STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 8 (April 19): In this lecture we will consider the "adaptive lasso" as discussed in the paper by Zou (2006).

Recall that the Lasso method introduced by Tibshirani (1996) seeks to

$$\text{minimize} \;\; S(\beta) = \frac{1}{2}\|Y - \mathbf{X}\beta\|^2 \qquad \text{subject to} \;\; \|\beta\|_1 \le t$$

for $0 < t < t_0 \equiv \|\widehat{\beta}^{LS}\|_1$. This is the "constrained form" of the optimization problem Putting this in penalized form yields the problem of minimizing

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i'\beta)^2 + \lambda \sum_{j=1}^k |\beta_j| \tag{1}$$

where $\lambda = \lambda_n$ is given. Here $\mathbf{x}_i \in \mathbb{R}^p$ for each i, so $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$ is $n \times p$.

Knight and Fu (2000) studied the lasso and bridge procedures under the following hypotheses:

Regularity conditions:
A1. $M_n \equiv n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' = n^{-1}\mathbf{X}'\mathbf{X} \to M$ where M is a $(p \times p)$ nonnegative definite matrix.
A2. $n^{-1}\max_{1\le i\le n} \mathbf{x}_i'\mathbf{x}_i \to 0$.
A3. $Y_i = \beta_0 + \mathbf{x}_i'\beta + \epsilon_i$ where ϵ_i are i.i.d. with $E\epsilon_i = 0$, $Var(\epsilon_i) < \infty$.

Define criterion functions $Z_n(\phi)$ by

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{x}_i'\phi)^2 + \frac{\lambda_n}{n}\sum_{j=1}^p |\phi_j|^\gamma, \qquad \phi \in \mathbb{R}^p. \tag{2}$$

Thus $Z_n(\phi)$ is minimized at $\phi = \widehat{\beta}_n(\gamma) \equiv \widehat{\beta}_n$.

Theorem 1. (Knight and Fu) If M is nonsingular and $\lambda_n/n \to \lambda_0 \ge 0$, then

$$\widehat{\beta}_n^{(\gamma)} \to_p \text{argmin}_\phi Z(\phi)$$

where

$$Z(\phi) = (\phi - \beta)'M(\phi - \beta) + \lambda_0 \sum_{j=1}^p |\phi_j|.$$

Thus if $\lambda_n = o(n)$ and $\lambda_0 = 0$, $\text{argmin}(Z) = \beta$ and $\widehat{\beta}_n^{(1)}$ is consistent.

They also proved the following result concerning asymptotic normality of the lasso estimator $\widehat{\beta}_n^{(1)}$.

Theorem 2. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and M is nonsingular, then

$$\sqrt{n}(\widehat{\beta}_n^{(1)} - \beta) \to_d \text{argmin}_\phi(V)$$

where, with $W \sim N(0, \sigma^2 M)$,

$$V(u) = -2u'W + u'Mu + \lambda_0 \sum_{j=1}^{p} \{u_j \text{sign}(\beta_j) 1\{\beta_j \neq 0\} + |u_j| 1\{\beta_j = 0\}\}.$$

On the other hand, for the case of "local alternatives" $\beta_n = \beta + n^{-1/2}t$ we have the following theorem:

Theorem 5. Suppose that the triangular array versions of conditions **A1** - **A3** hold with $\beta_n = \beta + t/\sqrt{n}$. Let $\widehat{\beta}_n^{(1)}$ minimize (1) where $\lambda_n/n^{1/2} \to \lambda_0 \geq 0$.
(a)

$$\sqrt{n}(\widehat{\beta}_n^{(1)} - \beta_n) \to_d \text{argmin}(V)$$

where

$$V(u) = -2u'W + u'Mu + \lambda_0 \sum_{j=1}^{p} \{u_j \text{sign}(\beta_j) 1\{\beta_j \neq 0\} + |u_j + t_j| 1\{\beta_j = 0\}.$$

Zou's further study of the lasso:

Let $\mathcal{A} \equiv \{j \in \{1, \ldots, p\} : \ \beta_j \neq 0\} = \{1, \ldots, p_0\}$ without loss of generality, and let $\mathcal{A}_n \equiv \{j \in \{1, \ldots, p\} : \ \widehat{\beta}_j^{(1)} \neq 0\}$.

Proposition 1. If $\lambda_n/\sqrt{n} \to \lambda_0$, then

$$\limsup_{n \to \infty} P(\mathcal{A}_n = \mathcal{A}) \leq c < 1$$

where c is a constant depending β and σ^2.

Proposition 2. If $\lambda_n/n \to 0$, $\lambda_n/\sqrt{n} \to \infty$ (e.g. $\lambda_n = n^{3/4}$, then

$$\frac{n}{\lambda_n}(\widehat{\beta}^{(1)} - \beta) \to_p \text{argmin}(V_3)$$

where

$$V_3(u) = u'Mu + \sum_{j=1}^{p} \left\{u_j \text{sign}(\beta_j) 1\{\beta_j \neq 0\} + |u_j| 1\{\beta_j = 0\}\right\}.$$

Note that:
(i) $n/\lambda_n = o(\sqrt{n})$ (i.e. the rate of convergence is slower than $\sqrt{n}$).
(ii) The limit is *not random*.

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

$$|M_{21}M_{11}^{-1}| \leq 1 \qquad \text{where} \qquad M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \tag{3}$$

Example 1. Suppose that $p_0 = 2m + 1 \geq 3$, $p = p_0 + 1$ (so there is just one irrelevant predictor). Let

$$\begin{aligned} M_{11} &= (1-\rho_1)I + \rho_1 J, \qquad J = \text{matrix of all } \ 1's, \\ M_{12} &= \rho_2 \underline{1}, \\ M_{22} &= 1. \end{aligned}$$

where

$$-\frac{1}{p_0 - 1} < \rho_1 < -\frac{1}{p_0}, \qquad \text{and} \qquad 1 + (p_0 - 1)\rho_1 < |\rho_2| < \sqrt{\frac{1 + (p_0 - 1)\rho_1}{p_0}}.$$

Then (3) fails and the lasso is model selection inconsistent.

Zou's Adaptive Lasso:

Consider a weighted version of the lasso problem as follows:

$$\widehat{\beta}^{(w)} \equiv \text{argmin}_\phi \left\{ \|Y - \mathbf{X}\phi\|^2 + \lambda \sum_{j=1}^{p} w_j |\phi_j| \right\}$$

for a vector $w = (w_1, \ldots, w_p)$ of weights. Note that increasing a particular weight w_j has the effect of increasing λ, but for just the j-th coefficient. If we let w depend on the data in some data-dependent way that accomplishes this, we are lead to the *adaptive lasso*. For $\gamma > 0$, define an "estimated weight vector" $\hat{w}$ by

$$\underline{\hat{w}} = \frac{1}{|\widehat{\beta}_n^{(0)}|^\gamma} = (1/|\widehat{\beta}_{n1}^{(0)}|^\gamma, \ldots, \widehat{\beta}_{np}^{(0)}|^\gamma$$

where $\widehat{\beta}_n^{(0)}$ is the least squares estimator. Then define

$$\begin{aligned} &\widehat{\beta}^{(\hat{w})} = \text{argmin}_\phi \|Y - \mathbf{X}\phi\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\phi_j|, \\ &\mathcal{A}_n^{(\hat{w})} \equiv \{j \in \{1, \ldots, p\} : \ \widehat{\beta}_j^{(\hat{w})} \neq 0\}. \end{aligned}$$

Theorem 2. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates satisfy:
(a) $P(\mathcal{A}_n^{(\hat{w})} = \mathcal{A}) \to 1$.
(b) $\sqrt{n}(\widehat{\beta}_{n,\mathcal{A}}^{(\hat{w})} - \beta_{\mathcal{A}}) \to_d N(0, \sigma^2 M_{11}^{-1})$.

Remark 0. If $\lambda_n = n^{3/8}$ and $\gamma =??$, then $\lambda_n n^{(\gamma-1)/2} = n^{3/8} n^{(\gamma-1)/2} \to \infty$ if $\gamma > 1/4$.

Remark 1. The least squares estimator need not be $\sqrt{n}-$consistent for the adaptive lasso procedure to work. If there is a sequence $\{a_n\}$ with $a_n \to \infty$ and $a_n(\widehat{\beta}_n^{(0)} - \beta) = O_p(1)$, then properties (a) and (b) in Theorem 2 continue to hold if $\lambda_n = o(\sqrt{n})$ and $a_n^\gamma \lambda_n/\sqrt{n} \to \infty$.

Remark 2. Note that the weights $\hat{w}_j$ for $j \notin \mathcal{A}$ converge to ∞, while for $j \in \mathcal{A}$, $\hat{w}_j \to_p 1/|\beta_j|^\gamma$.

Remark 3. The adaptive lasso solution is continuous. The bridge estimators with $0 < \gamma <!$ have oracle properties but are inconsistent.

Now consider the Gaussian means model of lectures 2-4:

$$Y_i = \mu_i + \epsilon_i, \qquad \epsilon_i \qquad \text{i.i.d.} \quad N(0,1).$$

Consider estimation of $\underline{\mu} = (\mu_1, \ldots \mu_n)$ by estimators $\hat{\mu}_n$. The risk of an estimator $\hat{\mu}_n$ is given by

$$R(\mu, \hat{\mu}_n) = E \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2,$$

while the "ideal risk" is given by

$$R(\mu, \hat{\mu}(\text{ideal}) = \sum_{i=1}^n \min\{\mu_i^2, 1\}.$$

Donoho and Johnstone (1994) showed that the soft-thresholding estimators $\hat{\mu}_i(\text{soft})$,

$$\begin{aligned} \hat{\mu}_i(\text{soft}) &= \text{argmin}_u \left(\frac{1}{2}(Y_i - u)^2 + \lambda |u| \right), \qquad i = 1, \ldots, n, \\ &= (|Y_i| - \lambda)_+ \text{sign} Y_i, \end{aligned}$$

have risk differing from the ideal risk by at most a factor of $2 \log n$, and that the $2 \log n$ factor is a sharp minimax bound.

Zou (2006) proposes the following penalized version of the naive estimators $\hat{\mu}(\text{naive}) = Y_i$:

$$\begin{aligned} \hat{\mu}_i^{(\hat{w})} &= \text{argmin}_u \left(\frac{1}{2}(Y_i - u)^2 + \lambda \frac{1}{|Y_i|^\gamma} |u| \right) \\ &= \left(|Y_i| - \frac{\lambda}{|Y_i|^\gamma} \right)_+ \text{sign}(Y_i), \qquad i = 1, \ldots, n. \end{aligned}$$

For the estimator $\hat{\mu}^{(\hat{w})}$, Zou proves the following oracle inequality:

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

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