}t^2 + o(t^2),$$

Thus $\varphi(t) \le \exp(t^2)$ for $0 \le t \le t_*$ for $t_* \le t_0$ sufficiently small, and we can bound $P(S_n > x)$ by $\exp(-tx + nt^2)$ for any $0 \le t \le t_*$. Now choose $c \equiv 2t_*$. To prove the first inequality, take $t = x/2n$; to prove the second inequality take $t = c/2 = t_0$, and note that $t_0 \le x/(2n)$. □

Lemma 1. If condition 1 holds, then

$$\inf_{P \in \mathcal{P}_n} Pr_{P_n}\left(-\sqrt{\frac{A \log n}{n}} \le \epsilon_{ij}^n \le \sqrt{\frac{A \log n}{n}} \ \text{ for all } \ 0 \le i, j \le n \right) \to 1.$$

Proof. First we use Lemma 0 to bound

$$\begin{aligned} Pr_{P_n}\left(|\epsilon_{jk}^n| > \sqrt{KCn^{-1}\log n}\right) &\leq 2\exp(-Kn\log n/(4n)) \qquad \text{if } \sqrt{Kn\log n} \leq cn \\ &= 2n^{-K/4}. \end{aligned}$$

Thus

$$\begin{aligned} &\sup_{P\in\mathcal{P}_n} Pr_{P_n}(|\epsilon_{jk}^n| > \sqrt{KCn^{-1}\log n} \text{ for some } 0 \leq j,k \leq p) \\ &\leq \sum_{0\leq j,k\leq p} \sup_{P\in\mathcal{P}_n} Pr_{P_n}(|\epsilon_{jk}^n| > \sqrt{KCn^{-1}\log n}) \\ &\leq 2(p_n+1)^2 n^{-K/4} \\ &\leq 8n^{2\alpha}n^{-K/4} = 8n^{-(K/4-2\alpha)} \leq p_n^{-2} \end{aligned}$$

if $K > 16\alpha$. □

Lemma 2. If condition 1 holds, then

$$\inf_{P\in\mathcal{P}_n} Pr_{P_n}\left(L_{P_n}(\beta) \leq \gamma'\Sigma_{\mathbb{P}_n}\gamma + |\gamma|'\hat{E}|\gamma| \text{ for all } \beta\in\mathbb{R}^p\right) \to 1 \tag{1}$$

where $\hat{E} \equiv J\sqrt{An^{-1}\log n}$ and $|\gamma| = (1, |\beta_1|, \ldots, |\beta_p|)$.

Proof of Lemma 2.

$$\begin{aligned} &\sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(L_{P_n}(\beta) > \gamma'\Sigma_{\mathbb{P}_n}\gamma + |\gamma|'\hat{E}|\gamma| \text{ for some } \beta\in\mathbb{R}^p\right) \\ &= \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(\gamma'\Sigma_{P_n}\gamma - \gamma'\Sigma_{\mathbb{P}_n}\gamma > |\gamma|'\hat{E}|\gamma| \text{ for some } \gamma=(-1,\beta)\in\mathbb{R}^{p+1}\right) \\ &= \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(\gamma'(\Sigma_{P_n} - \Sigma_{\mathbb{P}_n})\gamma > |\gamma|'\hat{E}|\gamma| \text{ for some } \gamma=(-1,\beta)\in\mathbb{R}^{p+1}\right) \\ &\leq \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(|\gamma'(\Sigma_{P_n} - \Sigma_{\mathbb{P}_n})\gamma| > |\gamma|'\hat{E}|\gamma| \text{ for some } \gamma=(-1,\beta)\in\mathbb{R}^{p+1}\right) \\ &\leq \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(|\gamma'|(|\hat{\sigma}_{jk} - \sigma_{jk}|)|\gamma| > |\gamma|'\hat{E}|\gamma| \text{ for some } \gamma=(-1,\beta)\in\mathbb{R}^{p+1}\right) \\ &\leq \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(\max_{1\leq j,k\leq p}|\hat{\sigma}_{jk} - \sigma_{jk}|)| > \sqrt{n^{-1}A\log n}\right) \\ &\to 0 \end{aligned}$$

by Lemma 1. □

Theorem 1. If condition 1 holds, then for any $B_{n,b_n} \subset \mathbb{R}^p$ with $b_n = o((n/\log n)^{1/4})$, there exists a persistent sequence of procedures $\hat{\beta}_n$. In particular,

$$\hat{\beta}_n \equiv \text{argmin}_{\beta:\ \|\beta\|_1\leq b_n} L_{\mathbb{P}_n}(\beta) \tag{2}$$

is persistent.

Proof. As in the proof of Lemma 2,

$$\begin{aligned}
&\sup_{P_n\in\mathcal{P}_n}\sup_{\beta\in B_{n,b_n}} Pr_{P_n}\left(|L_{P_n}(\beta)-L_{\mathbb{P}_n}(\beta)|>|\gamma|'\hat{E}|\gamma|\right)\\
&\quad\leq \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(|L_{P_n}(\beta)-L_{\mathbb{P}_n}(\beta)|>|\gamma|'\hat{E}|\gamma| \text{ for some } \gamma\right)\\
&\quad\leq \sup_{P_n\in\mathcal{P}_n} Pr_{P_n}\left(|\gamma'(\Sigma_{P_n}-\Sigma_{\mathbb{P}_n})\gamma|>|\gamma|'\hat{E}|\gamma| \text{ for some } \gamma\right)\\
&\quad\to 0.
\end{aligned}$$

Since

$$|\gamma|'\hat{E}|\gamma| = \sqrt{An^{-1}\log n}|\gamma|'\underline{11}'|\gamma| = \sqrt{An^{-1}\log n}\|\gamma\|_1^2,$$

this implies that

$$\begin{aligned}
&\inf_{P_n\in\mathcal{P}_n}\inf_{\beta\in B_{n,b_n}} Pr_{P_n}\left(|L_{P_n}(\beta)-L_{\mathbb{P}_n}(\beta)|\leq|\gamma|'\hat{E}|\gamma|\right)\\
&\quad= \inf_{P_n\in\mathcal{P}_n}\inf_{\beta\in B_{n,b_n}} Pr_{P_n}\left(|L_{P_n}(\beta)-L_{\mathbb{P}_n}(\beta)|\leq\sqrt{An^{-1}\log n}\|\gamma\|_1^2\right)\to 1.
\end{aligned}$$

But for sequences of vectors of order $b_n = o((n/\log n)^{1/4})$, the sequence $|\gamma|'\hat{E}|\gamma| = \sqrt{An^{-1}\log n}\|\gamma\|_1^2$ converges to 0. The result follows immediately from the definition of persistence.

Before we show that $\widehat{\beta}_n$ given in (2) is persistent, we need an inequality.

Lemma 3. (Nemirovski's inequality). Suppose that $V_1,\ldots,V_n$ are independent random vectors in $\mathbb{R}^m$, $m\geq 3$, with $E(V_i)=0$ and $E\|V_i\|_2^2 = EV_i'V_i<\infty$ for $i=1,\ldots,n$. Then for every $r\in[2,\infty]$,

$$E\Big\|\sum_{i=1}^n V_i\Big\|_r^2 \leq C\min\{r,\log(m)\}\sum_{i=1}^n E\|V_i\|_r^2$$

where $\|\cdot\|_r$ is the ℓ_r-norm, $\|x\| = \{\sum_{i=1}^m |x_i|^r\}^{1/r}$ for $x\in\mathbb{R}^m$, and C is an absolute constant.

Now we continue with the proof of Theorem 1. It remains to show that $\widehat{\beta}_n$ given by (2), namely the (constrained version of the) lasso estimator with $b_n = o((n/\log n)^{1/4})$, is persistent.

Consider the matrix $\Sigma_{\mathbb{P}_n}-\Sigma_P$ as a $(p+1)^2-$dimensional vector, and write

$$\Sigma_{\mathbb{P}_n}-\Sigma_P = \sum_{i=1}^n V_i = \sum_{i=1}^n \frac{1}{n}\left(X_0^iX_0^i - E(X_0^iX_0^i), X_0^iX_1^i - E(X_0^iX_1^i),\ldots,\right).$$

Suppose that

$$\max_{0\leq j,k\leq p}|X_j^iX_k^i - E(X_j^iX_k^i)| \equiv F(Z_i)$$

satisfies

$$E_{P_n} F(Z_i)^2 \leq M < \infty.$$

Then, by Nemirovski's inequality with $r = \infty$,

$$\begin{aligned}
\left\{ E \Big\| \sum_{i=1}^n V_i \Big\|_\infty \right\}^2 &\leq E \Big\| \sum_{i=1}^n V_i \Big\|_\infty^2 \leq C \log((p_n+1)^2) \sum_{i=1}^n E\|V_i\|_\infty^2 \\
&\leq C' \log(4n^{2\alpha}) \frac{1}{n^2} \sum_{i=1}^n E F(Z_i)^2 \\
&\leq C' M \frac{(2\alpha)\log n + \log 4}{n}, \qquad (3)
\end{aligned}$$

and hence

$$E\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = E \Big\| \sum_{i=1}^n V_i \Big\|_\infty \leq \sqrt{\frac{C'' \log n}{n}}. \qquad (4)$$

Remark. Alternatively, this last inequality follows almost immediately from Theorem 2.14.2, van der Vaart and Wellner (1996), page 240:

$$E\|\mathbb{G}_n\|_{\mathcal{F}}^* \lesssim J_{[]}(1, \mathcal{F}, L_2(P)) \|F\|_{P,2}$$

where $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ and

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) \equiv \int_0^\delta \sqrt{\log(1 + \log N_{[]}(\epsilon, \mathcal{F}, L_2(P)))} d\epsilon.$$

In this application $\mathcal{F} = \{f_{j,k}(z) = x_j x_k,\ 0 \leq j,k \leq p\}$ is a finite list of functions of cardinality $\#(\mathcal{F}) = (p_n+1)^2$, and hence $N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq (p_n+1)^2$ by simply choosing $\epsilon-$ brackets $[l_{jk}, u_{jk}]$ given by $u_{jk}(z) = f_{jk}(z) + \epsilon/2$, $l_{jk}(z) = f_{jk}(z) - \epsilon/2$. Thus the bound becomes

$$E_{P_n}\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \sqrt{1 + \log[(p_n+1)^2]} \|F\|_{P_n,2} \lesssim \sqrt{\log n},$$

or, equivalently

$$E\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = E\|\mathbb{P}_n - P\|_{\mathcal{F}} \lesssim \sqrt{n^{-1} \log n},$$

in agreement with the bound given by Nemirovski's inequality.

Now note that

$$|L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| = \gamma'(\Sigma_{\mathbb{P}_n} - \Sigma_{P_n})\gamma \leq \|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty \|\gamma\|_1^2,$$

so for $B_{n,b_n} = \{\beta \in \mathbb{R}^p : \ \|\beta\| \leq b_n\}$ with $b_n = o((n/\log n)^{1/4})$, by Markov's inequality followed by (4)

$$\begin{aligned}
Pr\left(\sup_{\beta \in B_{n,k_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| > \epsilon \right) &\leq Pr(\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty b_n^2 > \epsilon) \\
&\leq \epsilon^{-1} b_n^2 E\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty \\
&= \epsilon^{-1} b_n^2 \sqrt{\frac{C'' \log n}{n}} = o(1). \qquad (5)
\end{aligned}$$

Define $\hat{\beta}_n$ by

$$\hat{\beta}_n \equiv \text{argmin}_{\beta \in B_{n,b_n}} L_{\mathbb{P}_n}(\beta).$$

Then, since

$$L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \geq 0,$$
$$L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \leq 0,$$

it follows that

$$\begin{aligned} 0 &\leq L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \\ &= L_{P_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\hat{\beta}_n) + L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \\ &\qquad + L_{\mathbb{P}_n}(\beta_n^*) - L_{P_n}(\beta_n^*) \\ &\leq 2 \sup_{\beta \in B_{n,b_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| \\ &\to_p 0 \end{aligned}$$

by (5).

Condition 2. Let B_{n,k_n} be the set of all vectors $\beta \in \mathbb{R}^p$ with $k_n = o((n/\log n)^{1/2})$ non-zero entries. Suppose there is a constant $C < \infty$ such that $\beta_n^* \equiv \text{argmin}_{\beta \in B_{n,k_n}} L_{P_n}(\beta)$ satisfies $\|\beta_n^*\|_2 \leq C$ for all $P_n \in \mathcal{P}_n$.

Lemma 4. If $E_{P_n} Y^2 \leq M < \infty$ for all $P_n \in \mathcal{P}_n$, condition 2 holds if the minimal eigenvalue $\lambda_1 \equiv \lambda_{1,n}$ of $Cov_{P_n}(X)$ satisfies $\lambda_{1,n} \geq \delta > 0$.

Theorem 2. Suppose conditions 1 and 2 hold. There exists a persistent sequence of procedures with respect $\{B_{n,k_n}\}$ with $k_n = o((n/\log n)^{1/2})$.

Proof. Consider the same procedures as defined in Theorem 1. By condition 2 we can restrict attention to $\beta \in \mathbb{R}^p$ with $\|\beta\|_2 \leq C < \infty$. But for $\beta \in \mathbb{R}^p$ with $\|\beta\|_2 \leq C$ and fewer than k_n non-zero entries,

$$\begin{aligned} \|\beta\|_1 &= \sum_{j=1}^p |\beta_j| = \sum_{j=1}^p |\beta_j| 1\{\beta_j \neq 0\} \\ &\leq \sqrt{\sum_{j=1}^p |\beta_j|^2} \sqrt{\sum_1^p 1\{\beta_j \neq 0\}} \qquad \text{by Cauchy-Schwarz} \\ &\leq C\sqrt{k_n} \equiv b_n. \end{aligned}$$

Thus the results holds by virtue of Theorem 1 since under condition 2 $B_{n,k_n} \subset B_{n,b_n}$ with $b_n = o((n/\log n)^{1/4})$. □