STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 7 (April 17): This lecture begins discussion of statistical aspects of Lasso and related methods, as discussed in the paper by Knight and Fu (2000).

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

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

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) treat the whole family of "bridge estimators" $\widehat{\beta} = \widehat{\beta}^{(\gamma)}$ defined for $\gamma > 0$ by

$$\widehat{\beta}^{(\gamma)} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2 + \lambda_n \sum_{j=1}^{n} |\beta_j|^\gamma \right\}.$$

These estimators were introduced by Frank and Friedman (1993). The lasso estimator of Tibshirani (1996) corresponds to $\gamma = 1$, while the classical ridge estimator of Hoerl and Kennard (1970) corresponds to $\gamma = 2$.

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 $E\epsilon_i = 0$, $Var(\epsilon_i) < \infty$.

Under conditions **A1 - A3** with M nonsingular, it is well-known that the Least Squares Estimator $\widehat{\beta}_n^{(0)}$ is consistent and that

$$\sqrt{n}(\widehat{\beta}_n^{(0)} - \beta) \to_d N(0, \sigma^2 M^{-1}).$$

Conditions **A1 - A3** can be weakened considerably without destroying asymptotic normality; see e.g. Srivastava (1971). Knight and Fu (2000) work under the assumption that the limit matrix M is non-singular.

Define criterion functions $Z_n(\phi)$ by

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

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

Theorem 1. If M is nonsingular and $\lambda_n/n \to \lambda_0 \geq 0$, then

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

where

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

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

Proof. With Z_n defined as in (1) we need to show that

$$\sup_{\phi\in K}|Z_n(\phi) - Z(\phi) - \sigma^2| \to_p 0 \tag{2}$$

for any compact set $K \subset \mathbb{R}^p$ and that

$$\widehat{\beta}_n^{(\gamma)} = O_p(1). \tag{3}$$

Once (2) and (3) have been proved, then

$$\text{argmin}(Z_n) \to_p \text{argmin}(Z)$$

follows by the argmin continuous mapping theorem (see e.g. van der Vaart and Wellner (1996), section 3.2, pages 285 - 289, or Geyer (1996)). For $\gamma \geq 1$, Z_n is convex; in this case (2) and (3) follows from the pointwise convergence in probability of $Z_n(\phi)$ to $Z(\phi) + \sigma^2$ by applying standard results (e.g. Andersen and Gill (1982) or Pollard (1991)). (See van der Vaart and Wellner (1996), page 208 for two key exercises concerning convex/concave criterion functions.) For $\gamma < 1$ the criterion function is not convex, but (2) still holds via direct empirical process arguments. To prove (3), note that

$$Z_n(\phi) \geq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\phi)^2 \equiv Z_n^{(0)}(\phi)$$

for all ϕ. Since $\text{argmin}(Z_n^{(0)}) = O_p(1)$, it follows that $\text{argmin}(Z_n) = O_p(1)$. □

Now we can treat asymptotic normality of $\widehat{\beta}_n^{(\gamma)}$ assuming that $\lambda_n = o(n)$. What will be needed for asymptotic normality is that $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. But here we will see that if $\lambda_0 = 0$, the limiting distribution is just the same as that of the least squares estimator.

Theorem 2. Suppose that $\gamma \geq 1$. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and M is nonsingular, then

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

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

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

Proof. Define $V_n(u)$ by

$$\begin{aligned} V_n(u) &= n(Z_n(\beta + n^{-1/2}u) - Z_n(\beta)) \\ &= \sum_{i=1}^n \{(\epsilon_i - \mathbf{x}_i'u/\sqrt{n})^2 - \epsilon_i^2\} + \lambda_n \sum_{j=1}^p \{|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma\}. \end{aligned}$$

Note that V_n is minimized at $\sqrt{n}(\widehat{\beta}_n^{(\gamma)} - \beta)$. Now the first term in $V_n(u)$ converges in distribution easily:

$$\begin{aligned} \sum_{i=1}^n \{(\epsilon_i - \mathbf{x}_i'u/\sqrt{n})^2 - \epsilon_i^2\} &= -2\frac{u'}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i \epsilon_i + u' \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' u \\ &\to_d -2u'W + u'Mu \end{aligned}$$

by **A1 - A3**. If $\gamma > 1$, then the second term satisfies

$$\lambda_n \sum_{j=1}^p \{|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma\} \to \lambda_0 \sum_{j=1}^p u_j \text{sign}(\beta_j)|\beta_j|^{\gamma-1},$$

and for $\gamma = 1$

$$\lambda_n \sum_{j=1}^p \{|\beta_j + u_j/\sqrt{n}| - |\beta_j|\} \to \lambda_0 \sum_{j=1}^p \{u_j \text{sign}(\beta_j) 1\{\beta_j \neq 0\} + |u_j| 1\{\beta_j = 0\}\}.$$

Thus $V_n(u) \to_d V(u)$ with the finite-dimensional convergence holding trivially. Since V_n is convex and V has a unique minimum, it follows from Geyer (1966) (or argmin-continuous mapping) that

$$\text{argmin}(V_n) = \sqrt{n}(\widehat{\beta}_n - \beta) \to_d \text{argmin}(V).$$

Note that if $\lambda_0 = 0$, then $\text{argmin}(V) = M^{-1}W \sim N(0, \sigma^2 M^{-1})$, the same as the limiting distribution of the least squares estimator. □

Corollary. If $\gamma = 2$, and **A1 - A3** hold, then

$$\sqrt{n}(\widehat{\beta}^{(2)} - \beta) \to_d M^{-1}(W - \lambda_0 \beta) \sim N(-\lambda_0 M^{-1}\beta, \sigma^2 M^{-1}).$$

The corollary suggests that the asymptotic mean square error of the ridge regression estimator $\widehat{\beta}_n^{(2)}$ is larger than that of the ordinary least squares estimator. But of course the current situation (with M non-singular) is not putting the ridge estimator in its preferred setting.

Theorem 3. Suppose that $\gamma < 1$. If $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$, then

$$\sqrt{n}(\widehat{\beta}_n^{(\gamma)} - \beta) \to_d \text{argmin}_u V_\gamma(u)$$

where

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

Proof. This goes along the lines of the proof of Theorem 2, but with some complications arising from the lack of convexity of the objective function. Define

What if $\gamma < 1$, $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$, but $\lambda_n/n^{\gamma/2} \to \infty$? For example, suppose that $\lambda_n = \lambda_0\sqrt{n}$ with $\lambda_0 > 0$. Then $\lambda_n/\sqrt{n} = \lambda_0$, and $\lambda_n/n^{\gamma/2} = \lambda_0 n^{1/2-\gamma/2} \to \infty$.

$$\begin{aligned} V_n(u) &= n(Z_n(\beta + n^{-1/2}u) - Z_n(\beta)) \\ &= \sum_{i=1}^{n}\{(\epsilon_i - \mathbf{x}_i'u/\sqrt{n})^2 - \epsilon_i^2\} + \lambda_n \sum_{j=1}^{p}\{|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma\} \end{aligned}$$

as before. Since $\lambda_n = O(n^{\gamma/2}) = o(\sqrt{n})$, it follows that

$$\lambda_n\{|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma\} \to 0$$

if $\beta_j \neq 0$. Thus

$$\lambda_n \sum_{j=1}^{p}\{|\beta_j + u_j/\sqrt{n}|^\gamma - |\beta_j|^\gamma\} \to \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma 1\{\beta_j \neq 0\}$$

where the convergence is uniform over u in compact sets. Thus it follows that

$$V_n \Rightarrow V$$

on the space of functions topologized by uniform convergence on compact sets. To show that $\text{argmin}(V_n) \to_d \text{argmin}(V)$, it suffices to show that $\text{argmin}(V_n) = O_p(1)$ [see e.g. Kim and Pollard (1990) or van der Vaart and Wellner (1996), section 3.2]. But we note that

$$\begin{aligned} V_n(u) &\geq \sum_{i=1}^{n}\{(\epsilon_i - \mathbf{x}_i'u/\sqrt{n})^2 - \epsilon_i^2\} - \lambda_n \sum_{j=1}^{p} |u_j/\sqrt{n}|^\gamma \\ &\geq \sum_{i=1}^{n}\{(\epsilon_i - \mathbf{x}_i'u/\sqrt{n})^2 - \epsilon_i^2\} - (\lambda_0 + \delta) \sum_{j=1}^{p} |u_j|^\gamma \\ &= V_n^{(l)}(u) \end{aligned}$$

for all u and n sufficiently large. Since the quadratic terms in $V_n^{(l)}$ grow faster than the $|u_j|^\gamma$ terms, it follows that $\text{argmin}(V_n^{(l)}) = O_p(1)$; hence it follows that $\text{argmin}(V_n) = O_p(1)$. But $\text{argmin}(V)$ is unique with probability 1, the conclusion follows.

Now suppose that

$$Y_{ni} = \beta_n' \mathbf{x}_{ni} + \epsilon_{ni}, \qquad i = 1, \ldots, n, \tag{4}$$

where $\epsilon_{n1}, \ldots, \epsilon_{nn}$ are i.i.d. random variables with mean 0 and variance σ^2. We assume that

$$n^{-1} \sum_{i=1}^{n} \mathbf{x}_{ni}' \mathbf{x}_{ni}' \to M, \qquad \text{positive definite}, \tag{5}$$

and

$$n^{-1} \max_{1 \le i \le n} \mathbf{x}_{ni}' \mathbf{x}_{ni} \to 0. \tag{6}$$

Suppose that $\beta_n = \beta + n^{-1/2} t$ and define $\widehat{\beta}_n^{(\gamma)}$ to minimize (1) with $Y_i,\ \mathbf{x}_i$ replaced by $Y_{ni},\ \mathbf{x}_{ni}$.

Theorem 4. Suppose that (4) holds with $\beta_n = \beta + t/\sqrt{n}$ and assume that (5) and (6) holds. Let $\widehat{\beta}_n^{(\gamma)}$ minimize (1) for some $\gamma > 1$.
(a) If $\lambda_n/\sqrt{n} \to \lambda_0 \ge 0$, then

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

where

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

(b) If $\beta = 0$ and $\lambda_n / n^{\gamma/2} \to \lambda_0 \ge 0$, then

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

where

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

On the other hand, for the case $\gamma \le 1$ we have the following theorem:

Theorem 5. Suppose that (4) holds with $\beta_n = \beta + t/\sqrt{n}$ and assume that (5) and (6) holds. Let $\widehat{\beta}_n^{(\gamma)}$ minimize (1) for some $\gamma \le 1$ where $\lambda_n / n^{\gamma/2} \to \lambda_0 \ge 0$.
(a) For $\gamma = 1$,

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

where

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

(b) For $\gamma < 1$,

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

where

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

Theorem 4 suggests that the advantages of using a penalty with $\gamma > 1$ occur only when all the regression parameters β_j are small relative to n: here is the corollary for ridge regression in this case.

Corollary. Suppose that $\beta_n = 0 + n^{-1/2}\ t$, $\lambda_n/\sqrt{n} \rightarrow \lambda_0 > 0$, and $\gamma = 2$. Then

$$\begin{aligned} \sqrt{n}(\widehat{\beta}_n^{(2)} - t/\sqrt{n}) &\rightarrow_d (M + \lambda_0 I)^{-1}(W - \lambda_0 t) \\ &\sim N(-\lambda_0 (M + \lambda_0 I)^{-1} t,\ \sigma^2((M + \lambda_0 I)M(M + \lambda_0 I)^{-1}). \end{aligned}$$

This suggests that we can choose λ_0 to make the mean square error of $\mathbf{x}'\widehat{\beta}_n^{(\gamma)}$ smaller than the MSE of $\mathbf{x}'\widehat{\beta}_n^{(2)}$. If some of the β_j's are non-zero (or "large"), then part (a) of Theorem 4 shows that the bias suggested by Theorem 2 would still persist.

On the other hand, when $\gamma \leq 1$, Theorem 5 suggests that "small" parameters may be estimated exactly as 0 even in the presence of some "large" parameters.

Example: Suppose that $\beta_{nj} = t_j/\sqrt{n}$ and let $\gamma \leq 1$. Then the limiting distribution of $\sqrt{n}(\widehat{\beta}_{nj}^{(\gamma)} - t_n/\sqrt{n})$ puts positive probability mass at $-t_j$, and hence the limiting distribution of $\sqrt{n}\widehat{\beta}_{nj}$ puts positive probability mass at 0.

What if the design matrix is singular?

Define $\hat{\beta}_\lambda$ to minimize the objective function

$$\sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\phi)^2 + \lambda \sum_{j=1}^{p} |\phi_j|^\gamma. \tag{7}$$

The $\hat{\beta}_\lambda$ minimizes (7), it also minimizes

$$h_\lambda(\phi) = \frac{1}{\lambda} \left\{ \sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\phi)^2 - \sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\hat{\beta}^{(0)})^2 \right\} + \sum_{j=1}^{p} |\phi_j|^\gamma \tag{8}$$

where $\hat{\beta}^{(0)}$ is a LS estimator of β; i.e. $\hat{\beta}^{(0)}$ satisfies

$$\sum_{i=1}^{n} \mathbf{x}_i(Y_i - \mathbf{x}_i'\hat{\beta}^{(0)}) = 0.$$

As $\lambda \to 0$, h_λ in (8) epi-converges to the function

$$h_0(\phi) = \begin{cases} \sum_{j=1}^{p} |\phi_j|^p, & \text{if } \sum_{i=1}^{n} \mathbf{x}_i(Y_i - \mathbf{x}_i'\phi) = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

Hence if argmin(h_0) is unique (as it certainly is when $\gamma > 1$), then

$$\hat{\beta}_\lambda \to \hat{\beta}_0 \equiv \text{argmin} \left\{ \sum_{j=1}^{p} |\phi_j|^\gamma : \quad \sum_{i=1}^{n} \mathbf{x}_i(Y_i - \mathbf{x}_i'\phi) = 0 \right\} \qquad \text{as} \quad \lambda \to 0.$$

When $\gamma = 1$, this is related to the optimization problem used by Candès and Tao to define the Dantzig selector:

$$\text{minimize} \quad \sum_{i=1}^{p} |\phi_j| \quad \text{subject to} \quad \|\mathbf{X}'(Y - \mathbf{X}\phi)\|_\infty \le (1 + t^{-1})\sqrt{2 \log p} \cdot \sigma.$$