STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 1 (March 27): Introduction and brief overview

An important set of problems in statistics center around the problems involved in regression problems: suppose we observe

$$(X_1, Y_1), \ldots, (X_n, Y_n) \qquad \text{i.i.d. in } \mathbb{R}^{k+1}$$

where

$$\begin{aligned} Y_i &= \text{response for unit } \ i, \\ X_i &= \text{vector of covariates or predictors for unit } \ i. \end{aligned}$$

Suppose that k is "large", say $k > 50$, but potentially much larger. The usual linear model is

$$Y_i = X_i'\beta + \epsilon_i \qquad \text{or} \qquad \underline{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}$$

where ϵ_i is independent of X_i and $E(\epsilon_i) = 0$, $Var(\epsilon_i) < \infty$. The usual least squares estimator of β is

$$\begin{aligned} \hat{\beta} &= \text{argmin}_\beta \sum_1^n (Y_i - \underline{X_i'}\underline{\beta})^2 \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y} \\ &= \text{argmin}_\beta \mathbb{P}_n (Y - \underline{X}'\underline{\beta})^2. \end{aligned}$$

One direction for model selection: consider all models based on subsets of $\underline{X}$, i.e. $\{X_{i_1}, \ldots, X_{i_q}\}$, $q \geq 1$, where $1 \leq i_1 < \cdots < i_q \leq k$ are distinct. How many models do we get this way?

Example: With $k = 3$, so $\underline{X} = (X_1, X_2, X_3)$, there are $2^3 = 8$ possible models.

Another direction: enlarge the set of predictor variables by considering powers of the X_i's and interactions.

Example: $k = 3$, $\underline{X} = (X_1, X_2, X_3)$, and include quadratic terms and all first order interactions. This gives $\underline{Z} = (X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_2X_3)$, with $k = 9$, and we could consider models of the form

$$Y_i = \underline{Z_i'}\beta + \epsilon_i$$

where now have $k = 9$ and $2^k = 2^9 = 512$ submodels. If we do this for an initial $k = 100$, then $k' = 100 + 100 + \binom{100}{2} = 200 + 100 \cdot 99/2 = 200 + 50 \cdot 99 = 5150$, and then the number of sub-models is 2^{5150}, an enormous number. Moreover, $k' = 5150$ can easily be larger than n, and then $\mathbf{X}'\mathbf{X}$ becomes singular and $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist.

One solution to singularity of $\mathbf{X}'\mathbf{X}$: ridge regression (Hoerl and Kennard, 1970):

$$\text{minimize} \;\; \mathbb{P}_n(Y - \underline{X}'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|^2$$

where $\lambda > 0$, or, equivalently,

$$\text{minimize} \;\; \mathbb{P}_n(Y - \underline{X}'\beta)^2 \qquad \text{subject to} \;\; \sum_{j=1}^{k} |\beta_j|^2 \leq t,$$

where $t > 0$ is in a one-to-one correspondence with λ. The solution is

$$\underline{\hat{\beta}}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda I)^{-1}\mathbf{X}'Y.$$

A family of solutions: Alternatively, we can consider the following family of minimization problems: for $\gamma > 0$

$$\text{minimize} \;\; \mathbb{P}_n(Y - \underline{X}'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|^\gamma$$

where $\lambda > 0$, or, equivalently,

$$\text{minimize} \;\; \mathbb{P}_n(Y - \underline{X}'\beta)^2 \qquad \text{subject to} \;\; \sum_{j=1}^{k} |\beta_j|^\gamma \leq t,$$

where $t > 0$ is in a one-to-one correspondence with λ. The resulting $\widehat{\beta}_n = \widehat{\beta}_{n,\lambda,\gamma}$ are the *bridge estimators* of Frank and Friedman (1993). Here

- $\gamma = 2$ gives *ridge regression*;
- $\gamma = 1$ gives *the lasso* proposed by Tibshirani (1996);
- $\gamma \geq 1$: a criterion function which is convex in β_j's;
- $\gamma = 0$: in this case the criterion function becomes (in the limit as $\gamma \searrow 0$)

$$\begin{aligned} \text{minimize} \qquad & \mathbb{P}_n(Y - \underline{X}'\beta)^2 + \lambda \sum_{j=1}^{k} 1\{|\beta_j| > 0\} \\ & = \mathbb{P}_n(Y - \underline{X}'\beta)^2 + \lambda \#\{j \leq k : \ |\beta_j| > 0\} \end{aligned}$$

where $\lambda > 0$, or, equivalently,

$$\text{minimize} \ \ \mathbb{P}_n(Y - \underline{X}'\beta)^2 \qquad \text{subject to} \ \ \sum_{j=1}^{k} 1\{|\beta_j| > 0\} \leq t.$$

Because of the lack of convexity of the objective function in this case we are faced with combinatorial optimization problems, as opposed to convex optimization problems which occur if $\gamma \geq 1$. On the other hand, fixing λ at particular values gives rise to several well-known classical model selection methods such as AIC and BIC: when $\lambda = 2\sigma^2$ (or perhaps $2\widehat{\sigma}^2$) and $\gamma = 0$, then we are essentially considering AIC model selection; when $\lambda = (1/2)(\log n)\sigma^2$, then this gives BIC, or Schwarz's Bayesian Information Criterion.

We will begin by studying some of the classical model selection methods such as AIC, FPE, PSS, BIC and its generalization(s) GIC, all of which involve (more or less) fixed choices of λ (up to some data- dependent choice of scale). After building some understanding in the classical cases, we will study some of the newer methods based on the lasso and its variants.