STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 9 (April 24): In this lecture we will consider several proofs of the results for the "adaptive lasso" as discussed in the paper by Zou (2006) before moving on to the paper by Greenshtein and Ritov (2004).

Zou (2006) gives a necessary condition for model selection consistency of the lasso as follows:

Theorem 1. (Necessary condition for model selection consistency of lasso). Suppose that $P(\mathcal{A}_n = \mathcal{A}) \to 1$. Then there is a sign vector $s = (s_1, \ldots, s_{p_0})'$, $s_j = \pm 1$, such that

$$|M_{21}M_{11}^{-1}s| \le 1 \qquad \text{where} \qquad M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \tag{1}$$

Proof. First suppose that $\lambda_n/\sqrt{n} \to \lambda_0$ and $\lambda_n/n \to 0$. Then the lasso is selection inconsistent by Proposition 1 (via Knight and Fu (2000)). This contradicts the assumed selection consistency, and hence this case does not occur. This leaves three cases: either
(i) $n^{-1}\lambda_n \to \infty$; or
(ii) $n^{-1}\lambda_n \to \lambda_0 \in (0, \infty)$; or
(iii) $n^{-1}\lambda_n \to 0$, but $n^{-1/2}\lambda_n \to \infty$.

If (i) occurs, then it is easily checked that $\hat\beta_n^{(1)} \to_p 0$, which contradicts selection consistency. Hence this case can not occur.

Suppose that (ii) occurs. By Theorem 1 of Knight and Fu (2000) (Lemma 1 of Zou (2006)),

$$\hat\beta_n^{(1)} \to_p \beta_* \equiv \text{argmin}_u V_1(u)$$

where $V_1(u) = (u - \beta^*)'M(u - \beta^*) + \lambda_0\|u\|_1$. Since $P(\mathcal{A}_n = \mathcal{A}) \to 1$, $\beta_{*j} = 0$ for all $j \notin \mathcal{A}$. Fix $j \in \mathcal{A}$ and consider $\{j \in \mathcal{A}_n\}$. By the Karush-Kuhn-Tucker conditions,

$$-2\mathbf{x}_j'(Y - \mathbf{X}\hat\beta_n^{(1)}) + \lambda_n \text{sign}(\hat\beta_n^{(1)}) = 0.$$

Hence

$$P(j \in \mathcal{A}_n) \le P(|-2\mathbf{x}_j' n^{-1}(Y - X\hat\beta_n^{(1)})| = n^{-1}\lambda_n)$$

where

$$\begin{aligned} -2\mathbf{x}_j' n^{-1}(Y - X\hat\beta_n^{(1)}) &= -2\mathbf{x}_j' n^{-1} X(\beta^* - \hat\beta_n^{(1)}) - 2n^{-1}\mathbf{x}x_{j\underline{\in}}' \\ &\to_p -2(M(\beta^* - \beta_*))_j. \end{aligned}$$

Thus $P(j \in \mathcal{A}_n) \to 1$ implies that

$$2(M(\beta^* - \beta_*))_j = \lambda_0.$$

Similarly, fix $j' \notin \mathcal{A}$. Then $P(j' \notin \mathcal{A}_n) \to 1$. Consider $\{j' \notin \mathcal{A}_n\}$; by the KKT conditions

$$| - 2\mathbf{x}_j'(Y - \mathbf{X}\hat{\beta}_n^{(1)})| \leq \lambda_n.$$

Thus

$$P(j' \notin \mathcal{A}_n) \leq P(| - 2\mathbf{x}_j' n^{-1}(Y - X\hat{\beta}_n^{(1)})| \leq n^{-1}\lambda_n),$$

and this yields

$$|2(M(\beta^* - \beta_*))_{j'}| \leq \lambda_0.$$

Note that

$$M(\beta^* - \beta_*) = \begin{pmatrix} M_{11}(\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}}) \\ M_{21}(\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}}) \end{pmatrix}.$$

Thus we find that

$$M_{11}(\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}}) = \frac{\lambda_0}{2} s_*, \qquad \text{and}$$
$$|M_{21}(\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}})| \leq \frac{\lambda_0}{2}$$

where $s_* = \text{sign}(M_{11}(\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}}))$. Solving the first of these relations for $\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}}$ yields

$$\beta^*_{\mathcal{A}} - \beta_{*\mathcal{A}} = \frac{\lambda_0}{2} M_{11}^{-1} s_*,$$

and plugging this back into the second relation gives

$$|M_{21}(\frac{\lambda_0}{2} M_{11}^{-1} s_*)| \leq \frac{\lambda_0}{2},$$

or

$$|M_{21} M_{11} s_*| \leq 1.$$

This proves the claim in case (ii).

As discussed in Lecture 8 on 19 April, Zou (2006) claims that the following oracle inequality holds:

Theorem 3. Let $\lambda_n \equiv (2 \log n)^{(1+\gamma)/2}$. Then

$$R(\mu, \hat{\mu}^{(\hat{w})}) \leq \left(2 \log n + 5 + 4\gamma^{-1}\right) \left(R(\mu, \hat{\mu}(\text{ideal})) + \frac{1}{2\sqrt{\pi}} \frac{1}{\sqrt{\log n}} \right).$$

I claim that this should be revised to the following:

Theorem 3 revised. Let $\lambda_n \equiv (2 \log n)^{(1+\gamma)/2}$. Then

$$R(\mu, \hat{\mu}^{(\hat{w})}) \leq (2 \log n + 5 + 4\gamma) \left(R(\mu, \hat{\mu}(\text{ideal})) + \frac{1}{2\sqrt{\pi}} \frac{1}{\sqrt{\log n}} \right).$$

Proof. We will first show that

$$E(\hat{\mu}_i^{(\hat{w})} - \mu_i)^2 \leq \left(\lambda^{2/(1+\gamma)} + 5 + 4\gamma\right)\left(\min\{\mu_i^2, 1\} + \frac{\phi(\lambda^{1/(1+\gamma)})}{\lambda^{1/(1+\gamma)}}\right) \tag{2}$$

where $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. Then from (2) it follows that

$$R(\mu, \hat{\mu}^{(\hat{w})}) \leq (\lambda^{2/(1+\gamma)} + 5 + 4\gamma)\left(R(\text{ideal}) + n\frac{\phi(\lambda^{1/(1+\gamma)})}{\lambda^{1/(1+\gamma)}}\right).$$

Note that when $\lambda = (2\log n)^{(1+\gamma)/2}$, $\lambda^{2/(1+\gamma)} = 2\log n$, and hence

$$n\frac{\phi(\lambda^{1/(1+\gamma)})}{\lambda^{1/(1+\gamma)}} = n\frac{1}{\sqrt{2\pi}}\frac{e^{-\log n}}{\sqrt{2\log n}} = \frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{\log n}}.$$

Hence the theorem is proved if we show that (2) holds. To do this, we decompose the expectation on the left side as

$$\begin{aligned} E(\hat{\mu}_i^{(\hat{w})} - \mu_i)^2 &= E(\hat{\mu}_i^{(\hat{w})} - Y_i)^2 + E(Y_i - \mu_i)^2 \\ &\quad + 2E(\hat{\mu}^{(\hat{w})}(Y_i - \mu_i)) - 2E(Y_i(Y_i - \mu_i)) \\ &= E(\hat{\mu}_i^{(\hat{w})} - Y_i)^2 + 1 + Eg'(\hat{\mu}_i) - 2 \end{aligned} \tag{3}$$

by Stein's lemma (see Lemma 5.7.2, Stat 580's notes, chapter 5 page 26, or Stein (1981). with $g(Y_i) = \hat{\mu}_i^{(\hat{w})} = \hat{\mu}_i^{(\hat{w})}(Y_i)$. Now

$$\begin{aligned} g(Y_i) &\equiv \hat{\mu}_i^{(\hat{w})}(Y_i) = (|Y_i| - \lambda/|Y_i|^\gamma)_+\text{sign}(Y_i) \\ &= \begin{cases} (|Y_i| - \lambda/|Y_i|^\gamma)\text{sign}(Y_i), & \text{if } |Y_i| \geq \lambda/|Y_i|^\gamma \\ 0, & \text{if } |Y_i| \leq \lambda/|Y_i|^\gamma \end{cases} \\ &= \begin{cases} (|Y_i| - \lambda/|Y_i|^\gamma)\text{sign}(Y_i), & \text{if } |Y_i|^{1+\gamma} \geq \lambda \\ 0, & \text{if } |Y_i|^{1+\gamma} \leq \lambda \end{cases} \\ &= \begin{cases} (|Y_i| - \lambda/|Y_i|^\gamma)\text{sign}(Y_i), & \text{if } |Y_i| \geq \lambda^{1/(1+\gamma)} \\ 0, & \text{if } |Y_i| \leq \lambda^{1/(1+\gamma)}. \end{cases} \end{aligned}$$

Therefore

$$\begin{aligned} g'(Y_i) &= \begin{cases} 1 + \lambda\gamma|Y_i|^{-\gamma-1}, & \text{if } |Y_i| \geq \lambda^{1/(1+\gamma)} \\ 0, & \text{if } |Y_i| \leq \lambda^{1/(1+\gamma)}. \end{cases} \\ &\leq 1 + \gamma. \end{aligned}$$

Plugging this into (3) yields

$$\begin{aligned} E(\hat{\mu}_i^{(\hat{w})} - \mu_i)^2 &= EY_i^2 1\{|Y_i| \leq \lambda^{1/(1+\gamma)}\} + E\left(\frac{\lambda^2}{|Y_i|^{2\gamma}} 1\{|Y_i| \geq \lambda^{1/(1+\gamma)}\}\right) \qquad (4) \\ &\quad + 2E\left\{\left(1 + \lambda\gamma|Y_i|^{-\gamma-1}\right)1\{|Y_i| \geq \lambda^{1/(1+\gamma)}\}\right\} - 1 \\ &\leq \lambda^{2/(1+\gamma)} P(|Y_i| \leq \lambda^{1/(1+\gamma)}) + \lambda^{2/(1+\gamma)} P(|Y_i| > \lambda^{1/(1+\gamma)}) \\ &\quad + 2(1+\gamma)P(|Y_i| > \lambda^{1/(1+\gamma)}) - 1 \\ &\leq \lambda^{2/(1+\gamma)} + 2(1+\gamma) - 1 = \lambda^{2/(1+\gamma)} + 1 + 2\gamma. \qquad (5) \end{aligned}$$

On the other hand, starting again from (5) again and decomposing slightly differently yields

$$\begin{aligned}
E(\hat{\mu}_i^{(\hat{w})} - \mu_i)^2 &= EY_i^2 1\{|Y_i| \le \lambda^{1/(1+\gamma)}\} + E\left(\frac{\lambda^2}{|Y_i|^{2\gamma}} 1\{|Y_i| \ge \lambda^{1/(1+\gamma)}\}\right) \\
&\quad + 2E\left\{\left(1 + \lambda\gamma |Y_i|^{-\gamma-1}\right) 1\{|Y_i| \ge \lambda^{1/(1+\gamma)}\}\right\} - 1 \\
&= E(Y_i^2) - 1 + E\left\{\frac{\lambda^2}{|Y_i|^{2\gamma}} + 2\left(1 + \frac{\lambda\gamma}{|Y_i|^{1+\gamma}}\right) - Y_i^2\right\} 1\{|Y_i| > \lambda^{1/(1+\gamma)}\} \\
&= \mu_i^2 + E\left\{\frac{\lambda^2}{|Y_i|^{2\gamma}} + 2\left(1 + \frac{\lambda\gamma}{|Y_i|^{1+\gamma}}\right) - Y_i^2\right\} 1\{|Y_i| > \lambda^{1/(1+\gamma)}\} \\
&\le \mu_i^2 + E\left\{2\left(1 + \frac{\lambda\gamma}{|Y_i|^{1+\gamma}}\right)\right\} 1\{|Y_i| > \lambda^{1/(1+\gamma)}\} \\
&\le \mu_i^2 + 2(1+\gamma) P(|Y_i| > \lambda^{1/(1+\gamma)}). \qquad (6)
\end{aligned}$$

Now

$$\begin{aligned}
P(|Y_i| > t) &\equiv g_t(\mu_i) = g_t(0) + g_t'(0)\mu_i + (1/2) g_t''(\xi^*)\mu_i^2 \le g_t(0) + (1/2)\|g_t''\|_\infty \mu_i^2 \\
&\le P_0(|Y_i| > t) + 2\mu_i^2, \qquad \text{since } \|g''\|_\infty \le 1/2 < 4 \\
&\le 2t^{-1}\phi(t) + 2\mu_i^2, \qquad \text{by Mills' ratio inequality, } 1 - \Phi(x) \le x^{-1}\phi(x).
\end{aligned}$$

Here

$$\begin{aligned}
g_t(\mu) &= P_\mu(|Y| > t) = P_\mu(Y > t) + P_\mu(-Y > t) \\
&= P_\mu(Y - \mu > t - \mu) + P_\mu(-(Y-\mu) > t + \mu) \\
&= 1 - \Phi(t - \mu) + 1 - \Phi(t + \mu),
\end{aligned}$$

so that

$$\begin{aligned}
&g_t'(\mu) = \phi(t-\mu) - \phi(t+\mu) \overset{\mu=0}{=} \phi(t) - \phi(t) = 0, \qquad \text{and} \\
&g_t''(\mu) = -\phi'(t-\mu) - \phi'(t+\mu) = (t-\mu)\phi(t-\mu) + (t+\mu)\phi(t+\mu), \qquad \text{so} \\
&\|g_t''\|_\infty \le 2 \sup_x |x\phi(x)| = 2\phi(1) = 2(.241971...) \le .5 < 4.
\end{aligned}$$

Taking $t = \lambda^{1/(1+\gamma)}$ and combining with (6) yields

$$\begin{aligned}
E(\hat{\mu}_i^{(\hat{w})} - \mu_i)^2 &\le \mu_i^2 + 2(1+\gamma) P(|Y_i| > \lambda^{1/(1+\gamma)}) \\
&\le \mu_i^2 + 2(1+\gamma)\left\{2\mu_i^2 + 2\lambda^{-1/(1+\gamma)}\phi(\lambda^{1/(1+\gamma)})\right\} \\
&= (1 + 4(1+\gamma))\mu_i^2 + 4(1+\gamma)\lambda^{-1/(1+\gamma)}\phi(\lambda^{1/(1+\gamma)}) \\
&= (5 + 4\gamma)\mu_i^2 + 4(1+\gamma)\lambda^{-1/(1+\gamma)}\phi(\lambda^{1/(1+\gamma)}). \qquad (7)
\end{aligned}$$

Combining (5) and (7), we find that (2) holds.

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 $Y_{ij} \equiv X_i X_j$, $0 \leq i, j \leq p$ satisfy $Var_P(Y_{ij}) \leq C$ 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 $\sup_{|t| \leq \epsilon} |\phi_{ij}^{(3)}(t)| \leq C_2$ for all $P \in \mathcal{P}_n$ and all i, j for some small $\epsilon > 0$.

Lemma 1. If condition 1 holds, then

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

Lemma 2. If condition 1 holds, then

$$\inf_{P \in \mathcal{P}_n} Pr_{P_n} \left(L_{P_n}(\beta) \leq \gamma' \Sigma_{\mathbb{P}_n} \gamma + |\gamma|' \hat{E} |\gamma| \text{ for all } \beta \in \mathbb{R}^p \right) \to 1 \tag{8}$$

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

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,

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

works.

Proof. As in (8),

$$\inf_{P_n \in \mathcal{P}_n} \inf_{\beta \in B_{n,b_n}} Pr_{P_n} \left(|L_{P_n}(\beta) - L_{\mathbb{P}_n}(\beta)| \leq |\gamma|' \hat{E} |\gamma| \right) \to 1.$$

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