# STATISTICS 593C: Spring, 2007

## Model Selection and Regularization

**Jon A. Wellner**

**Lecture 12 (May 3):** In this lecture we will begin discussing the paper by Candès and Tao (2007) on *the Dantzig selector*.

Once again consider the linear model in which we observe $\mathbf{Y}$ where

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \qquad\qquad \epsilon \sim N_n(0, \sigma^2 I)$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is the $n \times p$ design matrix (regarded primarily as non-random) with $p > n$ and $\epsilon$ is $n \times 1$. Candès and Tao regard $\sigma^2$ as "known".

Candès and Tao begin with a discussion of the "noiseless case" in which $\sigma^2 = 0$, or equivalently $\epsilon = 0$. In a series of papers (Candès and Tao (2005), Candès, Romberg, and Tao (2006), Candès and Tao (2006)) they have considered reconstruction or identification of $\beta$ when some further knowledge about $\beta$ is available. The following definitions give various ways in which we might have further knowledge about $\beta$:

**Definition 1.** The coefficient vector $\beta \in \mathbb{R}^p$ is "$S$-sparse" if $\#\{j \le p : \ \beta_j \neq 0\} \le S$.

**Definition 2.** The coefficient vector $\beta \in \mathbb{R}^p$ is "compressible of order $s$" if $|\beta_{(j)}| \le Rj^{-1/s}$ for all $j \ge 1$ for some $0 < R < \infty$ and $0 < s \le 1$.

The results in Candès and Tao (2005), Candès, Romberg and Tao (2006), Candès and Tao (2006) all depend on certain hypotheses about the design matrix $\mathbf{X}$ which they call *Uniform Uncertainty Principles*:

**Definition 3.** The design matrix $\mathbf{X}$ satisfies a *restricted isometry hypothesis* with $S-$restricted isometry constant $\delta_S$ if $\delta_S$ is the smallest $\delta > 0$ satisfying

$$(1-\delta)\|c\|_2^2 \le \|\mathbf{X}_T c\|_2^2 \le (1+\delta)\|c\|_2^2$$

for all subsets $T \subset \{1, \ldots, p\}$ with $|T| \le S$ and $c = (c_j)_{j \in T} \in \mathbb{R}^{|T|}$. Here $\mathbf{X}_T$ is the $n \times |T|$ sub-matrix obtained by extracting the columns of $\mathbf{X}$ corresponding to the indices in $T$.

Candès and Tao describe this as meaning that "every set of columns of $\mathbf{X}$ with cardinality less than $S$ approximately behaves as an orthonormal system.

**Definition 4.** For integers $S, S'$ with $S + S' \le p$, define the *restricted orthogonality constant* $\theta_{S,S'}$ to be the smallest $\theta$ such that

$$|\langle \mathbf{X}_T b, \mathbf{X}_{T'} c \rangle| \le \theta \|b\|_2 \|c\|_2$$

for all $b \in \mathbb{R}^{|T|}$, $c \in \mathbb{R}^{|T'|}$ and all disjoint sets $T, T' \subset \{1, \ldots, p\}$ with $|T| \le S$, $|T'| \le S'$.

Small values of $\theta_{S,S'}$ indicate that disjoint subsets of coordinates span nearly orthogonal subspaces of $\mathbb{R}^n$.

Note that

$$\langle \mathbf{X}_T b, \mathbf{X}_{T'} c \rangle = b' X_T' X_{T'} c$$

is a bilinear mapping from $\mathbb{R}^{|T|} \times \mathbb{R}^{|T'|}$ to $\mathbb{R}$, while $\mathbf{X}_T$ is a linear map from $\mathbb{R}^{|T|}$ to $\mathbb{R}^n$.

Candès and Tao (2005) then prove the following theorem in the noiseless setting:

**Theorem 1.** Suppose that $\beta$ is $S-$sparse with $S$ such that

$$\delta_S + \theta_{S,S} + \theta_{S,2S} < 1.$$

Then $\beta$ is the unique solution of

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \|\tilde{\beta}\|_1 \qquad \text{subject to} \qquad \mathbf{Y} = \mathbf{X}\tilde{\beta}. \tag{1}$$

**Noisy data:** When $\sigma^2 > 0$, then Candès and Tao propose changing the minimization problem (1) as follows:

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \|\tilde{\beta}\|_1 \qquad \text{subject to} \qquad \|\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\tilde{\beta})\|_\infty = \sup_{1 \le i \le p} |(\mathbf{X}^T r)_i| \le \lambda_p \sigma; \tag{2}$$

Here $r \equiv r(\tilde{\beta}) \equiv \mathbf{Y} - \mathbf{X}\beta$ and $\lambda_p > 0$. They call the solution $\widehat{\beta}$ of this minimization problem the *Dantzig selector*.

Candès and Tao describe this minimization problem as "looking for the vector $\tilde{\beta}$ with minimum complexity measured by the $\ell_1$ norm among all coefficient vectors consistent with the data". The constraint on $r$ guarantees that the residuals are "within the noise level".

The current minimization problem makes sense with the convention that all the columns of $\mathbf{X}$ have the same Euclidean (i.e. $\|\cdot\|_2$) size: Candès and Tao assume that the columns of $\mathbf{X}$ all satisfy $\|\mathbf{x}_j\|_2 = 1$ where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. If this is not the case, some modifications need to be made.

It is important to note that the minimization problem in (2) is convex; moreover it can be easily reformulated as a linear programming problem:

$$\min \sum_{i=1}^{p} u_i \qquad \text{subject to} \qquad -u \le \tilde{\beta} \le u \ \text{ and}$$
$$-\lambda_p \sigma \underline{1} \le \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \le \lambda_p \sigma \underline{1}.$$

Here is the first theorem of Candès and Tao (2007):

**Theorem 2.** (Theorem 1.1, C & T). Suppose $\beta \in \mathbb{R}^p$ is any $S-$sparse vector in $\mathbb{R}^p$ with $\delta_{2S} + \theta_{S,2S} < 1$. Let $\lambda_p = \sqrt{2 \log p}$ in (2). Then with large probability the solution $\widehat{\beta}$ of the minimization problem (2) satisfies

$$\|\widehat{\beta} - \beta\|_2^2 \le C_1^2 (2 \log p) \cdot S \cdot \sigma^2 \tag{3}$$

with $C_1 = 4/(1-\delta_S-\theta_{S,2S})$. In fact, with this choice of $\lambda$ the inequality (3) holds with probability at least $1-1/\sqrt{\pi \log p}$. If $\lambda = \sqrt{2(1+a)\log p}$ in (2), then the inequality in (3) holds with probability at least

$$1 - \frac{1}{p^a \sqrt{\pi \log p}}.$$

Why is the (oracle-) inequality of Theorem 2 "reasonable" or "good"? If we knew *exactly which subset $T_0$ of size $S$ had non-zero coefficients $\beta_j$*, then would use this information and estimate $\beta$ via least squares base on the predictors $\mathbf{x}_j$, $j \in T_0$: the resulting "ideal estimator" $\beta^*$ based on the *location of non-zero coefficients oracle* is

$$\beta^*_{T_0} = (\mathbf{X}_{T_0}\mathbf{X}_{T_0})^{-1}\mathbf{X}^T_{T_0}\mathbf{Y}$$

where $\beta^*_{T_0}$ is the restriction of $\beta^*$ to $T_0$ and $\beta_{T_0^c} \equiv 0$. Note that

$$\beta^* = \beta + (\mathbf{X}^T_{T_0}\mathbf{X}_{T_0})^{-1}\mathbf{X}^T_{T_0}\epsilon,$$

and hence

$$\begin{aligned} E\|\beta^* - \beta\|^2 &= E\|(\mathbf{X}^T_{T_0}\mathbf{X}_{T_0})^{-1}\mathbf{X}^T_{T_0}\epsilon\|_2^2 \\ &= \sigma^2 \mathrm{trace}((\mathbf{X}^T_{T_0}\mathbf{X}_{T_0})^{-1}) \geq \sigma^2 \frac{1}{1+\delta_S} S. \end{aligned}$$

Thus the Dantzig-selector $\widehat{\beta}$ is achieving the MSE of the "location of non-zero coefficients oracle" up to a factor $C_1^2(2\log p)$; this is the price of not knowing the locations or positions of the non-zero coefficients in $\beta$.

**More oracle inequalities:**

On the other hand, the bound obtained in Theorem 1 is naive in that it does not take the size of the non-zero coefficients $\beta$ into account. If $\beta$ is small, with $|\beta_i| << \sigma$ for all $1 \leq i \leq p$, then we could consider the trivial estimator $\hat{\beta} = 0$; for this estimator

$$E\|\hat{\beta} - \beta\|_2^2 = \|\beta\|_2^2 << S \cdot \sigma^2.$$

This suggests that we might be able to improve on Theorem 1 "coordinate by coordinate".

- **Simple case: orthogonal design** When $X$ = identity so that $\mathbf{Y} = N_p(\beta, \sigma^2 I)$, consider the "size oracle" which tells us in advance the set

$$T_0 \equiv \{j \leq p : \ |\beta_j| > \sigma\}.$$

If we know $T_0$, then we can use the estimator $\beta^*$ given by

$$\beta^*_j = \begin{cases} Y_j, & j \in T_0, \\ 0, & j \in T_0^c. \end{cases}$$

Then it is easily computed that

$$E\|\beta^* - \beta\|_2^2 = \sum_{j=1}^{p} (\beta_j^2 \wedge \sigma^2). \tag{4}$$

It is well-known that threshholding estimators with threshholding level $\sigma\sqrt{\log p}$ achieve the ideal risk in (4) up to logarithmic factors $\log p = \log n$ in this case.

- **Linear model:** For the linear model which we want to treat here, an "ideal estimator" might be the Least Squares Estimator based on choosing the coefficient vector $\beta^*_{ideal,allsubsets}$ defined by

$$\beta^*_{ideal,allsubsets} = \beta^*_{\hat{I}}, \qquad \hat{I}^* \equiv \text{argmin}_{I \subset \{1,\ldots,p\}} E\|\hat{\beta}_I - \beta_I\|_2^2 \tag{5}$$

where $\hat{\beta}_I$ the the least squares estimator based on $\mathbf{X}_I$. Note that $\beta^*_{ideal,allsubsets}$ is not a real estimator since it depends on the true coefficient vector.

**Question:** Does there exist a real estimator $\widehat{\beta}$ such that

$$\|\widehat{\beta} - \beta\|_2^2 = O(\log p) E\|\beta^*_{ideal,allsubsets} - \beta\|^2?$$

Candès and Tao argue that (4) is a suitable substitute or proxy for the "ideal all subsets regression risk" $E\|\beta^*_{\hat{I}} - \beta\|_2^2$ in (5) as follows. Let $I \subset \{1, \ldots, p\}$, and consider regressing $\mathbf{Y}$ onto $\mathbf{X}_I$. Then

$$\|\hat{\beta}_I - \beta\|_2^2 = \|\hat{\beta}_I - \beta_I\|_2^2 + \|\beta_I - \beta\|_2^2.$$

**check this last identity!** Here the first term involves

$$\hat{\beta}_I - \beta_I = (\mathbf{X}_I^T \mathbf{X}_I)^{-1} \mathbf{X}_I^T \mathbf{X} \beta_{I^c} + (\mathbf{X}_I^T \mathbf{X}_I)^{-1} \mathbf{X}_I^T \epsilon,$$

so that

$$\begin{aligned} E\|\hat{\beta}_I - \beta_I\|_2^2 &= \|(\mathbf{X}_I^T \mathbf{X}_I)^{-1} \mathbf{X}_I^T \mathbf{X} \beta_{I^c}\|_2^2 + \sigma^2 \text{trace}((\mathbf{X}_I^T \mathbf{X}_I)^{-1}) \\ &\geq \frac{\sigma^2 |I|}{1 + \delta_{|I|}}. \end{aligned}$$

Since the second term is just $\sum_{j \in I^c} \beta_j^2$, we conclude that

$$E\|\hat{\beta}_I - \beta\|_2^2 \geq \frac{1}{2} \left\{ \sum_{j \in I^c} \beta_j^2 + |I|\sigma^2 \right\}$$

for all $I$ such that $|I| \leq S$ with $\delta_S < 1$ (say). Now note that

$$\sum_{j=1}^{p} \min\{\beta_j^2, \sigma^2\} = \min_{I \subset \{1,\ldots,p\}} \left\{ \|\beta - \beta_I\|_2^2 + |I|\sigma^2 \right\}.$$

**Theorem 3.** (Theorem 1.2, C & T). Choose $t > 0$ and set $\lambda_p = (1 + t^{-1})\sqrt{2 \log p}$. Suppose that $\beta$ is $S-$sparse with $\delta_{2S} + \theta_{S,2S} < 1 - t$. Then the solution $\widehat{\beta}$ of (2) satisfies

$$\|\widehat{\beta} - \beta\|_2^2 \leq C_2^2 \lambda_p^2 \left( \sigma^2 + \sum_{j=1}^{p} \beta_j^2 \wedge \sigma^2 \right) \tag{6}$$

with high probability. Here

$$C_2 = 2\frac{C_0}{1 - \delta - \theta} + 2\frac{\theta(1+\delta)}{(1-\delta-\theta)^2} + \frac{1+\delta}{1-\delta-\theta},$$
$$C_0 = 2\sqrt{2}\left(1 + \frac{1-\delta^2}{1-\delta-\theta}\right) + (1 + 2^{-1/2})\frac{(1+\delta)^2}{1-\delta-\theta}$$

where $\delta \equiv \delta_{2S}$, $\theta = \theta_{S,2S}$. When $\delta$ and $\theta$ are small, we see that

$$C_0 = 2\sqrt{2}(1+1) + (1 + 2^{-1/2}),$$
$$C_2 = 2C_0 + 1 = 8\sqrt{2} + 2 + \sqrt{2} + 1 = 9\sqrt{2} + 3 \approx 15.78....$$

**Question: what is the probability bound in terms of $t$ and $p$?**