STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 2 (March 29): General Notation and Some Examples

Here is some notation and terminology that I will try to use (more or less) systematically throughout the course.

Suppose that we have:

- **Data:** $\xi^{(n)} \sim P_{n,\theta} = P_\theta$ for $\theta \in \Theta$; $\xi^{(n)} \in \Xi_n$.
- **Parameter space:** Θ. Sometimes $\Theta \subset \mathbb{R}^k$ for some k; often Θ is some (large) collection of functions.
- **A sieve (or finite-dimensional subsets of Θ):** $\{\Theta_m : \ m \in \mathcal{M}_n\}$ where $\Theta_m \subset \Theta$ with $\dim(\Theta_m) = D_m < \infty$.
- **A collection of models:** $\{\mathcal{P}_m : \ m \in \mathcal{M}_n\}$, with $\mathcal{P}_m = \{P_\theta : \ \theta \in \Theta_m\}$.
- **An empirical contrast function:** $\gamma_n : \Theta \times \Xi_n \to \mathbb{R}$, $\gamma_n(\theta, \xi^{(n)}) \equiv \gamma_n(\theta)$, $\theta \in \Theta$.
- **Empirical contrast estimators:** $\widehat{\theta}_{n,m} = \mathrm{argmin}_{\theta \in \Theta_m} \gamma_n(\theta)$ for $m \in \mathcal{M}_n$.
- **Risk functions:** $R_{n,m}(\theta, \widehat{\theta}_{n,m}) = E_\theta \left[d^2(\theta, \widehat{\theta}_{n,m}) \right] =$ risk of $\widehat{\theta}_{n,m}$ at θ for $m \in \mathcal{M}_n$ and $\theta \in \Theta$.

Here are several examples:

Example 1. (density estimation) Suppose that $\xi^{(n)} = (X_1, \ldots, X_n)$ where X_i are i.i.d. p_θ with respect to a dominating measure μ. Here are two common choices for a contrast function γ:

- **Maximum likelihood:** $\gamma_n(\theta) = \mathbb{P}_n \gamma(X, \theta)$ with $\gamma(x, \theta) = -\log p_\theta(x)$. Then

$$\begin{aligned} \widehat{\theta}_{n,m} &= \mathrm{argmin}_{\theta \in \Theta_m} \gamma_n(\theta) \\ &= \text{maximum likelihood estimator of } \ \theta \ \text{ over } \Theta_{n,m}. \end{aligned}$$

- **Least squares:** In this case we take

$$\gamma(x, \theta) = \int p_\theta^2 d\mu - 2 p_\theta(x),$$

so that

$$\begin{aligned} \gamma_n(\theta) \quad &= \quad \mathbb{P}_n \gamma(X, \theta) = \int p_\theta^2 d\mu - 2 \int p_\theta d\mathbb{P}_n \\ \text{“}=\text{”} \quad &\int (p_\theta - p_n)^2 d\mu - \int p_n^2 d\mu \qquad \text{if } p_n = d\mathbb{P}_n/d\mu \ \text{ exists.} \end{aligned}$$

Example 2. (regression) In this case $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ where $\xi_i = (X_i, Y_i)$ where

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

$X_i \sim G_i$ are independent, ϵ_i are independent with $E(\epsilon_i | X_i) = 0$ for $1 \le i \le n$. If $\mu \equiv \overline{G}_n = n^{-1} \sum_{i=1}^n G_i$, then for $\theta \in L_2(\mu)$ set

$$\gamma(\xi, \theta) = \gamma((x, y), \theta) = (y - \theta(x))^2.$$

Then

$$\widehat{\theta}_{n,m} = \text{argmin}_{\theta \in \Theta_{n,m}} \gamma_n(\theta)$$

is the least squares estimator of θ over Θ_m.

Example 3. (binary classification) In this case $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$ where $\xi_i = (X_i, Y_i)$ where $Y_i \in \{0, 1\}$,

$$\gamma(\xi, \theta) = \gamma((x, y), \theta) = (y - f_\theta(x))^2.$$

as in Example 2, $\theta(x) \equiv E(Y | X = x)$, and $f_\theta(x) = 1\{\theta(x) \ge 1/2\}$.

Example 4. (Gaussian white noise model) Let $\xi^{(n)}$ be the process on $[0, 1]^d$ defined by

$$d\xi^{(n)}(x) = \theta(x) dx + \frac{\sigma}{\sqrt{n}} dW(x),$$

where W is a Brownian sheet on $[0, 1]^d$ (i.e. a mean zero Gaussian process with $EW(x)W(y) = |x \wedge y| \equiv \prod_{j=1}^d (x_j \wedge y_j)$). For $\theta \in L_2([0, 1]^d, \lambda)$, define

$$\gamma_n(\theta) = \|\theta\|^2 - 2 \int_0^1 \theta(x) d\xi^{(n)}(x).$$

Then

$$\widehat{\theta}_{n,m} = \text{argmin}_{\theta \in \Theta_m} \gamma_n(\theta)$$

is the MLE of θ over Θ_m and is also the least squares estimator of θ over Θ_m.

Here is a natural choice of d^2 for each of these problems:

$$d^2(\theta, \theta') = E_\theta \gamma_n(\theta') - E_\theta \gamma_n(\theta).$$

In Example 1 with the maximum likelihood empirical contrast function,

$$d^2(\theta, \theta') = \int p_\theta \log(p_\theta / p_{\theta'}) d\mu = K(P_\theta, P_{\theta'}).$$

For the Least squares empirical contrast function,

$$d^2(\theta, \theta') = \int (p_{\theta'} - p_\theta)^2 d\mu = \|p_{\theta'} - p_\theta\|^2_{L_2(\mu)}.$$

In Example 2, $d^2(\theta, \theta') = \|\theta - \theta'\|^2 \equiv \|\theta - \theta'\|^2_{L_2(\mu)}$. In Example 3, we find, after some computation

$$d^2(\theta, \theta') = E_\theta\{|2\theta(X) - 1| |f_{\theta'}(X) - f_\theta(X)|\}.$$

In Example 4, $d^2(\theta,\theta') = \|\theta-\theta'\|^2_{L_2(\lambda)}$.

Now we continue with some computations for the white noise model, Example 4. Suppose that

$$\Theta_m = [\phi_1,\ldots,\phi_m] = \text{linear span of } \phi_1,\ldots,\phi_D \qquad \text{in } \; L_2([0,1]^d,\lambda)$$

where $D = D_m$ and $\{\phi_j\}_{j=1}^\infty$ is an orthonormal basis for $L_2([0,1]^d)$. Then the least squares estimator $\widehat{\theta}_{n,m}$ is given by

$$\widehat{\theta}_{n,m}(x) = \sum_{j=1}^{D}\left(\int_0^1 \phi_j d\xi^{(n)}\right)\phi_j(x);$$

note that

$$\int_0^1 \phi_j d\xi^{(n)} = \int_0^1 \phi_j\theta d\lambda + \frac{\sigma}{\sqrt{n}}Z_j, \qquad j\in\{1,\ldots,D\}$$

where Z_j are i.i.d. $N(0,1)$. Note that this can be re-written as a "Gaussian sequence model":

$$Y_j = \mu_j + \epsilon_j, \qquad j\in\{1,\ldots,D\}$$

where $Y_j = \int_0^1 \phi_j d\xi^{(n)}$, $\mu_j = \int_0^1 \phi_j\theta d\lambda$, and $\epsilon_j = \sigma Z_j/\sqrt{n} \sim N(0,\sigma^2/n)$, $j=1,\ldots,D$. Thus

$$\theta - \widehat{\theta}_{n,m} = \sum_{j=D+1}^{\infty}\left(\int \phi_j\theta d\lambda\right)\phi_j + \frac{\sigma}{\sqrt{n}}\sum_{j=1}^{D}\phi_j\int\phi_j dW,$$

so

$$\|\theta-\widehat{\theta}_{n,m}\|^2 = \sum_{j=D+1}^{\infty}\left(\int\phi_j\theta d\lambda\right)^2 + \frac{\sigma^2}{n}\sum_{j=1}^{D}Z_j^2,$$

and hence

$$\begin{aligned} R(\theta,\widehat{\theta}_{n,m}) &= E_\theta\{\|\theta-\widehat{\theta}_{n,m}\|^2\} = \|\theta-\Pi(\theta|\Theta_m)\|^2 + \frac{\sigma^2 D}{n} \\ &= \min_{\theta'\in\Theta_m}\|\theta-\theta'\|^2 + \frac{\sigma^2 D}{n} \equiv \|\theta-\theta_m\|^2 + \frac{\sigma^2 D}{n}. \end{aligned}$$

This is a classical formula involving a bias versus variance trade-off via the choice of D: increasing D leads to smaller bias but larger variance.

Model selection via penalization: Consider

$$\gamma_n(\widehat{\theta}_{n,m}) + \text{pen}(m), \qquad m\in\mathcal{M}_n. \tag{1}$$

For model selection via Mallows C_p or AIC, $\text{pen}(m) = 2D_m\sigma^2/n$ assuming that σ^2 is known. (If σ^2 is unknown, then we should estimate it using a low-bias model.) Then we choose $\hat{m}\in\mathcal{M}_n$ to minimize the penalized contrast function in (1): i.e.

$$\widehat{m} = \text{argmin}_{m\in\mathcal{M}_n}\left\{\gamma_n(\widehat{\theta}_{n,m}) + \text{pen}(m)\right\},$$

and then $\widehat{\Theta} \equiv \Theta_{\widehat{m}}$ and $\widehat{\theta} \equiv \widehat{\theta}_{\widehat{m}}$.

Heuristics for Mallows C_p penalty:

One way to proceed: an ideal model m^* should minimize the quadratic risk we calculated above:

$$m^* = \text{argmin}_{m \in \mathcal{M}_n} \left\{ \|\theta - \theta_m\|^2 + \frac{\sigma^2 D_m}{n} \right\} = \text{argmin}_{m \in \mathcal{M}_n} \left\{ \|\theta\|^2 - \|\theta_m\|^2 + \frac{\sigma^2 D_m}{n} \right\},$$

or, equivalently, minimize

$$-\|\theta_m\|^2 + \frac{\sigma^2 D_m}{n}.$$

This depends on the true θ through $\theta_m = \Pi(\theta|\Theta_m)$. But we can estimate $\|\theta_m\|^2$ by its natural unbiased estimator, namely

$$\|\widehat{\theta}_{n,m}\|^2 - \frac{\sigma^2 D_m}{n}.$$

Thus we can choose $\hat{m}$ by minimizing

$$-\|\widehat{\theta}_{n,m}\|^2 + \frac{2 D_m \sigma^2}{n}. \tag{2}$$

This is (very nearly) equivalent to minimizing

$$\gamma_n(\widehat{\theta}_{n,m}) + \frac{2\sigma^2 D_m}{n}. \tag{3}$$

General version of heuristics:

The following is from the introduction of Barron, Birgé, and Massart (1999).

An "ideal model" might be taken to be one that

$$\text{minimizes} \quad R_{n,m}(\theta, \widehat{\theta}_{n,m}) \qquad \text{over} \quad m \in \mathcal{M}_n.$$

(However, even if $\theta \in \Theta_{m_0}$ is true, this true model be far from the "ideal" model.)

Since θ is unknown, we cannot determine such an ideal model exactly.

Goal 1: Find $\widehat{m} \in \mathcal{M}_n$ based on the data, such that

$$R_{n,m}(\theta, \widehat{\theta}_{n,\widehat{m}}) \asymp \inf_{m \in \mathcal{M}_n} R_{n,m}(\theta, \widehat{\theta}_{n,m}) \equiv \text{minimal risk}.$$

Unfortunately, goal 1 is too hard in most problems. Instead, consider replacing the target of minimal risk by some appropriate "accuracy index"

$$\begin{aligned} a_n(\theta) &= \inf_{m \in \mathcal{M}_n} \{ d^2(\theta, \Theta_m) + pen_n(m) \} \\ &= \inf_{m \in \mathcal{M}_n} \{ \inf_{\theta' \in \Theta_m} d^2(\theta, \theta') + pen_n(m) \} \\ &\geq \text{minimal risk}. \end{aligned}$$

Goal 2: Find $\widehat{m} \in \mathcal{M}_n$ based on the data, such that

$$E_\theta d^2(\theta, \widehat{\theta}_{n,\widehat{m}}) \leq C(\theta) a_n(\theta) \qquad \text{for all } \ n.$$

One way to do this: choose

$$\widehat{m} = \text{argmin}_{m \in \mathcal{M}_n} \left\{ \gamma_n(\widehat{\theta}_{n,m}) + pen_n(m) \right\};$$

typically

$$pen_n(m) = \begin{cases} \frac{\kappa L_m D_m}{n}, & L_m = \text{“weights”} \\ K_m(\xi^{(n)}) \frac{L_m D_m}{n}, & L_m = \text{“weights”}, \quad K_m \ \text{ a function of the data} \end{cases}$$

where $\sum_{m \in \mathcal{M}_n} \exp(-L_m D_m) \leq 1$. Then the results of Birgé and Massart (1997), (2001) and Barron, Birgé and Massart (1999) are of the form

$$E_\theta d^2(\theta, \widehat{\theta}_{n,\widehat{m}}) \leq C(\theta) \inf_{m \in \mathcal{M}_n} \left\{ d^2(\theta, \Theta_m) + \frac{\kappa L_m D_m}{n} \right\}.$$

We will return to results of this type later in the quarter.