STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 15 (May 15): In this lecture I will finish discussing the paper by Candès and Tao (2007) on *the Dantzig selector*, and then attempt to survey some of the related contemporaneous and further work. This centers on the papers by Meinshausen and Bühlmann (2006), Zhang and Huang (2006), Huang, Ma, and Zhang (2006), Zhao and Yu (2006), and Meinshausen (2006).

A brief survey of papers on the lasso and Dantzig selector:

- **Meinshausen and Buhlmann (2006)** study Gaussian graphical model selection methods via the lasso.
- **Meinshausen (2006)** introduces and studies the "relaxed lasso", another possible two-stage procedure which uses the lasso at the first stage and then introduces another parameter ϕ to then reduce the penalty in a second stage procedure.
- **Huang, Ma, and Zhang (2006)** study Zou's adaptive lasso when $\log p_n = O(n^a)$ with $0 < a < 1$; the main difficulty is in construction of zero consistent initial or preliminary estimators.
- **Leng, Lin, and Wahba (2006)** show that the lasso is not model selection consistent when λ_n is chosen to minimize prediction error.
- **Zhang and Huang (2006)** study the lasso estimator under $p_n = n^\alpha$ with $\alpha > 1$ under a sparsity condition and a "partial Riesz condition" which is closely related to the "restricted isometry condition" of Candès and Tao (2007).
- **Zhao and Yu (2006)** show that the lasso estimator is model selection consistent if (and almost only if) the predictors no in the model are *irrepresentable* in a sense which they define.
- **Fan and Lv (2006)** study two stage model selection methods for the case of "ultra-high dimension", $\log p_n = O(n^a)$ with $0 < a < \infty$. Note that in this case the "cost of model selection" $2 \log p_n$ in the theorems of Candès and Tao can be quite large.

More on Zhang and Huang (2006):

Suppose that $\widehat{\beta} \equiv \widehat{\beta}(\lambda)$ is the lasso estimator of β in the usual linear model,

$$\widehat{\beta} = \mathrm{argmin}_{\tilde{\beta}} \left\{ \frac{1}{2} \|Y - \mathbf{X}\tilde{\beta}\|_2^2 + \lambda \|\tilde{\beta}\|_1 \right\},$$

and let

$$\widehat{A} \equiv \widehat{A}(\lambda) = \{j \leq p : \ \widehat{\beta}_j \neq 0\}.$$

Huang and Zhang (2006) assume throughout that $\mathbf{X}$ has columns $\mathbf{x}_j$ normalized so that $\|\mathbf{x}_j\|_2^2/n \asymp 1$.

Sparsity condition: There exists a set $A_0 \subset \{1, \ldots, p\}$ such that

$$\#(A_0^c) = q, \qquad \sum_{j \in A_0} |\beta_j| \leq \eta_1.$$

here A_0 is the set of "small β_j's". Thus with $|\beta|_{(1)} \geq \cdots \geq |\beta|_{(p)}$ we can write

$$\sum_{j=q+1}^{p} |\beta|_{(j)} \leq \eta_1, \qquad A_0 = \{(q+1), \ldots, (p)\}.$$

Here are several measures of how well the lasso (or any other procedure) performs:

- $\widehat{q} \equiv \#(\widehat{A})$, the number of non-zero $\widehat{\beta}_j$'s included in the model.
- $\tilde{B} = \tilde{B}(\lambda) \equiv \|(I - \widehat{P})\mathbf{X}\beta\|_2$ where $\widehat{P} = \widehat{P}_{\widehat{A}} = \Pi(\cdot|[\mathbf{x}_j : \ j \in \widehat{A}])$. $\tilde{B}$ measures the "bias" in the estimation using the set of predictors $\widehat{A}$ determined by the estimator $\widehat{\beta}$.
- For $0 \leq \alpha \leq \infty$ set

$$\zeta_\alpha(\lambda) = \left(\sum_{j \in A_0^c} |\beta_j|^\alpha 1\{\widehat{\beta}_j = 0\} \right)^{1/\alpha}.$$

  This measures the size of the "big" regression coefficients left out (or omitted) by the estimator $\widehat{\beta}$.

Huang and Zhang (2006) argue that appropriate benchmarks for $\tilde{B}^2$ and $n\zeta_2^2$ are provided by the numbers

$$\lambda\eta_1, \qquad \eta_2^2, \qquad \text{and} \qquad \frac{q\lambda^2}{n} \tag{1}$$

where

$$\eta_2 \equiv \max_{A \subset A_0} \|\sum_{j \in A} \beta_j \mathbf{x}_j\|_2 \leq \max_{j \leq p} \|\mathbf{x}_j\|_2 \eta_1 \asymp \sqrt{n}\eta_1.$$

Example: Suppose that $n^{-1}\mathbf{X}'\mathbf{X} = I_p$ and $\epsilon \sim N_n(0, I_n)$. Then as we have seen, the lasso estimator $\widehat{\beta}$ is given by soft-thresholding:

$$\widehat{\beta}_j = \text{sign}(Z_j)(|Z_j| - \lambda/n)_+, \qquad Z_j \equiv \mathbf{x}_j' Y \sim N(\beta_j, 1/n).$$

If $|\beta_j| = \lambda/n$ for $j = 1, \ldots, q + \eta_1 n\lambda$, and $\lambda/\sqrt{n} \to \infty$, then $P(\widehat{\beta}_j = 0) \approx 1/2$ so that $\tilde{B}^2 \approx 2^{-1}(q + \eta_1 n/\lambda)n(\lambda/n)^2 = 2^{-1}(q\lambda^2/n + \eta_1\lambda)$. Thus we see that $\tilde{B}^2$ cannot be smaller than the first and third of the quantities in (1), while that second quantity η_2^2 is a natural choice of $\tilde{B}^2$ as the maximum mean effect of variables with small coefficients.

In the proof of their Theorem 1 Huang and Zhang (2006) show that $\sqrt{n}\zeta_2$ is of the same order as $\tilde{B} + O(\eta_2)$. This leads them to the following definition:

Definition: The lasso is *rate consistent in model selection* if

$$\widehat{q} = O_p(q), \qquad \tilde{B} = O_p(B), \qquad \text{and} \quad \sqrt{n}\zeta_2 = O_p(B)$$

where

$$B \equiv \max\{\sqrt{\eta_1\lambda},\ \eta_2,\ \sqrt{q\lambda^2/n}\}.$$

Much as in our discussion of Candès and Tao, let $\mathbf{X}_A \equiv (\mathbf{x}_j : j \in A)$ and $\Sigma_A \equiv \mathbf{X}_A'\mathbf{X}_A/n$.

Definition. The design matrix $\mathbf{X}$ satisfies the *partial Riesz condition* (or PRC) with rank q^* and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \le \frac{\mathbf{X}_A v\|_2}{n\|v\|_2} \le c^* \qquad \text{for all} \quad A \quad \text{with} \quad \#(A) = q^*, \quad \text{all } v \in \mathbb{R}^{q^*}.$$

One motivation for this terminology is the following: if $\{\xi_j\}_{j\ge1}$ is a sequence of random variables, then we say the $\{\xi_j\}$ satisfies the *Riesz condition* if there exist $0 < \rho_* \le \rho* < \infty$ sudch that

$$\rho_* \sum_{j=1}^{\infty} b_j^2 \le E\Big|\sum_{j=1}^{\infty} b_j\xi_j\Big|^2 \le \rho^* \sum_{j=1}^{\infty} b_j^2$$

for all $\{b_j\}_{j\ge1} \in \ell_2$.

Since $n^{-1}\|\mathbf{X}_A v\|_2^2 = v'\Sigma_A v$, if $\mathbf{X}$ satisfies the partial Riesz condition, then all the eigenvalues of Σ_A are in $[c_*, c^*]$ when $|A| \le q^*$.

To prepare for the statement of the main theorem of Huang and Zhang (2006), set

$$\begin{aligned}
r_1 &\equiv r_1(\lambda) \equiv \left(\frac{c^*\eta_1 n}{q\lambda}\right)^{1/2},\\
r_2 &\equiv r_2(\lambda) \equiv \left(\frac{c^*\eta_2^2 n}{q\lambda^2}\right)^{1/2},\\
C &\equiv \frac{c^*}{c_*},\\
M_1^* &\equiv M_1^*(\lambda) = 2 + 4r_1^2 + 4\sqrt{C}r_2 + 4C,\\
M_2^* &\equiv M_2^*(\lambda) = \frac{8}{3}\left\{\frac{1}{4} + r_1^2 + r_2\sqrt{2C}(1+\sqrt{2C}) + C\left(\frac{1}{2} + \frac{4}{3}C\right)\right\},\\
M_3^* &\equiv M_3^*(\lambda) \equiv \frac{8}{3}\left\{\frac{1}{4} + r_1^2 + r_2\sqrt{C}(1+2\sqrt{1+C}) + \frac{3r_2^2}{4} + C\left(\frac{5}{6} + \frac{2}{3}C\right)\right\}.
\end{aligned}$$

Recall that an upper bound for the penalty level λ is given by $\lambda^* \equiv \max_{j\le p} |\mathbf{x}_j' Y|$. An appropriate lower bound for the penalty level λ is given by

$$\lambda_* \equiv \inf\{\lambda : \ M_1^*(\lambda)q + 1 \le q^*\}, \qquad \inf \emptyset \equiv \infty.$$

Consider the lasso path in the interval

$$\max\left\{\lambda_*, 2\sqrt{2(1+c_0)c^* n \log(p \wedge a_n}\right\} \le \lambda \le \lambda^* \tag{2}$$

where $c_0 \ge 0$ is a constant and a_n chosen so that $p/(p \wedge a_n)^{1+c_0} \approx 0$. With this preparation, the main theorem of Huang and Zhang (2006) is as follows:

Theorem 1. Suppose that the sparsity and partial Riesz conditions hold. Then there exists a set Ω_1 in the sample space of $(\mathbf{X}, \epsilon)$ such that

$$P(\Omega_1) \ge 1 - \frac{2p}{(p \wedge a_n)^{1+c_0}}$$

and the following inequalities hold on Ω_1 for all λ in the range given by (2):

$$\begin{aligned}
&\widehat{q}(\lambda) \le M_1^*(\lambda)q, \\
&\tilde{B}^2(\lambda) \equiv \|(I - \widehat{P}(\lambda))\mathbf{X}\beta(\lambda)\|_2^2 \le M_2^*(\lambda)\frac{q\lambda^2}{c^* n}, \\
&\zeta_2^2(\lambda) \equiv \sum_{j \in A_0^c} |\beta_j|^2 1\{\widehat{\beta}_j = 0\} \le M_3^*(\lambda)\frac{q\lambda^2}{c^* c_* n^2}.
\end{aligned}$$

Remark: When $\eta_1 = 0$ it follows that $r_1 = r_2 = 0$, and hence

$$\begin{aligned}
M_1^* &= 2 + 4C, \\
M_2^* &= \frac{2}{3} + \frac{4}{3}C + \frac{32}{9}C^2, \\
M_3^* &= \frac{2}{3} + \frac{20}{9}C + \frac{16}{9}C^2,
\end{aligned}$$

all depend only on $C = c^*/c_*$. Then the lower bound $\lambda_* = 0$ for $(2 + 4C)q + 1 \le q^*$ and $\lambda_* = \infty$ otherwise. Thus an implicit requirement in Theorem 1 is that $(2 + rC)q + 1 \le q^*$.

Huang and Zhang go on to develop a number of results giving conditions which imply their partial Riesz condition. The following two propositions are of this type.

Proposition 1. Suppose that $\mathbf{X}$ is deterministic and standardized with $\|\mathbf{x}_j\|_2^2/n = 1$. Let $\rho_{jk} = \mathbf{x}_j'\mathbf{x}_k/n$ be the correlation for $1 \le j \ne k \le p$. If

$$\max_{|A|=q^*} \inf_{\alpha \ge 1} \left\{ \sum_{j \in A} \left(\sum_{k \in A, k \ne j} |\rho_{jk}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right\}^{1/\alpha} \le \delta < 1,$$

then the partial Riesz condition holds with rank q^* and spectrum bounds $c_* = 1/(1+\delta)$, $c^* = 1/(1-\delta)$. In particular, the partial Riesz condition holds if

$$\max_{1\le j<k\le p} |\rho_{jk}| \le \frac{\delta}{q^*-1}, \qquad \delta < 1.$$

Remark: If $\delta = 1/3$, then $C = c^*/c_* = 2$ and Theorem 1 is applicable if $10q+1 \le q^*$ and $\eta_1 = 0$.

Now suppose that $\{\xi_j\}_{j\ge 1}$ is an infinite sequence of possible covariates, and we observe i.i.d. copies $(Y_i, \mathbf{x}^i)$ of $(Y, \xi_{k_1}, \ldots, \xi_{k_p}) \equiv (Y, x_1, \ldots, x_p)$ for certain integers $1 \le k_1 < \cdots < k_p$. Thus $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})$ are the row - vectors of $\mathbf{X} = (x_{ij}) = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. Since $\mathbf{x}^i$ are i.i.d. copies of $(\xi_{k_1}, \ldots, \xi_{k_p})$, if we assume that $\{\xi_j\}$ satisfies the (full) Riesz condition, then it follows that

$$\rho_*\|\mathbf{b}\|_2^2 \le \frac{E\|\mathbf{X}b\|^2}{n} = E\sum_{i=1}^n \frac{(\mathbf{b}'\mathbf{x}^i)^2}{n} = E\Big|\sum_{j=1}^p b_j\xi_{k_j}\Big|^2 \le \rho^*\|\mathbf{b}\|^2.$$

But this does note necessarily ensure that $0 < \kappa \le c_*(m) \le c^*(m) \le 1/\kappa$ for some $\kappa > 0$ with large probability for all m. The following proposition gives bounds that guarantee this when $\{\xi_j\}$ is Gaussian.

Proposition 2. Suppose that the n rows of the random matrix $\mathbf{X}_{n\times p}$ are i.i.d. copies of a sub-vector of a zero-mean Gaussian sequence $\{\xi_j\}$ satisfying the (full) Riesz condition with spectral bounds ρ_* and ρ^* respectively. Let $c_*(m)$ and $c^*(m)$ be the spectral bounds for submatrices with $m \le p$ rows defined by

$$c_*(m) \equiv \min_{|A|=m}\min_{\|v\|=1} \|\mathbf{X}_A v\|^2/n,$$
$$c^*(m) \equiv \max_{|A|=m}\max_{\|v\|=1} \|\mathbf{X}_A v\|^2/n.$$

Let ϵ_k, $k = 1, \ldots, 4$ be numbers in $(0,1)$ satisfying $2\epsilon_1 + 3\epsilon_2 \le 1$ and $\epsilon_3 + \epsilon_r = \{\epsilon_2 - \log(1-\epsilon_2)\}/2$. Then, for all (m,n,p) satisfying $m \le \min\{p,\ \epsilon_1 n\}$ and $\log\{\binom{p}{m}(2m-1)\} \le \epsilon_3 n$,

$$P(\tau_*\rho_* \le c_*(m) \le c^*(m) \le \tau^*\rho^*) \ge 1 - 2e^{-n\epsilon_4}$$

where $\tau_* = 1 + \epsilon_1 - \sqrt{\epsilon_1+\epsilon_2}(\sqrt{1+\epsilon_2} + \sqrt{1-\epsilon_2})$ and $\tau^* = (\sqrt{1+\epsilon_2} + \sqrt{\epsilon_1+\epsilon_2})^2$.