Econometrics **April 15, 2023**

Topic 13: Non-convex Learning + Lasso

by Sai Zhang

Key points: Combining the best of the two, we can use **Lasso plus Concave** method, with Lasso screening and concave component selecting variables, achieving a coordinated intrinsic two-scale learning.

Disclaimer: *The note is built on Prof. Jinchi Lv's lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.*

We are facing a tradeoff:

- **Convex** methods: have appealing prediction power and oracle inequalities, but challenging to provide tight false sign rate control
- **Concave** methods: have good variable selection properties, but challenging to establish global properties and risk properties

Here, we take advantage of the linearity of Lasso (convex *and* concave) and try to combine it with concave regularization to get the best of both.

13.1 Model Setup

Again, consider a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

- response vector $(n \times 1)$: $\mathbf{y} = (y_1, \cdots, y_n)'$
- design matrix $(n \times p)$: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$, with each column rescaled to have L_2-norm $n^{1/2}$

here, we consider a scenario where

- $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})'$ is ***sparse*** (with many 0 components)
- ultra-**high** dimensions: $\log p = O(n^a)$, for some $0 < a < 1$

and consider the penalized least squares

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_0 \|\boldsymbol{\beta}\|_1 + \|p_\lambda(\boldsymbol{\beta})\|_1 \right\} \tag{13.1}$$

where

- $\lambda_0 = c \left(\frac{\log p}{n} \right)^{1/2}$ for some $c > 0$
- $p_\lambda(\boldsymbol{\beta}) = p_\lambda(|\boldsymbol{\beta}|) = (p_\lambda(|\beta_1|), \cdots, p_\lambda(|\beta_p|))'$, with $|\boldsymbol{\beta}| = (|\beta_1|, \cdots, |\beta_p|)'$; the concave penalty $p_\lambda(t)$ is defined on $t \in [0, \infty)$, indexed by $\lambda \geq 0$, increasing in **both** t and λ, $p_\lambda(0) = 0$

the 2 penalty components

- L_1-component: minimum amount of regularization for *removing noise* in prediction
- concave component $\|p_\lambda(\boldsymbol{\beta})\|_1$: adapt model sparsity for *variable selection*

Under this set up, we can derive the hard-thresholding property as

Proposition 13.1.1: Hard-Thresholding Property

Assume the $p_\lambda(t)$, $t \geq 0$, is **increasing and concave** with

- $p_\lambda(t) \geq p_{H,\lambda}(t) = \frac{1}{2}\left[\lambda^2 - (\lambda - t)_+^2\right]$ on $[0, \lambda]$
- $p'_\lambda\left((1-c_1)\lambda\right) \leq c_1\lambda$ for some $c_1 \in [0,1)$
- $-p''_\lambda(t)$ decreasing on $[0, (1-c_1)\lambda]$

then any local minimizer of 13.1 that is also a global minimizer in each coordinate has the **hard-thresholding** feature that each component is either 0 or of magnitude **larger** than $(1-c_1)\lambda$

Such property is shared by a wide class of concave penalties, including hard-thresholding penalty $p_{H,\lambda}(t)$ with $c_1 = 0$, L_0-penalty, and SICA (with suitable c_1).

How to *understand* this proposition? Let $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)'$, then **each** $\hat{\beta}_j$ is the glocal minimizer of the corresponding univariate penalized least-square problem along the $j-$th coordinate. These univariate problems share a common form with (generally) different scalars z

$$\hat{\beta}(z) = \arg\min_{\beta\in\mathbb{R}}\left\{\frac{1}{2}(z-\beta)^2 + \lambda_0|\beta| + p_{H,\lambda}(|\beta|)\right\}$$

after we rescale all covariates to have L_2-norm $n^{1/2}$. The solution to these univariate problems are

$$\hat{\beta}(z) = \operatorname{sgn}(z)(|z| - \lambda_0)\cdot \mathbf{1}_{|z|>\lambda+\lambda_0}$$

these solutions have the same feature as the hard-thresholded estimator: each component is either 0 or of magnitude larger than λ. This provides a better distinction between insignificant and significant covariates then soft-thresholding by L_1 penalty.

With the hard-thresholding property of Prop. 13.1.1, we can prove a basic constraint for the global optimum $\hat{\boldsymbol{\beta}}$ on an event with significant probability (Fan and Lv, 2014)

$$\|\boldsymbol{\delta}_2\|_1 \leq 7\|\boldsymbol{\delta}_1\|_1 \tag{13.2}$$

where $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \left(\hat{\boldsymbol{\beta}}'_1, \hat{\boldsymbol{\beta}}'_2\right)' - \left(\boldsymbol{\beta}'_{0,1}, \boldsymbol{\beta}'_{0,2}\right)' = \left(\boldsymbol{\delta}'_1, \boldsymbol{\delta}'_2\right)'$, with $\boldsymbol{\delta}_1 \in \mathbb{R}^s$. Where does this constraint come from? For the penalized least square quesion 13.1

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{(2n)^{-1}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_0\|\boldsymbol{\beta}\|_1 + \|p_\lambda(\boldsymbol{\beta})\|_1\right\}$$

the global minimizer $\hat{\boldsymbol{\beta}}$ leads to

$$\begin{aligned}
(2n)^{-1}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda_0\|\hat{\boldsymbol{\beta}}\|_1 + \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 =& (2n)^{-1}\|\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda_0\|\hat{\boldsymbol{\beta}}\|_1 + \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \\
=& (2n)^{-1}\|\boldsymbol{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \lambda_0\|\hat{\boldsymbol{\beta}}\|_1 + \|p_\lambda(\hat{\boldsymbol{\beta}})\|_1 \\
\leq& (2n)^{-1}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2 + \lambda_0\|\boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 \\
=& (2n)^{-1}\|\boldsymbol{\epsilon}\|_2^2 + \lambda_0\|\boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1
\end{aligned}$$

then, plug in $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, we get

$$(2n)^{-1}\|\boldsymbol{\epsilon} - \mathbf{X}\boldsymbol{\delta}\|_2^2 + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{\delta}\|_1 + \|p_\lambda(\boldsymbol{\beta}_0 + \boldsymbol{\delta})\|_1 \leq (2n)^{-1}\|\boldsymbol{\epsilon}\|_2^2 + \lambda_0\|\boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1$$
$$(2n)^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\epsilon}'\mathbf{X}\boldsymbol{\delta} + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{\delta}\|_1 + \|p_\lambda(\boldsymbol{\beta}_0 + \boldsymbol{\delta})\|_1 \leq \lambda_0\|\boldsymbol{\beta}_0\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1$$

since $\boldsymbol{\beta}_{0,2} = \mathbf{0}$, $\boldsymbol{\delta}_2 = \boldsymbol{\beta}_{0,2} + \boldsymbol{\delta}_2$, we have

$$\|\boldsymbol{\beta}_0 + \boldsymbol{\delta}\|_1 = \|\boldsymbol{\beta}_{0,1} + \boldsymbol{\beta}_{0,2} + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2\|_1 = \|\boldsymbol{\beta}_{0,1} + \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2\|_1 \leq \|\boldsymbol{\beta}_{0,1} + \boldsymbol{\delta}_1\|_1 + \|\boldsymbol{\delta}_2\|_1$$

hence

$$(2n)^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\epsilon}'\mathbf{X}\boldsymbol{\delta} + \lambda_0\|\boldsymbol{\delta}_2\|_1 \leq \lambda_0\|\boldsymbol{\beta}_{0,1}\|_1 - \lambda_0\|\boldsymbol{\beta}_{0,1} + \boldsymbol{\delta}_1\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0 + \boldsymbol{\delta})\|_1$$

and by the reverse triangle inequality $\|\boldsymbol{\beta}_{0,1}\|_1 - \|\boldsymbol{\beta}_{0,1} + \boldsymbol{\delta}_1\|_1 \leq \|\boldsymbol{\delta}_1\|_1$, we get

$$(2n)^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 - n^{-1}\boldsymbol{\epsilon}'\mathbf{X}\boldsymbol{\delta} + \lambda_0\|\boldsymbol{\delta}_2\|_1 \leq \lambda_0\|\boldsymbol{\delta}_1\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0 + \boldsymbol{\delta})\|_1$$

If assume the distribution of the model error $\boldsymbol{\epsilon}$ as

$$\Pr\left(\left\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty > \frac{\lambda_0}{2}\right) = O\left(p^{-c_0}\right)$$

conditional on the event $\mathcal{E} = \left\{\left\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty \leq \lambda_0/2\right\}$, we have

$$-n^{-1}\boldsymbol{\epsilon}'\mathbf{X}\boldsymbol{\delta} + \lambda_0\|\boldsymbol{\delta}_2\|_1 - \lambda_0\|\boldsymbol{\delta}_1\|_1 \geq -\frac{\lambda_0}{2}\|\boldsymbol{\delta}\|_1 + \lambda_0\|\boldsymbol{\delta}_2\|_1 - \lambda_0\|\boldsymbol{\delta}_1\|_1 = \frac{\lambda_0}{2}\|\boldsymbol{\delta}_2\|_1 - \frac{3\lambda_0}{2}\|\boldsymbol{\delta}_1\|_1$$

plug this result back, get

$$\frac{1}{2n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{\lambda_0}{2}\|\boldsymbol{\delta}_2\|_1 \leq \frac{3\lambda_0}{2}\|\boldsymbol{\delta}_1\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0 + \boldsymbol{\delta})\|_1 \tag{13.3}$$

Now, if we further impose 2 conditions:

- **Condition 1 (eigenvalue condition)**: for some positive constant κ_0

$$\min_{\|\boldsymbol{\delta}\|_2=1, \|\boldsymbol{\delta}\|_0<2s} \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 \geq \kappa_0 \tag{A}$$

$$\kappa = \kappa(s,7) = \min_{\boldsymbol{\delta}\neq\mathbf{0}, \|\boldsymbol{\delta}_2\|_1\leq 7\|\boldsymbol{\delta}_1\|_1} \left\{\frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}_1\|_2 \vee \|\tilde{\boldsymbol{\delta}}_2\|_2}\right\} > 0 \tag{B}$$

  where $\tilde{\boldsymbol{\delta}}_2$ is the subvector of $\boldsymbol{\delta}_2$ consisting of the components with the s **largest** absolute values. Here

  - Condition **(A)** is a mild sparse eigenvalue condition
  - Condition **(B)** combines the restricted eigenvalue assumptions in Bickel et al. (2009)[1]. The intuition is, for OLS estimation, $\mathbf{X}'\mathbf{X}$ should be **positive definite**, that is

$$\min_{0\neq\boldsymbol{\delta}\in\mathbb{R}^p} \left\{\frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2}\right\} > 0$$

    however, when $p > n$, this condition **never** holds, hence we replace $\|\boldsymbol{\delta}\|_2$ with the L_2−norm of $\|\boldsymbol{\delta}_1\|_2$, a subvector of $\boldsymbol{\delta}$

$$\kappa = \kappa(s,7) \min_{\boldsymbol{\delta}\neq\mathbf{0}, \|\boldsymbol{\delta}_2\|_1\leq 7\|\boldsymbol{\delta}_1\|_1} \left\{\frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2}\right\} > 0$$

    and for L_q loss with $q \in (1,2]$, we further bound $\|\tilde{\boldsymbol{\delta}}_2\|_2$, which leads to condition **(B)**.

[1]Introduced by Candes and Tao (2007) for studying the oracle inequalities for the Lasso estimator and Dantzig selector.

- **Condition 2 (hard-thresholding condition)**: The penalty $p_\lambda(t)$ satisfies the conditions of Prop. 13.1.1 with

$$p'_\lambda\left\{(1-c_1)\lambda\right\} \le \lambda_0/4$$
$$\min_{j=1,\cdots,s}|\beta_{0,j}| > \max\left\{(1-c_1)\lambda, 2\kappa_0^{-1}p_\lambda^{1/2}(\infty)\right\}$$

Now, look back at the condition 13.3, we can upper-bound $\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0+\boldsymbol{\delta})\|_1$ by $\frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1$. Consider 2 cases:

- **Case 1**: $\|\hat{\boldsymbol{\beta}}\|_0 \ge s$. By the hard-thresholding condition, we have $|\beta_{0,j}| > (1-c_1)\lambda$ and $p'_\lambda\left\{(1-c_1)\lambda\right\} \le \lambda_0/4$. Hence, for $j = 1, \cdots, s$,
  - if $\hat{\beta}_j \neq 0$, we must have $|\hat{\beta}_j| > (1-c_1)\lambda$. And by the mean-value theorem, we have

  $$|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_j|)| = p'_\lambda(b)(|\hat{\beta}_j| - |\beta_{0,j}|) \le p'_\lambda(b)|\delta_{0,j}|$$

  where b is between $|\beta_{0,j}|$ and $|\hat{\beta}_j|$, hence, $b > |\beta_{0,j}| > (1-c_1)\lambda$, by the concavity of p_λ, we have $p'(b) < p'((1-c_1)\lambda) \le \lambda_0/4$, which leads to $|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_j|)| \le \frac{1}{4}\lambda_0|\delta_j|$.
  - if $\hat{\beta}_j = 0$, since $\|\hat{\boldsymbol{\beta}}_0\| \ge s$, there must exist some $j' > s$ s.t. $\hat{\beta}_{j'} \neq 0$, similarly

  $$\begin{aligned}
  |p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_{j'}|)| &\le |p_\lambda\left(|\beta_{0,j}|\right) - p_\lambda\left((1-c_1)\lambda\right)| + |p_\lambda\left(|\hat{\beta}_{j'}|\right) - p_\lambda\left((1-c_1)\lambda\right)| \\
  &= p'_\lambda(b_1)\left(|\beta_{0,j}| - (1-c_1)\lambda\right) + p'_\lambda(b_2)\left(|\hat{\beta}_{j'}| - (1-c_1)\lambda\right) \\
  &\le p'_\lambda(b_1)\left(|\beta_{0,j}| - \underbrace{|\hat{\beta}_j|}_{=0}\right) + p'_\lambda(b_2)\left(|\hat{\beta}_{j'}| - \underbrace{|\beta_{0,j'}|}_{=0}\right) \\
  &= p'_\lambda(b_1)|\delta_j| + p'_\lambda(b_2)|\delta_{j'}| \le \frac{\lambda_0}{4}\left(|\delta_j| + |\delta_{j'}|\right)
  \end{aligned}$$

  together, we have

  $$\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0+\boldsymbol{\delta})\|_1 \le \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1 \le \frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1$$

- **Case 2**: $\|\hat{\boldsymbol{\beta}}\|_0 = s - k$ for some $k \ge 1$. Then we must have $\|\boldsymbol{\delta}\|_0 \le \|\hat{\boldsymbol{\beta}}\|_0 + \|\boldsymbol{\beta}_0\|_0 \le s - k + s < 2s$, and $\|\boldsymbol{\delta}\|_2 \ge \sqrt{k}\min_{j=1,\cdots,s}|\beta_{0,j}|$. Also, there are at least k null estimates ($\hat{\beta}_j = 0$), thus

  $$\underbrace{\frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \ge \frac{\kappa_0^2}{4}\|\boldsymbol{\delta}\|_2^2}_{\text{Condition 1(A)}} \ge \underbrace{\frac{\kappa_0^2}{4}\left(\sqrt{k}\min_{j=1,\cdots,s}|\beta_{0,j}|\right)^2 \ge kp_\lambda(\infty)}_{\text{Condition 2}} \ge kp_\lambda(|\beta_{0,j}|)$$

  similar to Case 1, we have the desired upper bound

  $$\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0+\boldsymbol{\delta})\|_1 \le kp_\lambda(\infty) + \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1 \le \frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1$$

Combining Case 1 and 2, we have $\|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0+\boldsymbol{\delta})\|_1 \le \frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{4}\lambda_0\|\boldsymbol{\delta}\|_1$, plug this back in 13.3, get

$$\begin{aligned}
\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \lambda_0\|\boldsymbol{\delta}_2\|_1 &\le 3\lambda_0\|\boldsymbol{\delta}_1\|_1 + \|p_\lambda(\boldsymbol{\beta}_0)\|_1 - \|p_\lambda(\boldsymbol{\beta}_0+\boldsymbol{\delta})\|_1 \\
&\le 3\lambda_0\|\boldsymbol{\delta}_1\|_1 + \frac{1}{2n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{2}\lambda_0\underbrace{\|\boldsymbol{\delta}\|_1}_{=\|\boldsymbol{\delta}_1\|_1+\|\boldsymbol{\delta}_2\|_1} \\
&\le 7\lambda_0\|\boldsymbol{\delta}_1\|_1
\end{aligned}$$

which leads to the constraint in 13.2 and $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq 7\lambda_0\|\boldsymbol{\delta}_1\|_1$.

Now, look back at Condition 1(**B**)

$$\kappa = \kappa(s,7) = \min_{\boldsymbol{\delta}\neq\mathbf{0},\|\boldsymbol{\delta}_2\|_1\leq 7\|\boldsymbol{\delta}_1\|_1} \left\{ \frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}_1\|_2 \vee \|\tilde{\boldsymbol{\delta}}_2\|_2} \right\} > 0$$

we have

$$\frac{1}{4}\kappa^2(s,7)\|\boldsymbol{\delta}_1\|_2^2 \leq \frac{1}{4}\kappa^2(s,7)\left(\|\boldsymbol{\delta}_1\|_2^2 \vee \|\tilde{\boldsymbol{\delta}}_2\|_2^2\right) \leq \frac{1}{4n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq \underbrace{\frac{7}{4}\lambda_0\|\boldsymbol{\delta}_1\|_1 \leq \frac{7}{4}\lambda_0\sqrt{s}\|\boldsymbol{\delta}_1\|_2}_{\text{Cauchy-Schwartz inequality}}$$

hence

$$\|\boldsymbol{\delta}_1\|_2 \leq \frac{7\lambda_0\sqrt{s}}{\kappa^2(s,7)} \qquad \|\boldsymbol{\delta}_1\|_1 \leq \sqrt{s}\|\boldsymbol{\delta}_1\|_1 \leq \frac{7\lambda_0 s}{\kappa^2(s,7)} \qquad \|\boldsymbol{\delta}_2'\|_2 \leq \frac{\sqrt{7\lambda_0\sqrt{s}\|\boldsymbol{\delta}_1\|_2}}{\kappa(s,7)}$$

Notice that the k−th largest absolute component of $\boldsymbol{\delta}_2$ is bounded from above by $\|\boldsymbol{\delta}_2\|_1/k$, then for $\boldsymbol{\delta}_{2_s}$, the subvector of $\boldsymbol{\delta}_2$ consisting of components **excluding** those with the s largest magnitudes, we have

$$\|\boldsymbol{\delta}_{2_s}\|_2^2 \leq \sum_{k=s+1}^{p-s} \frac{1}{k^2}\|\boldsymbol{\delta}_2\|_1^2 \leq s^{-1}\|\boldsymbol{\delta}_2\|_1^2 \Rightarrow \|\boldsymbol{\delta}_{2_s}\|_2 \leq \frac{1}{\sqrt{s}}\|\boldsymbol{\delta}_2\|_1 \overset{13.2}{\leq} \frac{7}{\sqrt{s}}\|\boldsymbol{\delta}_1\|_1 \overset{\text{C-S}}{\leq} 7\|\boldsymbol{\delta}_1\|_2$$

since $\boldsymbol{\delta}_{2_s}$ and $\boldsymbol{\delta}_2'$ are a partition of $\boldsymbol{\delta}$, we have

$$\|\boldsymbol{\delta}_2\|_2 \leq \|\boldsymbol{\delta}_{2_s}\|_2 + \|\boldsymbol{\delta}_2'\|_2 \leq 7\|\boldsymbol{\delta}_1\|_2 + \frac{\sqrt{7\lambda_0\sqrt{s}\|\boldsymbol{\delta}_1\|_2}}{\kappa(s,7)} \leq 7\|\boldsymbol{\delta}_1\|_2 \leq \frac{56\lambda_0\sqrt{s}}{\kappa^2(s,7)}$$

Hence, we have the following theorem:

Theorem 13.1.2: 1

Assume that Condition 1 and 2 and the model error bound $\Pr\left(\left\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty > \frac{\lambda_0}{2}\right) = O\left(p^{-c_0}\right)$

References

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Yingying Fan and Jinchi Lv. Asymptotic properties for combined l 1 and concave regularization. *Biometrika*, 101(1):57–70, 2014.