STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 5 (April 10): This lecture concludes the discussion of (the discussion of) Shao (1997), and begins treatment of lasso, bridge, and other procedures based on ℓ_p and ℓ_1 penalty (or regularization) terms.

Beran's discussion of Shao: Suppose that

$$\begin{aligned} &Y_{n,i} = \mu_{n,i} + \epsilon_i, \qquad \text{where } \ \epsilon_i \sim N(0,\sigma^2), \\ &\text{ave}(f) \equiv n^{-1}\sum_1^n f(i) \qquad \text{for } \ i \in \{1,\ldots,n\} = \{t_1,\ldots,t_n\}, \\ &\mu_n = (\mu_{n,1},\ldots,\mu_{n,n}). \end{aligned}$$

The *time-averaged quadratic loss* of an estimator $\hat\mu_n$ of μ_n is

$$L_n(\hat\mu_n,\mu_n) = n^{-1}\sum_{i=1}^n (\hat\mu_{n,i}-\mu_{n,i})^2,$$

and the corresponding risk is

$$R_n(\hat\mu_n,\mu_n,\sigma^2) = EL_n(\hat\mu_n,\mu_n).$$

Let $u \in [0,1]$ index candidate models and corresponding estimators in the nested case, with

$$\begin{aligned} &\mu_n(u) \equiv \{\mu_{n,i}(u): \ i=1,\ldots,n\}, \qquad \text{with } \ \mu_{n,i}(u) = \mu_{n,i}1\{i/(n+1)\le u\}, \\ &\hat\mu_n(u) \equiv \{\hat\mu_{n,i}(u): \ i=1,\ldots,n\}, \qquad \text{with } \ \hat\mu_{n,i}(u) = Y_{n,i}1\{i/(n+1)\le u\}. \end{aligned}$$

Choose u by the GIC (λ_n) method: suppose that $\hat\sigma_n^2 \to_p \sigma^2$ and

$$\sup_{n^{-1}\sum_1^n \mu_{n,i}^2/\sigma^2 \le r} E|\hat\sigma_n^2 - \sigma^2| \to 0 \quad \text{for all } \ r \in [0,\infty). \tag{1}$$

Then the GIC(λ_n) criterion is

$$\Gamma_n(u,\lambda_n) = \widehat\sigma_n^2(u) + \frac{\lambda_n\hat\sigma_n^2[(n+1)u]}{n}$$

where

$$\widehat\sigma_n^2(u) = n^{-1}\sum_{i/(n+1)>u} Y_{n,i}^2.$$

Let

$$\widehat u_n \equiv \text{argmin}_{u\in[0,1]}\Gamma_n(u,\lambda_n).$$

Proposition 1. In the signal plus noise model with $\hat{\sigma}_n^2$ satisfying (1),

$$\begin{aligned}
&\sup_{\overline{\mu_n^2}/\sigma^2 \le r} R_n(\widehat{\mu}_{n,2}, \mu_n, \sigma^2) \to \sigma^2(r \wedge 1), \\
&\sup_{\overline{\mu_n^2}/\sigma^2 \le r} R_n(\widehat{\mu}_{n,\lambda_n}, \mu_n, \sigma^2) \to \sigma^2 r, \qquad \text{if } \lambda_n \to \infty, \\
&\sup_{\overline{\mu_n^2}/\sigma^2 \le r} R_n(Y_n, \mu_n, \sigma^2) \to \sigma^2
\end{aligned}$$

Note that if $\mu_n = (\mu_{n,1}, \ldots, \mu_{n,k_n}, 0, \ldots, 0)$ has k_n- non-zero entries, then

$$\frac{\overline{\mu_n^2}}{\sigma^2} = \frac{n^{-1}\sum_1^{k_n} \mu_{ni}^2}{\sigma^2} \le \frac{k_n \max_{1 \le i \le k_n} \mu_{ni}^2}{n\sigma^2}.$$

On the other hand, how well do model selection estimators do in the class of *all estimators of* μ_n? The following result which follows from Pinsker (1980) and/or Stein (1956) says that they are asymptotically inadmissible in the current setting.

Proposition 2. In the signal plus noise model with $\hat{\sigma}_n^2$ satisfying (1), the following equality holds for every $r \in [0, \infty)$:

$$\lim_{n\to\infty} \inf_{\hat{\mu}_n} \sup_{\overline{\mu_n^2}/\sigma^2 \le r} R_n(\hat{\mu}_n, \mu_n, \sigma^2) = \sigma^2 \frac{r}{r+1}.$$

Proof. This follows from Pinsker (1980) and/or Stein (1956); see also Beran (1996). □

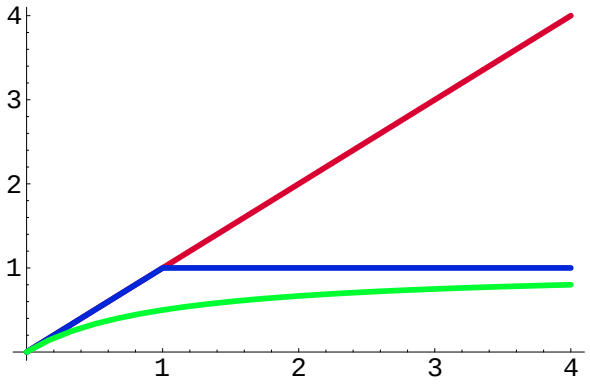

Figure 1: Maximum risk comparisons, in multiples of variance versus signal to noise ratio r

Thus to be asymptotically minimax, an estimator $\widehat{\mu}_n$ must satisfy

$$\sup_{\overline{\mu_n^2}/\sigma^2 \le r} R_n(\hat{\mu}_n, \mu_n, \sigma^2) \to \sigma^2 \frac{r}{r+1}.$$

The simplest such estimator is the James-Stein estimator

$$\widehat{\mu}_{n,S} \equiv \left(1 - \frac{\hat{\sigma}_n^2}{\text{ave}(Y_n^2)}\right)^+ Y_n.$$

The Lasso and variants: The following material is based on Tibshirani (1996).

Suppose that $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$ are independent. Here

$$\begin{aligned} &\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \qquad \text{is a vector of predictors or covariates,} \\ &Y_i \qquad \text{is a response for item } i, \quad i = 1, \ldots, n. \end{aligned}$$

Suppose that the x_{ij}'s are standardized so that

$$\frac{1}{n}\sum_{i=1}^{n} x_{i,j} = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 = 1.$$

Let $\beta = (\beta_1, \ldots, \beta_p)'$. The *lasso estimator* of β is given by

$$(\widehat{\alpha}, \widehat{\beta}) = \text{argmin}_{\alpha,\beta} \left\{ \sum_{i=1}^{n} (Y_i - \alpha - \mathbf{x}_i'\beta)^2 : \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \right\}.$$

For all t the solution for α is $\widehat{\alpha} = \overline{Y}$. So assume without loss of generality (at least for computational purposes) that $\overline{Y} = 0$ and omit α. Let $\widehat{\beta}^0$ denote that full least squares estimator (with the constraint omitted), and set

$$t_0 \equiv \sum_{j=1}^{p} |\widehat{\beta}_j^0|.$$

A family of problems generalizing the lasso estimator was proposed by Frank and Friedman (1993): for $\gamma > 0$,

$$(\widehat{\alpha}, \widehat{\beta})_\gamma = \text{argmin}_{\alpha,\beta} \left\{ \sum_{i=1}^{n} (Y_i - \alpha - \mathbf{x}_i'\beta)^2 : \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j|^\gamma \leq t \right\}.$$

As noted in lecture 1, $\gamma = 1$ corresponds to the lasso, while $\gamma = 2$ corresponds to ridge regression, and $\gamma \searrow 0$ yields more classical model selection methods based on ℓ_0 penalty terms.

In the meantime, a large number of variants and extensions of the lasso have been proposed:

- **Relaxed lasso** (Meihshausen, 2006)
- **Adaptive lasso** (Zou, 2006; Huang, Ma, Zhang, 2006)
- **SCAD methods** (Fan and Li, 2001)
- **Dantzig selector** (Candès and Tao, 2007)
- **Fused lasso** (Tibshirani, Rosset, Zhu, Knight, 2005)

We will spend most of the rest of the quarter studying these various variants of the lasso and their properties. But first it will be helpful to consider some optimization and computational aspects of the plain vanilla lasso. We will begin with the work of Osborne, Presnell, and Turlach (2000), who gave a careful study of the dual optimization problem. We will then consider the computational scheme devised by Efron, Hastie, Tibshirani, and Johnstone based on least angle regression (lars).

Lasso: the primal and dual problems

The original *constrained regression problem* as formulated by Tibshirani (1996) is

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \le t.$$

As noted by Tibshirani (1996), page 277, this is equivalent to the following *penalized regression problem*:

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

Tibshirani refers to Gill, Murray, and Wright (1981), chapter 5, but chapter 5 of GMW treats only linear constraints while the current constraint is non-linear. Chapter 6 of GMW treats nonlinear constraints such as $\sum|\beta_j| \le t$; see their example 6.2.2.1, page 215, but I have not yet succeeded in developing the precise correspondence from the material there. On the other hand, Osborne, Presnell, and Turlach (2000) give a very precise treatment of the dual problem which is related.

We will write $A = \mathbf{X}^T\mathbf{X}$, and assume that $\mathbf{X}$ has maximal rank $n \wedge p$, and write $N(\mathbf{X}) \subset \mathbb{R}^p$ for the null space of $\mathbf{X}$.

- If $p \le n$, $N(\mathbf{X}) = \{0\}$, and $\widehat{\beta}^0 = A^{-1}\mathbf{X}'Y$ is unique.

- If $p > n$, then $N(\mathbf{X})$ has dimension $p - n$, $\widehat{\beta}^0$ is not unique, and $X(\widehat{\beta}^0 + \eta) = Y$ for $\eta \in N(\mathbf{X})$.

In either case, define

$$t_0 \equiv \min_{\eta \in N(\mathbf{X})} \|\widehat{\beta}^0 + \eta\|_1$$

where $\|\beta\|_1 \equiv \sum_1^p |\beta_j|$. Then the lasso is equivalent to ordinary least squares when $t \ge t_0$, so assume $t < t_0$.

Problem:

$$\text{minimize } f(\beta) \text{ subject to } g(\beta) \ge 0 \tag{2}$$

where

$$f(\beta) = \frac{1}{2}(Y - \mathbf{X}\beta)'(Y - \mathbf{X}\beta) = \frac{1}{2}r'r$$
$$g(\beta) = t - \sum_{j=1}^{p} |\beta_j|,$$
$$r = r(\beta) = Y - \mathbf{X}\beta.$$

Since f is continuous and the region of feasible β's, $\{\beta : \|\beta\|_1 \leq t\}$, is compact, a solution exists. Since $t < t_0$, all critical values of f occur outside the feasible region, and any solution β^* must lie on its boundary, i.e. $\|\beta^*\|_1 = t$. Since g is concave, the region of feasible values is convex, and since f is convex, the solution set is convex. If $p \leq n$, f is strictly convex and the solution β^* is unique.

Theorem 1. (Existence and uniqueness). If $t < t_0$, then the following hold:
(a) if $p \leq n$, then a unique solution β^* of (2) exists and $\beta^*\|_1 = t$.
(b) If $p > n$, then a solution β^* of (2) exists and $\|\beta^*\|_1 = t$ for any solution. If β_1^* and β_2^* are both solutions, then $\lambda\beta_1^* + (1-\lambda)\beta_2^*$ is also a solution for all $0 \leq \lambda \leq 1$.

Now we define the Lagrangian associated with the problem, namely

$$\begin{aligned}\mathcal{L}(\beta, \lambda) &= f(\beta) - \lambda g(\beta) = f(\beta) + \lambda(-g(\beta)) \\ &= f(\beta) + \lambda \sum_{1}^{p} |\beta_j| - \lambda t.\end{aligned}$$

Define

$$\begin{aligned}\mathcal{L}^*(\beta) &\equiv \sup_{\lambda \geq 0} \mathcal{L}(\beta, \lambda) \\ &= \begin{cases} f(\beta) & \text{if } g(\beta) \geq 0 \\ \infty & \text{if } g(\beta) < 0. \end{cases}\end{aligned}$$

Thus minimizing $\mathcal{L}^*(\beta)$ is equivalent to solving (2); this is the *primal problem*. For $\lambda \geq 0$ the dual objective function is

$$\mathcal{L}_*(\lambda) = \inf_{\beta} \mathcal{L}(\beta, \lambda)$$

and the dual problem is

$$\text{maximize}_{\lambda \geq 0}\ \mathcal{L}_*(\lambda). \tag{3}$$

If we fix $\lambda \geq 0$, $\mathcal{L}(\beta, \lambda)$ is a convex function of β and $\mathcal{L}(\beta, \lambda) \to \infty$ as $\|\beta\|_1 \to \infty$. Hence $\mathcal{L}(\cdot, \lambda)$ has at least one minimum and $\bar{\beta} = \bar{\beta}(\lambda)$ minimizes $\mathcal{L}(\beta, \lambda)$ if and only if the $p-$dimensional null-vector 0 is an element of the sub-differential $\partial_\beta \mathcal{L}(\bar{\beta}, \lambda)$. In our current problem

$$\partial_\beta \mathcal{L}(\bar{\beta}, \lambda) = -\mathbf{X}^T r + \lambda v$$

where v is a vector with components v_i satisfying

$$v_i = \begin{cases} 1, & \beta_i > 0, \\ -1, & \beta_i < 0, \\ \in [-1, 1], & \beta_i = 0. \end{cases}$$

Thus if $\overline{\beta}$ minimizes $\mathcal{L}(\beta, \lambda)$ for a given value of λ, then

$$0 = -\mathbf{X}^T \bar{r} + \lambda v, \qquad \bar{r} = Y - \mathbf{X}\bar{\beta} \tag{4}$$

for some v of the form above. But the form of v implies that $v'\bar{\beta} = \|\bar{\beta}\|_1$, and thus it follows from (4) that if $\bar{\beta}$ minimizes $\mathcal{L}(\beta, \lambda)$, then $\lambda = \bar{r}'\mathbf{X}\bar{\beta}/\|\bar{\beta}\|_1$. Alternatively, if $\bar{\beta} \neq 0$, which is the case whenever $t > 0$, then $\|v\|_\infty = 1$, and it follows from (4) that

$$\lambda = \|\mathbf{X}^T \bar{r}\|_\infty.$$

Using these two expressions for λ yields

$$\begin{aligned} \mathcal{L}_*(\lambda) &= \mathcal{L}(\bar{\beta}, \lambda) \\ &= \frac{1}{2} r'r - \frac{\bar{r}'\mathbf{X}\bar{\beta}}{\|\bar{\beta}\|_1}(t - \|\bar{\beta}\|_1) \\ &= \frac{1}{2} r'r + \bar{r}'\mathbf{X}\bar{\beta} - t\frac{\bar{r}'\mathbf{X}\bar{\beta}}{\|\bar{\beta}\|_1} \equiv \tilde{h}(\bar{\beta}) \\ &= \frac{1}{2} Y'Y - \frac{1}{2}\bar{\beta}' A \bar{\beta} - t\|\mathbf{X}\bar{r}\|_\infty \equiv \overline{h}(\bar{\beta}). \end{aligned}$$

Remark. In ℓ_2 penalized regression (ridge regression), it is typically true that the solution $\beta \to 0$ as $\lambda \to \infty$, but for any finite λ all entries in β are non-zero. In contrast, in ℓ_1-penalized regression it follows from (4) that as soon as $\lambda \geq \|\mathbf{X}^T Y\|_\infty$ is chosen, $\beta = 0$ is a solution of (2). To see this, note that if $\beta = 0$, then $r = r(\beta) = Y$, and if we choose $v = \mathbf{X}^T Y/\lambda$, then (4) holds and v is of the required form; that is, each of its components has an absolute value less than or equal to one. Thus if we chose the smoothing parameter λ in ℓ_1-penalized regression adaptively – e.g. by cross-validation – then the search for the optimal parameter λ can be conveniently restricted to the interval from 0 to $\|\mathbf{X}'Y\|_\infty$.

The following result relates the primal and dual problems:

Theorem 2. (weak duality). if β^* is a solution of (2) and $\bar{\lambda}$ is a solution of the dual problem (3), then $\mathcal{L}(\bar{\lambda}) \leq \mathcal{L}^*(\beta^*)$; that is $h(\bar{\beta}) \leq f(\beta^*)$, where $\bar{\beta}$ satisfies $\mathcal{L}_*(\bar{\lambda}) = \mathcal{L}(\bar{\beta}, \bar{\lambda})$.

An immediate consequence of this theorem is that $f(\beta^*) \geq h(\beta^*)$ for all solutions of (2). It is quite desirable that equality hold at solutions of (2), since this allows use of the *dual gap* $f(\beta) - h(\beta)$ to test for solution of (2). But by ?? of Nash and Sofer (1996), chapter 14.8, equality holds if and only if there is some point $(\beta*, \lambda^*)$ that satisfies the saddle-point condition

$$\mathcal{L}(\beta^*, \lambda) \leq \mathcal{L}(\beta^*, \lambda^*) \leq \mathcal{L}(\beta, \lambda^*)$$

for all $\beta \in \mathbb{R}^p$ and $\lambda \geq 0$. The next theorem shows that for this problem such points exist.

Theorem 3. (Strong duality). If β^* is a solution of (2) and λ^* is the Lagrange multiplier corresponding to $\beta*$ (i.e. $\lambda^* = r'(\beta^*)\mathbf{X}\beta^*/\|\beta^*\|_1$), then λ^* is a solution of the dual problem (3) and $\mathcal{L}_* = \mathcal{L}(\beta*, \lambda^*)$. It follows that the optimal primal and dual function values are equal; that is, $h(\beta^*) = f(\beta^*)$.

Proof. Based on results of Osborne (1985).

The rest of Osborne, Presnell, and Turlach (2000a):

- Section 3: characteristics of solutions (tests for uniquness and number of non-zero coefficients).
- Section 4: Standard errors of lasso estimates.
- Section 5: Algorithms (Remark 9 on page 331 seems to indicate the direction pursued by Efron et al. in developing LARS); this is connected to Osborne, Presnell, and Turlach (2000b).
- Smooth approximations of lasso: replace $|u|$ by an everywhere differentiable function $\rho_c(u)$ where $\rho_c(u) = |u|$ for $|u| \geq c$ or by $\sqrt{u^2 + c^2}$.