STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 13 (May 8): In this lecture we will continue discussing the paper by Candès and Tao (2007) on *the Dantzig selector*.

Identifiability and "aliasing": Consider a "rank deficient" sub-matrix $\mathbf{X}_{T\cup T'}$ with $2S$ columns and smallest eigenvalue $0 = 1 - \delta_{2S}$ and with indices in T and T' each of size S. Then there exists a vector h such that $\mathbf{X}h = 0$ and $h = \beta - \beta'$ where $\beta_j \neq 0$ for $j \in T$ and similarly for β' (i.e. $\beta'_j \neq 0$ for $j \in T'$?). Thus $\mathbf{X}\beta = \mathbf{X}\beta'$. This implies that β is not identifiable since both β and β' are $S-$sparse. Thus $\delta_{2S} < 1$ is necessary for identifiability.

Other constraints on β: Arrange the entries of β by decreasing order of magnitude:

$$|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|.$$

Suppose that

$$|\beta_{(j)}| \leq Rj^{-1/s} \quad \text{for all } \ j \geq 1 \tag{1}$$

and some $R \in (0,\infty)$, $s \in (0,1]$. Such a coefficient vector β is called *compressible of order* s.

Can we show that $\widehat{\beta} = \widehat{\beta}_{DS}$ achieves an error close to

$$E\|\beta^* - \beta\|_2^2 = \sum_{j=1}^{p} \min\{\beta_j^2, \sigma^2\}?$$

Two observations:

- Let $S = |\{j \leq p : \ |\beta_j| > \sigma\}|$. Then if $\delta_{2S} + \theta_{S,2S} < 1$, there is some hope of success.
- For $\beta \in \mathbb{R}^p$ satisfying (1),

$$\begin{aligned}
\sum_{j=1}^{p} \beta_j^2 \wedge \sigma^2 &= S \cdot \sigma^2 + \sum_{j \geq S+1} |\beta_{(j)}|^2 \\
&\leq S \cdot \sigma^2 + \sum_{j \geq S+1} R^2 j^{-2/s} \\
&\leq S \cdot \sigma^2 + \frac{R^2}{2/s - 1} S^{-2r} \\
&\equiv g(S)
\end{aligned}$$

where $r \equiv 1/s - 1/2 \geq 1/2$ for $s \leq 1$. Moreover, note that for $\beta \in \mathbb{R}^p$ satisfying (1), $|\{j \leq p : \ \beta_j > \sigma\}| \leq (R/\sigma)^s$: this follows easily since $Rj^{-1/s} \geq |\beta_{(j)}| > \sigma$ implies that $j^{-1/s} > \sigma/R$, and hence $j < (R/\sigma)^s$.

Now note that the function $g(S)$ in (2) is minimized over $1 \le S \le S^*$ for any $S^* \ge (2rR^2/\sigma^2)^{1/(2r+1)}$ by

$$S_{opt} \equiv (2rR^2/\sigma^2)^{1/(2r+1)}$$

and then

$$g(S_{opt}) = C_r R^{2/(2r+1)} \sigma^{2r/(2r+1)}$$

where $C_r \equiv (2r)^{1/(2r+1)} + (2r)^{-2r/(2r+1)}$.

In fact, Candès and Tao establish the following theorem:

Theorem 4. (Theorem 1.3 of C & T). Suppose that $\beta \in \mathbb{R}^p$ satisfies (1) and that S_* satisfies $\delta_{2S^*} + \theta_{S^*,2S^*} < 1$. Choose $\lambda_p = \sqrt{2 \log p}$ as in Theorem 1. Then the Dantzig selector $\widehat{\beta}$ satisfies

$$\begin{aligned} \|\widehat{\beta} - \beta\|_2^2 &\le \min_{1 \le S \le S^*} \left\{ C_3 (2 \log p) \left(S \cdot \sigma^2 + R^2 S^{-2r} \right) \right\} \\ &= O(\log p) R^{2/(2r+1)} \left(\sigma^2 \right)^{2r/(2r+1)} \qquad \text{if } \; S_* \ge (2rR^2/\sigma^2)^{1/(2r+1)} \end{aligned}$$

with high probability. If $S_* < (2rR^2/\sigma^2)^{1/(2r+1)}$, then the method "saturates", and the squared loss is bounded by

$$C_3 \cdot 2 \log p \cdot \left(S_* \sigma^2 + R^2 S_*^{-2r} \right).$$

When $\mathbf{X}$ is orthogonal (so $n = p$ and $X^T X = I$, $\widehat{\beta}$ is the ℓ_1 minimizer such that $\|\mathbf{X}^T Y - \widehat{\beta}\|_\infty \le \lambda_p \sigma$. This means that $\widehat{\beta}$ is the soft-thresholded version of $\mathbf{X}^T Y$ at level $\lambda_p \sigma$:

$$\widehat{\beta}_i = \max\{|(\mathbf{X}^T Y)_i - \lambda_p \sigma, \; 0\} \text{sign}((\mathbf{X}^T Y)_i).$$

This means that in this case $\widehat{\beta}$ is a shrinkage estimator, and hence may be biased down. This difficulty very likely carries over to the non-orthogonal case. In order to avoid this bias, Candès and Tao suggest the following two-stage procedure:
Step 1: Estimate $J = \{j \le p : \; \beta_j \ne 0\}$ by $\hat{J} = \{j \le p : \; \widehat{\beta}_n \ne 0\}$ or perhaps by $\tilde{J} = \{j \le p : \; |\widehat{\beta}_j| > \alpha\sigma\}$ for some $\alpha > 0$.
Step 2: Construct the usual least squares estimator $\widehat{\beta}_{\hat{J}}$:

$$\widehat{\beta}_{\hat{J}} = (\mathbf{X}_{\hat{J}}^T \mathbf{X}_{\hat{J}} \mathbf{X}_{\hat{J}}^T Y.$$

Candès and Tao call the resulting estimator $\widehat{\beta}_{\hat{J}}$ the *Gauss-Dantzig* selector.

Other error distributions than normal? Similar developments can be made for error distributions other than the normal distribution. The modification required is to set the threshold λ_p so that

$$Z^* \equiv \sup_{1 \le j \le p} |\langle \underline{x}^j, \epsilon \rangle| \le \lambda_p \sigma$$

with high probability. Another option might be to choose the thresholds to depend on the column index: i.e. choose $(\lambda_p^1, \lambda_p^2, \ldots, \lambda_p^p)$ so that

$$Z^{**} \equiv \sup_{1 \le j \le p} |\frac{\langle \underline{x}^j, \epsilon \rangle}{\lambda_p^j}| \le \sigma$$

with high probability.

Proofs:

Probability bound argument: Assume without loss of generality that $\sigma^2 = 1$. Then it follows that $Z_j \equiv \langle \underline{x}^j, \epsilon \rangle = \sum_{i=1}^n x_i^j \epsilon_i \sim N(0,1)$ for each j since $\|\underline{x}^j\|_2 = 1$ for each $j \in \{1, \ldots, p\}$. Thus

$$P(|Z_j| > z) \le \frac{2}{z}\phi(z)$$

by Mills' ratio. Thus (even though the Z_j are dependent),

$$P(\max_{1 \le j \le p} |Z_j| > z) \le p\frac{2}{z}\phi(z),$$

and hence for $\lambda_p = \sqrt{2 \log p}$

$$\begin{aligned} P(\max_{1 \le j \le p} |Z_j| > \lambda_p) &\le p\frac{2}{\lambda_p}\phi(\lambda_p) \\ &= \frac{p}{\lambda_p}\frac{2}{\sqrt{2\pi}}\exp(-\lambda_p^2/2) \\ &= \sqrt{\frac{2}{\pi 2 \log p}} = \frac{1}{\sqrt{\pi \log p}}. \end{aligned}$$

Taking $\lambda_p = \sqrt{2(1+a)\log p}$ yields

$$P(\max_{1 \le j \le p} |Z_j| > \lambda_p) \le \frac{1}{p^a}\frac{1}{\sqrt{\pi \log p}}.$$

High dimensional geometry:

First, consider the context of Theorem 1.1. Clearly the true coefficient vector β is feasible (with high probability, since $\|X^T(Y - \mathbf{X}\beta)\|_\infty = \|\mathbf{X}^T\epsilon\|_\infty \le \lambda_p$ with high probability as we have shown above), and hence

$$\|\widehat{\beta}\|_1 \le \|\beta\|_1.$$

Write $\widehat{\beta} = \beta + h$ and let $T_0 \equiv \{j \le p : \ \beta_j \ne 0\}$. Here are two geometric facts about the relations between h and β:

Fact 1: The vector $h = \widehat{\beta} - \beta$ must satisfy

$$\|\beta\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 \leq \|\beta + h\|_1 = \|\widehat{\beta}\|_1 \leq \|\beta\|_1.$$

Hence

$$\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1. \tag{2}$$

To see this, note that

$$\begin{aligned} \|\beta + h\|_1 - \|\beta\|_1 &= \sum_{j \in T_0^c} |h_j| + \sum_{j \in T_0^c} \{|\beta_j + h_j| - |\beta_j|\} \\ &\geq \sum_{j \in T_0^c} |h_j| - \sum_{j \in T_0^c} \{|h_j|\} \\ &= \|h_{T_0^c}\|_1 - \|h_{T_0}\|_1 \end{aligned}$$

since $|\beta_j| = |\beta_j + h_j - h_j| \leq |\beta_j + h_j| + |h_j|$.

Fact 2: Since $\epsilon = Y - \mathbf{X}\beta$ and $r = Y - \mathbf{X}\widehat{\beta}$,

$$\begin{aligned} &\langle \epsilon - r, \mathbf{x}^j \rangle = \langle \mathbf{X}\widehat{\beta} - \mathbf{X}\beta, \mathbf{x}^j \rangle = \langle \mathbf{X}h, \mathbf{x}^j \rangle, \text{or, equivalently} \\ &\langle \mathbf{X}h, \mathbf{x}^j \rangle = \langle \epsilon, \mathbf{x}^j \rangle - \langle r, \mathbf{x}^j \rangle, \end{aligned}$$

so it follows that

$$\begin{aligned} |\langle \mathbf{X}h, \mathbf{x}^j \rangle| &\leq |\langle \epsilon, \mathbf{x}^j \rangle| + |\langle r, \mathbf{x}^j \rangle| \\ &\leq \sup_{1 \leq j \leq p} |Z_j| + \|\langle r, \mathbf{X} \rangle\|_\infty \\ &\leq \lambda_p + \lambda_p = 2\lambda_p \end{aligned} \tag{3}$$

with high probability as described above and by definition of $\widehat{\beta}$.

Our goal is to show that Facts 1 and 2 imply that $h = \widehat{\beta} - \beta$ is small in the ℓ_2-norm: i.e. we will show that

$$\sup_{h \in \mathbb{R}^p} \|h\|_2^2 \qquad \text{such that } \|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1 \text{ and } \|\mathbf{X}^T\mathbf{X}h\|_\infty \leq 2\lambda_p$$

satisfies

$$\|h\|_2^2 \leq \lambda_p^2 |T_0|.$$

The situation in the context of Theorem 2 is more complicated. Now let $T_0 \equiv \{j \leq p : |\beta_j| > \sigma\}$ and let β_{T_0} be the vector that agrees with β for $j \in T_0$, but has zero entries for $j \in T_0^c$, and similarly for $\beta_{T_0^c}$; thus $\beta = \beta_{T_0} + \beta_{T_0^c}$.

If β_{T_0} were feasible, then we would have

$$\|\widehat{\beta}\|_1 \leq \|\beta_{T_0}\|_1.$$

Writing $\beta = \beta_{T_0} + h$, the same analysis as above would yield

$$\|\widehat{\beta} - \beta_{T_0}\|_2^2 = O(\log p)|T_0|\sigma^2.$$

Then we would have

$$\begin{aligned} \|\widehat{\beta} - \beta\|_2^2 &= \|\widehat{\beta} - \beta_{T_0} + \beta_{T_0} - \beta\|_2^2 \\ &\leq 2\|\widehat{\beta} - \beta_{T_0}\|_2^2 + \|\beta_{T_0} - \beta\|_2^2 \\ &= O(\log p)|T_0|\sigma^2 + 2 \sum_{j:\ |\beta_j| \leq \sigma^2} \beta_j^2. \end{aligned}$$

which is the desired conclusion for Theorem 2. Unfortunately, although β_{T_0} may be feasible for most $S-$sparse vectors β, it is not feasible for some such vectors, and the argument requires a further detour. Here is a key lemma in getting control: it is here that the $S-$restricted isometry constant δ_S and the restricted orthogonality constant $\theta_{S,2S}$ both appear.

Lemma 1. Suppose that T_0 is a set of cardinality S with $\delta + \theta < 1$. For $h \in \mathbb{R}^p$, let T_1 be the S-largest positions of h outside of T_0. Put $T_{01} \equiv T_0 \cup T_1$. Then

$$\|h\|_{\ell_2(T_{01})} \leq \frac{1}{1-\delta}\|\mathbf{X}_{T_{01}}^T X h\|_{\ell_2} + \frac{\theta}{(1-\delta)\sqrt{S}}\|h\|_{\ell_1(T_0^c)}$$

and

$$\|h\|_{\ell_2}^2 \leq \|h\|_{\ell_2(T_{01})}^2 + S^{-1}\|h\|_{\ell_1(T_0^c)}^2.$$

With Lemma 1 in hand, we can complete the proof of Theorem 1: by (2)

$$\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1 \leq \sqrt{S}\|h_{T_0}\|_2 \tag{4}$$

by the Cauchy-Schwarz inequality. On the other hand, (3) gives

$$\|\mathbf{X}_{T_{01}}^T \mathbf{X} h\|_2 \leq \sqrt{2S} \cdot 2\lambda_p.$$

Then by the first part of Lemma 1,

$$\begin{aligned} \|h\|_{\ell_2(T_{01})} &\leq \frac{1}{1-\delta}\sqrt{2S}2\lambda_p + \frac{\theta}{(1-\delta)\sqrt{S}}\sqrt{S}\|h_{T_0}\|_2 \\ &\leq \frac{1}{1-\delta}\sqrt{2S}2\lambda_p + \frac{\theta}{(1-\delta)}\|h_{T_{01}}\|_2. \end{aligned}$$

This yields

$$\left(1 - \frac{\theta}{1-\delta}\right)\|h\|_{\ell_2(T_{0,1})} \leq \frac{1}{1-\delta}\sqrt{2S}2\lambda_p,$$

and by rearranging we get

$$\|h\|_{\ell_2(T_{0,1})} \leq \frac{1}{1-\delta-\theta}\sqrt{2S}2\lambda_p.$$

Thus by the second part of Lemma 1 and then by (4),

$$\begin{aligned}
\|h\|_2^2 &\leq \|h\|^2_{\ell_2(T_{01})} + \frac{1}{S}\|h\|^2_{\ell_1(T_0^c)} \\
&\leq \|h\|^2_{\ell_2(T_{01})} + \frac{1}{S}S\|h\|^2_{\ell_1(T_0)} \\
&\leq 2\|h\|^2_{\ell_2(T_{01})} \\
&\leq 2\frac{1}{1-\delta-\theta}\sqrt{2S}2\lambda_p.
\end{aligned}$$