STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 3 (April 3): This lecture focuses on the papers of Nishii (1984), Shibata (1984), and heads toward the paper by Shao (1997).

Suppose that

$$Y_i = \theta(X_i) + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\theta(x) = x'\beta$, with $x \in \mathbb{R}^k$, $\beta \in \mathbb{R}^k$, and $\epsilon_i \sim N(0, \sigma^2)$. Thus in vector form

$$\underline{Y} = \mathbf{X}\beta + \underline{\epsilon}.$$

Here

$$\begin{aligned}
\Theta_m &= \{x'\beta : \ \beta_{j_1} \neq 0, \ldots, \beta_{j_d} \neq 0, \ \beta_i = 0, \ i \notin \{j_1, \ldots, j_d\}\} \\
&= \Theta_{j_1,\ldots,j_d}, \qquad j = m = \{j_1, \ldots, j_d\}, \ \ 1 \le j_1 < \cdots < j_d \le k, \\
\mathcal{M} &= \cup_{d=1}^k \{\{j_1, \ldots, j_d\} : \ 1 \le j_1 < \cdots < j_d \le k\}, \qquad \text{and} \\
\#(\mathcal{M}) &= \sum_{d=}^{k} \#(\mathcal{M}_{k,d}) = \sum_{d=1}^{k} \binom{k}{d} = 2^k - 1.
\end{aligned}$$

Let D_j be a $k \times d$ matrix of 0's and 1's such that

$$XD_j = \text{columns } \ j_1, \ldots, j_d \ \text{ of } \ X.$$

Then for model m,

$$\underline{Y} = \mathbf{X}\beta(j) + \epsilon$$

where $\beta(j) = D_j D_j' \beta = D_j(\beta_{j_1}, \ldots, \beta_{j_d})'$.

Assumption 1: The true model if $m_0 = \{1, \ldots, d_0\}$ and $\mathcal{M}_0 \subset \mathcal{M}$ contains m_0; i.e. $m_0 \in \mathcal{M}_0$.

Assumption 2: $\mathbf{X}'\mathbf{X}$ is positive definite and $M = \lim_{n\to\infty}(\mathbf{X}'\mathbf{X})/n$ exists and is positive definite.

Assumption 2 implies that $\text{rank}(\mathbf{X}D_j) = d_j$; i.e. $D_j'\mathbf{X}'\mathbf{X}D_j$ is positive definite.

Define:

$$\begin{aligned}
&\widehat{\beta}(j) = D_j(D_j'\mathbf{X}'\mathbf{X}D_j)^{-1}D_j'\mathbf{X}'Y = \text{MLE of } \ \beta(j), \\
&H(j) = \mathbf{X}D_j(D_j(D_j'\mathbf{X}'\mathbf{X}D_j)^{-1}D_j'\mathbf{X}' = \Pi(\cdot|\text{columns } j_1, \ldots, j_d \ \text{ of } \mathbf{X}), \\
&\widehat{\sigma}^2(j) = n^{-1}\underline{Y}'(I - H(j))\underline{Y} = \ \ \text{MLE of } \ \ \sigma^2 \ \ \text{for model } j = n^{-1}\|\underline{Y} - H(j)\underline{Y}\|^2.
\end{aligned}$$

Then the AIC, C_p, FPE, PSS, and GIC criteria (defined by Akaike (1970), (1973); Mallows (1973); Akaike (1974); Allen (1971, 1974); and Schwarz (??) and Nishii (1984) respectively), are defined as follows:

$$
\begin{aligned}
AIC(j) &= \log \widehat{\sigma}^2(j) + a\frac{d(j)}{n}, \\
C_p(j) &= \frac{\widehat{\sigma}^2(j)}{\widehat{\sigma}^2(\{1,\ldots,k\})} + a\frac{(d(j)-1)}{n}, \\
FPE(j) &= \left(1 + a\frac{d(j)-1}{n}\right)\widehat{\sigma}^2(j), \\
PSS(j) &= n^{-1}\underline{Y}(I - H(j))(I - \Lambda(j))^{-2}(I - H(j))\underline{Y}, \\
GIC(j) &= \log \widehat{\sigma}^2(j) + a_n\frac{d(j)}{n}
\end{aligned}
$$

where a is a positive constant (usually $a = 2$),

$$\Lambda(j) = \mathrm{diag}(H(j)_{ii}),$$

and a_n satisfies $a_n \to \infty$, $n^{-1}a_n \to 0$. (For BIC, $a_n = (1/2)\log n$.) It seems that PSS corresponds to what is known now as "delete-1 cross - validation"; c.f. Hastie, Friedman, and Tibshirani (2001), pages 214 - 217.

Assessment criteria:

Criterion 1. Selection probabilities: $\{p_n(m) \equiv P(\widehat{m} = m),\ m \in \mathcal{M}\}$ with $P = P_{m_0}$.

Criterion 2. Prediction (error) risk:

$$
\begin{aligned}
R_n &= E\left\{\|\mathbf{X}\beta - \mathbf{X}\widehat{\beta}(\widehat{m})\|^2\right\} \\
&= \sum_{m\in\mathcal{M}} E\left\{\|\mathbf{X}\beta - \mathbf{X}\widehat{\beta}(\widehat{m})\|^2 1\{\widehat{m} = m\}\right\} \\
&\equiv \sum_{m\in\mathcal{M}} R_n(m).
\end{aligned}
$$

Note that with $\theta(x,\beta) = x'\beta$ and $\widehat{\theta}(x,\beta) = x'\widehat{\beta}$ we have

$$\|\widehat{\theta} - \theta\|^2_{L_2(\mu)} = \int (\widehat{\theta} - \theta)^2 d\overline{G}_n,$$

while, with $\mathbb{G}_n \equiv n^{-1}\sum_{i=1}^n \delta_{x_i}$,

$$n\int (\widehat{\theta} - \theta)^2 d\mathbb{G}_n = \sum_{i=1}^n [x_i'(\widehat{\beta} - \beta)]^2 = \|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta\|^2.$$

Thus the quantity R_n is basically n times the quantity $R_{n,m}(\theta, \widehat{\theta}_{n,m})$ we discussed in Lecture 2 when the x_i's are regarded as fixed and non-random.

Two key sub-collections of $\mathcal{M}$ are defined as follows:

$$\mathcal{M}_1 \equiv \{m \in \mathcal{M}_0 : \quad m_0 \not\subseteq m\},$$
$$\mathcal{M}_2 \equiv \{m \in \mathcal{M}_0 : \quad m_0 \subseteq m\}.$$

Nishii (1984) organizes his paper around the following two conditions for a selection rule:

Condition 1. $np_n(m) \to 0$ for $m \in \mathcal{M}_1$.
Condition 2. $p_n(m) \to 0$ for $m \in \mathcal{M}_2 \setminus \{m_0\}$.

Theorem 1.
(i) If Condition 1 holds, then $R_n(m) \to 0$ for $m \in \mathcal{M}_1$.
(ii) If Condition 2 holds, then $R_n(m) \to 0$ for $m \in \mathcal{M}_2 \setminus \{m_0\}$.

Corollary. If Conditions 1 and 2 hold, then $R_n \to d_0\sigma^2$ and $R_n(m_0) \to d_0\sigma^2$.

Theorem 2. (AIC, FPE, C_p). If assumptions 1 and 2 hold, then:
(i) (a) For every $h > 0$, $n^h p_n(m) \to 0$ for $m \in \mathcal{M}_1$.
(i) (b) For $m \in \mathcal{M}_2$, $p_n(m) \to Pr(Q_m^{(a)} \geq Q_l^{(a)}$ for all $l \in \mathcal{M}_2) \equiv p_a(m)$ where

$$Q_l \equiv Z' A_l Z, \qquad Z \sim N_{k-d_0}(0, I),$$
$$Q_l^{(a)} \equiv Q_l - a d_l^* = Q_l - a(d_l - d_{m_0}),$$
$$A_l = L_l (L_l' L_l)^{-1} L_l',$$
$$M^{1/2} D_l = \begin{matrix} d_0 \\ k - d_0 \end{matrix} \begin{pmatrix} * & * \\ 0 & L_l \end{pmatrix}$$

where $M = (M^{1/2})' M^{1/2}$ and $M^{1/2}$ is an upper-triangular matrix of order k.
(ii)

$$R_n \quad \to \quad \sigma^2 \left(d_0 + \sum_{m \in \mathcal{M}_2} E(Q_m 1\{Q_m^a \geq Q_l^a, \text{ for all } l \in \mathcal{M}_2\}) \right) \equiv R(a).$$

To treat the PSS criterion, we need the following additional assumption:

Assumption 3: Let $c_i^{(n)} \equiv H(\{1, \ldots, k\})_{ii}$, $i = 1, \ldots, n$. Then

$$max_{1 \leq i \leq n} c_i^{(n)} \to 0.$$

This assumption implies that

$$max_{1 \leq i \leq n} c_i^{(n)}(m) \to 0$$

where $c_i^{(n)}(m) \equiv H(m)_{ii}$ for $i = 1, \ldots, n$.

Theorem 3. (PSS) If Assumption 3 holds, then
(i) (a) For every $h > 0$, $n^h p_n(m) \to 0$ for $m \in \mathcal{M}_1$.
(i) (b) For $m \in \mathcal{M}_2$, $p_n(m) \to Pr(Q_m^{(2)} \geq Q_l^{(2)}$ for all $l \in \mathcal{M}_2) \equiv p_2(m)$ where is $p_a(m)$ is as defined in Theorem 2.
(ii) $R_n \to R(2)$.

Finally, here is the result for GIC.

Theorem 4. (GIC) If Assumption 3 holds, then
(i) (a) For every $h > 0$, $n^h p_n(m) \to 0$ for $m \in \mathcal{M}_1$.
(i) (b) For $m \in \mathcal{M}_2 \setminus \{m_0\}$, $p_n(m) \to 0$.
(ii) $R_n \to d_0\sigma^2$.

All these results in Nishii (1984) are under the assumption that k is fixed with $k \leq n$ and $n \to \infty$ with $n^{-1}\mathbf{X}'\mathbf{X} \to M$ and $m_0 \in \mathcal{M}_0$. On the other hand, Shibata (1981) assumes that $k = k_n \to \infty$, and does not impose the assumption that $m_0 \in \mathcal{M}_0$. Shibata also assumes, along with Nishii, that the x_i's are fixed, nonrandom (and hence $\mathbf{X}$ is also nonrandom).

Let

$$R_n(m) \equiv E\|\mathbf{X}\beta - \mathbf{X}\widehat{\beta}(m)\|^2 = \|\mathbf{X}\beta - \mathbf{X}\beta(m)\| + k_n(m)\sigma^2$$

where $\beta(m)$ is the projection of β onto the space spanned by the columns $m_1, \ldots m_{k_n(m)}$ of $\mathbf{X}$.

Assumption 1 (Shibata): For each $m \in \mathcal{M}_{0,n}$, $\text{rank}(\mathbf{X}'(m)\mathbf{X}(m)) = k_n(m) = o(n)$.

Assumption 2 (Shibata): For $0 < \delta < 1$, $\sum_{m \in \mathcal{M}_{0,n}} \delta^{R_n(m)} \to 0$ as $n \to \infty$.

Suppose that $\widehat{m} \in \mathcal{M}_{0,n}$ is chosen via Akaike's FPE criterion; i.e.

$$\widehat{m} = \text{argmin}_{m \in \mathcal{M}_{0,n}} \left\{ 1 + \frac{2k_n(m)}{n} \right\} \widehat{\sigma}^2(m),$$

and suppose that m_n^* is the "optimal" or "best" model from the perspective of the risk $R_n(m)$ over $\mathcal{M}_{0,n}$; i.e.

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

Then Shibata (1981) proves the following theorem:

Theorem 1. If A1 and A2 hold, then

$$\frac{\|\mathbf{X}\beta - \mathbf{X}\widehat{\beta}(\widehat{m})\|^2}{R_n(m_n^*)} \to_p 1.$$

Shibata also proves an interesting lower bound result under weaker hypotheses:
Assumption 3 (Shibata): For each $m \in \mathcal{M}_{0,n}$, $k(m) \leq n$ and $\text{rank}(\mathbf{X}'(m)\mathbf{X}(m)) = k_n(m)$.
Assumption 4 (Shibata): Let $\alpha_n(m) \equiv R_n(m)/\{k(m)\sigma^2\}$ for $m \in \mathcal{M}_{0,n}$. Then for $0 < \delta < 1$,

$$\sum_{m \in \mathcal{M}_{0,n} : \delta\alpha_n(m) < 1} \{(1 - \delta\alpha_n(m)) \exp(\delta\alpha_n(m))\}^{k(m)/2} \to 0 \qquad \text{as} \quad n \to \infty.$$

Theorem 2. If A3 and A4 (δ) hold, then for *any* $\widetilde{m} \in \mathcal{M}_{0,n}$ depending possibly on the data,

$$Pr\left(\frac{\|\mathbf{X}\beta - \mathbf{X}\widehat{\beta}(\widetilde{m})\|^2}{R_n(m_n^*)} > 1 - \delta \right) \to 1.$$

Proof. From the definition of m_n^* it follows that

$$\begin{aligned}
&P\left\{\min_{m\in\mathcal{M}_{0,n}} \|\mathbf{X}\widehat{\beta}(m)-\mathbf{X}\beta\|^2/R_n(m_n^*)\le 1-\delta\right\}\\
&\le\ P\left\{\min_{m\in\mathcal{M}_{0,n}} \|\mathbf{X}\widehat{\beta}(m)-\mathbf{X}\beta\|^2/R_n(m)\le 1-\delta\right\}\\
&\le\ \sum_{m\in\mathcal{M}_{0,n}} P\left\{\frac{\|\mathbf{X}\widehat{\beta}(m)-\mathbf{X}\beta^n(m)\|^2}{\sigma^2}+\frac{\|\mathbf{X}\beta^{(n)}(m)-\mathbf{X}\beta\|^2}{\sigma^2}\le(1-\delta)R_n(m)/\sigma^2\right\}\\
&=\ \sum_{m\in\mathcal{M}_{0,n}} P\left\{\frac{\|\mathbf{X}\widehat{\beta}(m)-\mathbf{X}\beta^n(m)\|^2}{\sigma^2}\le k(m)-\delta R_n(m)/\sigma^2\right\}\\
&\le\ \sum_{m:k(m)>\delta R_n(m)/\sigma^2} \{(1-\delta\alpha_n(m))\exp(\delta\alpha_n(m))\}^{k(m)/2}
\end{aligned}$$

where the last inequality uses Lemma 2.1.

Lemma 2.1. Let $\chi_k^2\sim$ Chi-square with $k-$degrees of freedom. Then for any $\delta>0$,

$$\begin{aligned}
&P(\chi_k^2\le k-\delta)\le(1-\delta/k)^k\exp(\delta/2)\le\begin{cases}\exp\left(-\frac{\delta^2}{4k}\right), & k>\delta\\ 0, & k\le\delta,\end{cases}\\
&P(\chi_k^2\ge k+\delta)\le(1+\delta/k)^{k/2}\exp(-\delta/2)\le\exp(-\delta/4).
\end{aligned}$$

When do Shibata's Assumptions 2 and 4 hold? Shibata devotes Sections 3 and 4 of his paper to giving conditions which imply Assumption 2 in two important cases:

Example 1. (Nested models). Suppose that

$$\mathcal{M}_{n,0}=\{\{1\},\{1,2\},\{1,2,3\},\ldots,\{1,\ldots,k_n\}\}\equiv\{m_1,\ldots,m_{k_n}\}.$$

In this case, Assumption 2 can be replaced by one of the following conditions:

Condition 1: There exists a divergent sequence $\{r_n\}$ such that $r_n\le k_n$ and $\log r_n=o(\|\mathbf{X}\beta-\mathbf{X}\beta(r_n)\|^2)$.

To see that Assumption 2 follows from this condition, note that for any $0<\delta<1$, and

using $R_n(m_k) \geq k\sigma^2$,

$$\begin{aligned}
\sum_{k=1}^{k_n} \delta^{R_n(m)} &= \sum_{k=1}^{k_n} \exp(-R_n(k)\log(1/\delta)) \\
&= \sum_{k=1}^{r_n} \exp(-R_n(k)\log(1/\delta)) + \sum_{k=r_n+1}^{k_n} \exp(-R_n(k)\log(1/\delta)) \\
&\leq \sum_{k=1}^{r_n} \exp(-\|\mathbf{X}\beta - \mathbf{X}\beta(k)\|^2 \log(1/\delta)) \\
&\qquad + \exp(-r_n\sigma^2\log(1/\delta))/(1-\exp(\sigma^2\log\delta)) \\
&\leq r_n \exp(-\|\mathbf{X}\beta - \mathbf{X}\beta(r_n)\|^2 \log(1/\delta)) \\
&\qquad + \exp(-r_n\sigma^2\log(1/\delta))/(1-\exp(\sigma^2\log\delta)) \\
&= \exp(o(1)\|\mathbf{X}\beta - \mathbf{X}\beta(r_n)\|^2) \exp(-\log(1/\delta)\|\mathbf{X}\beta - \mathbf{X}\beta(r_n)\|^2) \\
&\qquad + \exp(-r_n\sigma^2\log(1/\delta))/(1-\exp(\sigma^2\log\delta)) \\
&= \exp(-(\log(1/\delta) - o(1))\|\mathbf{X}\beta - \mathbf{X}\beta(r_n)\|^2) \\
&\qquad + \exp(-r_n\sigma^2\log(1/\delta))/(1-\exp(\sigma^2\log\delta))
\end{aligned}$$

Condition 2: The sequence $\{k_n\}$ satisfies $k_n \to \infty$ and $\|\mathbf{X}\beta - \mathbf{X}\beta(k)\|^2 \to \infty$ for any fixed k.

Theorem 3. Suppose there exists a sequence $\{c_n\}$ with $c_n \to \infty$ such that $c_n^{-1}\mathbf{X}'\mathbf{X} \to M$ in the sense of the operator norm for maps from ℓ_2 to ℓ_2 whose $k \times k$ principal submatrices $M(k)$ have full rank for any $k > 0$. If β does not depend on n and has infinitely many non-zero coordinates, then for any fixed $k > 0$,

$$C^{-1} \leq \|\mathbf{X}\beta - \mathbf{X}\beta^{(n)}(k)\|^2/c_n \leq C$$

for some $0 < C < \infty$, and $m_n^* \to \infty$ as $n \to \infty$. Condition 2 is satisfied when $k_n \to \infty$.