STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 4 (April 5): This lecture continues with the paper of Shibata, and then discusses the results of Shao (1997).

Now suppose that

$$
\begin{aligned}
&\underline{Y} = (Y_1, \ldots, Y_n) \qquad \text{are independent,}\\
&\mathbf{X} \equiv \mathbf{X}_n = (\mathbf{x}_1', \ldots, \mathbf{x}_n') \qquad \text{a } \ n \times k_n \ \text{ matrix,}\\
&\underline{\mu}_n = E(\underline{Y}|\mathbf{X}_n),\\
&\underline{\mu}_n(m) = \mathbf{X}_n(m)\beta(m), \qquad m \in \mathcal{M}_{0,n},\\
&R_n^{pred}(m) = n^{-1}\|\mu_n - \widehat{\mu}_n(m)\|^2 = n^{-1}\|\mathbf{X}\beta = \mathbf{X}\widehat{\beta}(m)\|^2\\
&\qquad\qquad \neq E\{n^{-1}\|\mathbf{X}\beta = \mathbf{X}\widehat{\beta}(m)\|^2\}\\
&\qquad\qquad = n^{-1}\|\mu_n - H_n(m)\mu_n\|^2 + \frac{d(m)\sigma^2}{n}\\
&\qquad\qquad \equiv R_n(m) \equiv \Delta_n(m) + \frac{d(m)\sigma^2}{n},\\
&\widehat{\mu}_n(m) = \text{Least Squares Estimator of } \mu_n \text{ under model } m \in \mathcal{M}_n\\
&\qquad\qquad \equiv \mathbf{X}\widehat{\beta}(m).
\end{aligned}
$$

Note that *no normality assumption* has been imposed on the errors $Y_i = \mu_{n,i}$. Instead Shao imposes (later) moment conditions of the form $E(Y_1 - \mu_{n,1})^{4l} < \infty$ for some integer $l \geq 1$. Let

$$m_n^* \equiv \text{argmin}_{m \in \mathcal{M}_{0,n}} R_n^{pred}(m);$$

note that this is a random model depending on $\underline{Y}$ and β (or $\underline{\mu}_n$. Also let

$$\widehat{m}_n \equiv \text{a random } \ m \in \mathcal{M}_{0,n} \text{ depending only on the data.}$$

Shao's definition of a consistent model selection procedure is slightly different than Nishii's: Shao says that $\widehat{m}_n$ is *consistent* if

$$P(\widehat{m}_n = m_n^*) \to 1 \qquad \text{as } \ n \to \infty.$$

Furthermore, Shao says that a selection procedure $\widehat{m}_n$ is *Asymptotically Loss Efficient* if

$$\frac{R_n^{pred}(\widehat{m}_n)}{R_n^{pred}(m_n^*)} \to_p 1 \qquad \text{as } \ n \to \infty.$$

The following proposition connects these two definitions.

Proposition 1. Suppose that $k_n/n \to 0$, and that

$$\liminf_{n\to\infty} \min_{m\in\mathcal{M}_{1,n}} \Delta_n(m) > 0, \tag{1}$$

and $\mathcal{M}_{2,n} \neq \emptyset$ for sufficiently large n. Then consistency holds if and only if asymptotic loss efficiency holds if either $k_n(m_n^*) \not\to_p \infty$ or $\#(\mathcal{M}_{2,n}) = 1$ for n large.

Shao uses the following three examples to illustrate his results:

Example 1. (Linear regression). $\mu_n = \mathbf{X}_n\beta$ where $\mathbf{X}_n$ is $n \times k$, $\beta \in \mathbb{R}^k$. In a simple case $\mathcal{M}_n = \{m_1, m_2\}$ where $\mu_(m_1) = \mathbf{X}_{n,1}\beta_1$ and $\mu_n(m_2) = \mathbf{X}_n\beta$; here $\mathbf{X}_n = (\mathbf{X}_{n,1}, \mathbf{X}_{n,2})$ where $\mathbf{X}_{n,1}$ contains the first k_1 columns of $\mathbf{X}$, and $\beta = (\beta_1', \beta_2')'$. More generally $\mu_n(m) = \mathbf{X}_n(m)\beta_n(m)$ where $m \subset \{1, \ldots, k\}$.

Example 2. (One mean versus $k-$means) Suppose that $n = s\cdot r$ where $s = s_n$ is the number of groups and $r = r_n$ is the number of observations in each group. In this case

$$\mathbf{X}_n = \begin{pmatrix} \underline{1}_r & 0 & 0 & \cdots & 0 \\ \underline{1}_r & \underline{1}_r & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ \underline{1}_r & 0 & 0 & \cdots & \underline{1}_r \end{pmatrix},$$

$$\beta = (\mu_1, \mu_2 - \mu_1, \ldots, \mu_s - \mu_1)',$$

$$\mathcal{M}_{0,n} = \{m_1, m_s\}, \quad \text{with} \quad m_1 = \{1\},\ m_s = \{1, \ldots, s\}.$$

Example 3. (Linear approximation to response surface) This is very closely related to Example 2 of lecture 2. Now $\mu_n(m) = \mathbf{X}_n(m)\beta_n(m)$, $m \in \mathcal{M}_n$. One particular illustrative case is when the function to be estimated is based on a real-valued number t, we observe (Y_i, t_i), $i = 1, \ldots, n$, and we use the polynomial models $\theta_k(t) = \sum_{j=1}^k \beta_j t^{j-1}$ to approximate the true θ. Then

$$\mathbf{X}_n = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{k-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{k-1} \end{pmatrix},$$

and in this case $\mathcal{M}_{0,n} = \{m_d :\ d = 1, \ldots, k\}$ with $m_d = \{1, \ldots, d\}$, so the family of models under consideration is nested.

The GIC family of model selection methods:

Let

$$\Gamma_{n,\lambda_n} \equiv \widehat{\sigma}^2(m) + \frac{\lambda_n \widehat{\sigma}_n^2 k_n(m)}{n}$$

where

$$\widehat{\sigma}_n^2(m) \equiv n^{-1}\|\underline{Y} - \widehat{\mu}_n(m)\|^2 = n^{-1}\|\underline{Y}'(I - H(m))\underline{Y}\|^2,$$

$$\lambda_n \geq 2, \qquad \text{and} \qquad n^{-1}\lambda_n \to 0.$$

Shao calls $\widehat{m}_n \equiv \widehat{m}_{n\lambda_n}$ minimizing $\Gamma_{n,\lambda_n}(m)$ over $m \in \mathcal{M}_{0,n}$ as the *GIC* (λ_n) *model selection method*. He analyzes these methods in two primary cases: (i) $\lambda_n = 2$; (ii) $\lambda_n \to \infty$.

Here are some conditions:

Condition 1:

$$\sum_{m \in \mathcal{M}_{1,n}} \frac{1}{(nR_n(m))^l} \to 0 \tag{2}$$

where $l \geq 1$ is an integer such that $E(Y_1 - \mu_1)^{4l} < \infty$.

Condition 2: Condition 2(a) $\widehat{\sigma}_n^2 \to_p \sigma^2$.
Condition 2(b) $\widehat{\sigma}_n^2 = O_p(1)$ and $1/\widehat{\sigma}_n^2 = O_p(1)$.

Condition 3: $\sum_{m \in \mathcal{M}_{2,n}} d_n(m)^{-l} \to 0$ for some integer l with $E(Y_1 - \mu_1)^{4l} < \infty$.

Condition 4: $\sum_{m \in \mathcal{M}_{2,n} \setminus \{m_0\}} (d_n(m) - d_n(m_0))^{-l} \to 0$ for some l with $E(Y_1 - \mu_1)^{4l} < \infty$.

Condition 5:

Theorem 1. (GIC(2)) Suppose that conditions 1 and 2(a) hold.
(i) If $\#(\mathcal{M}_{2,n}) \leq 1$ for all n, then $\widehat{m}_2$ is asymptotically loss efficient. If $\mathcal{M}_{2,n}$ contains a unique model m_0 with $d_0 = d(m_0)$ fixed and finite, then $\widehat{m}_2$ is consistent.

(ii) Suppose that $\#(\mathcal{M}_{2,n}) > 1$ for all n large. If condition 3 or condition 4 holds, then $\widehat{m}_2$ is asymptotically loss efficient.

(iii) Suppose that $\#(\mathcal{M}_{2,n}) > 1$ for all n large and condition 5 holds. Then a necessary condition for $\widehat{m}_2$ to be asymptotically loss efficient is that $d_n(m_n^0) \to \infty$ or $\min_{m \in \mathcal{M}_{2,n} \setminus \{m_0\}} (d_n(m) - d_n(m_n^0)) \to \infty$.

(iv) If $\#(\mathcal{M}_{2,n}) < \infty$ as $n \to \infty$, or if $l = 2$ and $\mathcal{M}_{0,n} = \{m_d : d = 1, \ldots, k_n\}$ (i.e. the nested case), then $d_n(m_n^0) \to \infty$ or $\min_{m \in \mathcal{M}_{2,n} \setminus \{m_0\}} (d_n(m) - d_n(m_n^0)) \to \infty$ is sufficient for $\widehat{m}_2$ to be asymptotically loss efficient.

Theorem 2. (GIC(λ_n) with $\lambda_n \to \infty$) Suppose that conditions 1 and 2(b) hold.
(i) If (1) holds and both $\lambda_n \to \infty$ and $n^{-1}\lambda_n k_n \to 0$, then $\widehat{m}_{n,\lambda_n}$ is asymptotically loss efficient.

(ii) If $\mathcal{M}_{0,n}$ contains at least one correct model with fixed dimension for all n sufficiently large, $\lambda_n \to \infty$, and $n^{-1}\lambda_n \to 0$, then $\widehat{m}_{n,\lambda_n}$ is consistent.