STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 11 (May 1): In this lecture we will continue with the results of Greenshtein and Ritov (2004) and Greenshtein (2006).

It seems that Theorem 1 of Greenshtein and Ritov (2004) might really be two different theorems as follows:

Theorem 1A. If condition 1 holds and $B_{n,b_n} \subset \mathbb{R}^p$ with $b_n = o((n/\log n)^{1/4})$, there exists a persistent sequence of procedures $\hat{\beta}_n$ in the sense that

$$\begin{aligned} &\inf_{P_n \in \mathcal{P}_n} \inf_{\beta \in B_{n,b_n}} Pr_{P_n}\left(|L_{P_n}(\beta) - L_{\mathbb{P}_n}(\beta)| \le |\gamma|' \hat{E} |\gamma|\right) \\ &= \inf_{P_n \in \mathcal{P}_n} \inf_{\beta \in B_{n,b_n}} Pr_{P_n}\left(|L_{P_n}(\beta) - L_{\mathbb{P}_n}(\beta)| \le \sqrt{An^{-1}\log n}\|\gamma\|_1^2\right) \to 1. \end{aligned}$$

Theorem 1B. If $p_n = n^\alpha$ and

$$F(Z_i) \equiv \max_{0 \le j,k \le p} |X_j^i X_k^i - E(X_j^i X_k^i)|$$

satisfies $E_{P_n} F^2(Z_1) \le M < \infty$ for all $n \ge 1$, then

$$\widehat{\beta}_n \equiv \text{argmin}_{\beta \in B_{n,b_n}} L_{\mathbb{P}_n}(\beta) \tag{1}$$

is persistent with respect to B_{n,b_n}.

When does the envelope condition of Theorem 1B hold? The following condition is intended to provide a sufficient condition:

Condition 3′: There are constants C and L such that $\sup_n E_{P_n} Y^4 \le C$ for all $P_n \in \mathcal{P}_n$ and $|X_j| \le L$ with probability 1, $j = 1, \ldots, p = p_n$. (This differs slightly from the Condition 3 of Greenshtein and Ritov in which they impose only a uniform bound on the second moments of Y.)

Theorem 3. If condition 3′ holds, then for any sequence $\{B_{n,b_n}\} \subset \mathbb{R}^p$ with $b_n = o((n/\log n)^{1/4})$ there exists a persistent sequence of procedures. In particular, $\widehat{\beta}_n$ defined by (1) is persistent.

Proof. Note that if all but a fixed number p_0 of $X_0, \ldots, X_p$ are bounded by some constant L and all the other X_j's have uniformly bounded fourth moments, then the collection $\{X_j X_k :$

$0 \le j,k \le p\}$ has an envelope function $F(Z)$ which satisfies $E_{P_n}^2 F^2(Z) \le M$ for some $M < \infty$. For example, if condition 3′ holds, then

$$F(Z) = L^2 \wedge LX_0 \wedge X_0^2 = L^2 \wedge LY \wedge Y^2 \ge \max_{0 \le j,k \le p} |X_j X_k|$$

satisfies

$$\sup_n E_{P_n} F^2(Z) \le \text{some } M < \infty \qquad \text{if and only if} \qquad \sup_n E_{P_n} Y^4 \le \text{some} M.$$

The result then follows from Theorem 1.B. □

Theorem 4. Suppose that $\{X_j X_k,\ 0 \le j,k \le p\}$ has an envelope function with a uniformly bounded second moment under $P_n \in \mathcal{P}_n$, $n \ge 1$. Suppose that condition 2 holds. Then there exists a method which is persistent with respect to the "variable selection sets" B_{n,k_n} with $k_n = o((n/\log n)^{1/2})$.

Theorem 4 follows from Theorem 3 and in the same way as theorem 2 follows from theorem 1.

On the other hand, section 3 of G-R (2004) concerns the case in which $(Y, \underline{X}) \equiv Z \sim N_{p+1}(\mu_Z, \Sigma_Z)$ and concerns sets of predictors B_{n,k_n} with $k_n = o(n/\log n)$.

Condition 4: Suppose that the sets $\mathcal{P}_n$ consist of all multivariate normal distributions of $Z = (Y, \underline{X})$ with uniformly bounded variance of Y.

Theorem 5. If $p_n = n^\alpha$ with $\alpha > 1$, then if $k_n = o(n/\log n)$ there exists a persistent sequence of procedures with respect to B_{n,k_n}.

The proof of Theorem 5 involves rather different techniques than the proofs of Theorems 1-4; in particular it relies on results for the smallest eigenvalue of a Wishart matrix.

Moreover, Greenshtein and Ritov establish the following negative result in the Gaussian setting.

Theorem 6. Suppose that $p = n^\alpha$, $\alpha > 1$. If $k_n > c(n/\log n)$ with $c > 0$, then there exists no procedure which is persistent with respect to the corresponding B_{n,k_n}.

The proof of Theorem 6 used Fano's lemma or inequality; see e.g. Le Cam (1986), page 524; Yu (1997), page 429; or Ibragimov and Has'minskii (1981), page 323.

Section 4 of G-R (2004) concerns the computational complexity of persistent procedures. Along the way in this section one of their key lemmas is:

Lemma 4. Let $Z = (Y, X_1, \ldots, X_n) \sim P$ be a random vector with $E_P Y^2 < \infty$, and $|X_j| \le c$, $j = 1, \ldots, p$ with probability 1. Then for any $\beta \in \mathbb{R}^p$ with $\|\beta\|_1 = \nu$, there exists a vector β' such that $\#\{j \le p :\ \beta'_j \ne 0\} \le k$ and

$$L_P(\beta') \le L_P(\beta) + \frac{c^2 \nu^2}{k}.$$

Proof. Suppose first that the coordinates of β_j all positive. Set $p_j = \beta_j/\|\beta\|_1 = \beta_j/\nu$, for $j = 1, \ldots, p$. Now let $\hat{p}_j = N_j/k$ where $\underline{N} = (N_1, \ldots, N_p) \sim \text{Mult}_p(k, \underline{p})$. Note that $\hat{p} = (\hat{p_1}, \ldots, \hat{p}_p)$ has at most k non-zero entries. We will show that

$$EL_P(\nu\underline{\hat{p}}) \leq L_P(\beta) + \frac{c^2\nu^2}{k},$$

where the expectation is over the randomness in $\underline{\hat{p}}$, and this proves the claim. Let $Z = (Y, Z_1, \ldots, X_p)$ be independent of $\underline{\hat{p}}$. Now

$$\begin{aligned}
EL_P(\nu\underline{\hat{p}}) &= E(Y - \sum_{j=1}^{p} \nu\hat{p}_j X_j)^2 \\
&= E(Y - \sum_{j=1}^{p} \nu p_j X_j + \sum_{j=1}^{p} \nu p_j X_j - \sum_{j=1}^{p} \nu\hat{p}_j X_j)^2 \\
&= E(Y - \sum_{j=1}^{p} \nu p_j X_j)^2 + E\left(\sum_{j=1}^{p} \nu X_j(p_j - \hat{p}_j)\right)^2 \\
&\qquad +2E\left(Y - \sum_{j=1}^{p} \nu p_j X_j\right)\left(\sum_{j=1}^{p} \nu X_j(p_j - \hat{p}_j)\right) \\
&= L_P(\beta) + E\left(\sum_{j=1}^{p} \nu X_j(p_j - \hat{p}_j)\right)^2 \\
&\leq L_P(\beta) + \nu^2 c^2 \sum_{j=1}^{p} Var(\hat{p}_j), \qquad \text{since } \; Cov(\hat{p}_j, \hat{p}_k) \leq 0, \;\; j \neq k \\
&= L_P(\beta) + \nu^2 c^2 \sum_{j=1}^{p} \frac{p_j(1-p_j)}{k} \\
&\leq L_P(\beta) + \frac{\nu^2 c^2}{k},
\end{aligned}$$

proving the claim when all β_j's are non-negative. The argument is easily modified to handled the case of β_j's possibly < 0. □