Econometrics **May 2, 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

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

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

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

hence

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

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

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

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

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

plug this result back, get

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

Now, if we further impose 2 conditions:

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

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

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

  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

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

    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 $\boldsymbol{\delta}$

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

    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

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

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 $p'_\lambda\{(1-c_1)\lambda\} \le \lambda_0/4$. 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

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

  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 $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq 7\lambda_0\|\boldsymbol{\delta}_1\|_1$.

13.2 Asymptotic Properties of Global Optimum

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

$$\|\boldsymbol{\delta}_1\|_2 \leq \frac{7\lambda_0\sqrt{s}}{\kappa^2(s,7)} \qquad \|\boldsymbol{\delta}_1\|_1 \leq \sqrt{s}\|\boldsymbol{\delta}_1\|_1 \leq \frac{7\lambda_0 s}{\kappa^2(s,7)} \qquad \|\boldsymbol{\delta}_2'\|_2 \leq \frac{\sqrt{7\lambda_0\sqrt{s}\|\boldsymbol{\delta}_1\|_2}}{\kappa(s,7)} \tag{13.4}$$

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

$$\|\boldsymbol{\delta}_2\|_2 \leq \|\boldsymbol{\delta}_{2_s}\|_2 + \|\boldsymbol{\delta}_2'\|_2 \leq 7\|\boldsymbol{\delta}_1\|_2 + \frac{\sqrt{7\lambda_0\sqrt{s}\|\boldsymbol{\delta}_1\|_2}}{\kappa(s,7)} \leq \frac{56\lambda_0\sqrt{s}}{\kappa^2(s,7)} \tag{13.5}$$

Together, for the estimation loss $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, we have

- **L_2-covar-loss-correlation**: $\frac{1}{n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 \leq 7\lambda_0\|\boldsymbol{\delta}_1\|_1 \leq \frac{(7\lambda_0)^2 s}{\kappa^2(s,7)} \Rightarrow \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 \leq \frac{7\lambda_0\sqrt{s}}{\kappa(s,7)}$
- **L_2-loss**: $\|\boldsymbol{\delta}\|_2 \leq \|\boldsymbol{\delta}_1\|_2 + \|\boldsymbol{\delta}_2\|_2 \leq \frac{63\lambda_0\sqrt{s}}{\kappa^2(s,7)}$
- **L_q-loss**: $\|\boldsymbol{\delta}\|_q \leq \left(s^{(2-q)/2}\|\boldsymbol{\delta}_1\|_2^q\right)^{1/q} = s^{(2-q)/2q}\|\boldsymbol{\delta}_1\|_2 \leq s^{(2-q)/2q}\frac{7\lambda_0\sqrt{s}}{\kappa^2(s,7)} = \frac{7\lambda_0 s^{1/q}}{\kappa^2(s,7)}$ ←by Holder's inequality

Define the **number of falsely discovered signs** as[2]

$$\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right) = \left|\left\{j = 1, \cdots, p : \operatorname{sgn}\left(\hat{\beta}_j\right) \neq \operatorname{sgn}\left(\beta_{0,j}\right)\right\}\right|$$

we know from Prop.13.1.1 that $|\hat{\beta}_j| > (1-c_1)\lambda$ and from Condition 2 that $|\beta_{0,j}| > (1-c_1)\lambda$, then if $\operatorname{sgn}(\hat{\beta}_j) \neq \operatorname{sgn}(\beta_{0,j})$, we must have $|\delta_j| = |\hat{\beta}_j - \beta_{0,j}| \geq (1-c_1)\lambda$. Therefore, it follows that

$$\|\boldsymbol{\delta}\|_2 \geq \left(\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right)\right)^{1/2}(1-c_1)\lambda$$

[2] **Stronger** than the total number of false positives and false negatives.

hence

$$\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right) \leq \frac{\|\boldsymbol{\delta}\|_2^2}{(1-c_1)^2\lambda^2} \leq \left(\frac{63}{1-c_1}\right)^2\left(\frac{\lambda_0}{\lambda}\right)^2\frac{s}{\kappa^4(s,7)}$$

The results above are all conditional on the event $\mathcal{E} = \left\{\left\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty \leq \lambda_0/2\right\}$, hence hold simultaneously with proability $1-O(p^{-c_0})$.

Altogether, we have the following theorem:

Theorem 13.2.1: Properties of the Global Minimizer $\hat{\beta}$

Assume that Condition 1 and 2 and the model error bound $\Pr\left(\left\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right\|_\infty > \frac{\lambda_0}{2}\right) = O\left(p^{-c_0}\right)$, and $p_\lambda(t)$ is continuously differentiable. Then the global minimizer $\hat{\boldsymbol{\beta}}$ of 13.1 has the hard-thresholding property stated in Prop.13.1.1, and, with probability $1-O(p^{-c_0})$, statisfies simultaneously that

$$\frac{1}{\sqrt{n}}\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\|_2 = O(\kappa^{-1}\lambda_0 s^{1/2}) \tag{13.6}$$

$$\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_q = O(\kappa^{-2}\lambda_0 s^{1/q}), \qquad q \in [1,2] \tag{13.7}$$

$$\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right) = O\left(\kappa^{-4}\left(\frac{\lambda_0}{\lambda}\right)^2 s\right) \tag{13.8}$$

If in addition $\lambda \geq \frac{56\lambda_0\sqrt{s}}{(1-c_1)\kappa^2}$, then with probability $1-O(p^{-c_0})$, we also have that

$$\mathrm{sgn}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{sgn}\left(\boldsymbol{\beta}_0\right) \qquad \left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_\infty = O\left(\lambda_0\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\right)$$

where $\mathbf{X}_1$ is the $n\times s$ submatrix of $\mathbf{X}$ corresponding to s nonzero regression coefficients $\beta_{0,j}$.

The proof of the second part follows as such: by assuming $\lambda \geq \frac{56\lambda_0\sqrt{s}}{(1-c_1)\kappa^2}$, from Condition 2, we have $\min_{j=1,\cdots,s}|\beta_{0,j}| > \frac{56\lambda_0\sqrt{s}}{\kappa^2(s,7)}$, combined with 13.4, we know that

$$\mathrm{sgn}\left(\hat{\beta}_j\right) = \mathrm{sgn}\left(\beta_{0,j}\right),\ \forall j = 1,\cdots,s$$

by a simple contradiction argument. In view of 13.5 and the hard-thresholding feature of $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_{0,1}',\hat{\boldsymbol{\beta}}_{0,2}'\right)'$, with $\hat{\boldsymbol{\beta}}_{0,1} = \left(\hat{\beta}_1,\cdots,\hat{\beta}_s\right)'$, a similar contradiction argument leads to $\hat{\boldsymbol{\beta}}_{0,2} = \mathbf{0}$. Together, we have the sign consistency: $\mathrm{sgn}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{sgn}\left(\boldsymbol{\beta}_0\right)$. Under this result, applying Theorem 1 of Lv and Fan (2009), the estimation $\hat{\boldsymbol{\beta}}_{0,1}$ solves the following equation for $\boldsymbol{\gamma} \in \mathbb{R}^s$

$$\boldsymbol{\gamma} = \tilde{\boldsymbol{\beta}}_{0,1} - \left(n^{-1}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{b}$$

where

- $\mathbf{X}_1$ is the $n\times s$ submatrix of $\mathbf{X}$ corresponding to the s non-zero regression coefficients $\beta_{0,j}$
- $\mathbf{b} = \left\{\lambda_0\mathbf{1}_s + p_\lambda'(|\boldsymbol{\gamma}|)\right\}\circ\mathrm{sgn}(\tilde{\boldsymbol{\beta}}_{0,1}) - n^{-1}\mathbf{X}_1'\boldsymbol{\epsilon}$, with componentwise derivative and product.

From the concavity and monotonicity of $p_\lambda(t)$ and Condition 2, we have

$$0 \leq p_\lambda'(t) \leq p_\lambda'\left\{(1-c_1)\lambda\right\} \leq \lambda_0/4$$

this gives that each component of $\hat{\boldsymbol{\beta}}_{0,1}$ has magnitude larger than $(1-c_1)\lambda$. Since $\|n^{-1}\mathbf{X}_1'\boldsymbol{\epsilon}\|_\infty \le \|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\|_\infty \le \frac{\lambda_0}{2}$ on the event $\mathcal{E}$, hence we have

$$\operatorname{sgn}(\mathbf{b}) = \operatorname{sgn}(\tilde{\boldsymbol{\beta}}_{0,1}), \qquad \frac{\lambda_0}{2} \le \|\mathbf{b}\|_\infty \le \frac{7\lambda_0}{4}$$

which completes the proof for Theorem 13.2.1.

How to understand Theorem 13.2.1?

- False sign rate $\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right) = O\left(\kappa^{-4}\left(\frac{\lambda_0}{\lambda}\right)^2 s\right)$ is asymptotically vanishing when $\lambda_0/\lambda \to 0$, outperforming Lasso, whose false sign rate is generally bounded by $O\left(\lambda_{\max}\right)$ with $\lambda_{\max}$ being the largest eigenvalue of Gram matrix $n^{-1}\mathbf{X}'\mathbf{X}$; also outperforming concave method, whose false sign rate is generally of order $O(1)$. When **signal strength is stronger** and λ **is chosen suitably**, sign consistency is stronger as well.
- Convergence rates of $\frac{1}{\sqrt{n}}\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\|_2$ and $\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_q$ are the same as those in Bickel et al. (2009) for the L_1-component, and are consistent with the concave component of Zhang and Zhang (2012). The bounds $O(\kappa^{-1}\lambda_0 s^{1/2}), O(\kappa^{-2}\lambda_0 s^{1/q})$ depend only on the universal regularization parameter $\lambda_0 = c\sqrt{\frac{\log p}{n}}$ for L_1-component, and are independent of λ for concave component.
- The L_∞–bound $\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_\infty = O\left(\lambda_0\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\right)$ involves $\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty$, which is bounded from above by $\sqrt{s}\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_2 \le \sqrt{s}\kappa_0^{-2}$ and can be **dimension-free** in certain scenarios.
- **Oracle property**: Under all conditions of Theorem 13.2.1 hold, and let $\tilde{\boldsymbol{\beta}}$ be the refitted least-squares estimator given by covariates in $\operatorname{supp}\left(\hat{\boldsymbol{\beta}}\right)$, with $\hat{\boldsymbol{\beta}}$ being the estimator in Theorem 13.2.1. The with probability $1-O(p^{-c_0})$, $\tilde{\boldsymbol{\beta}}$ equals the oracle estimator, and has the oracle property if the oracle estimaor is asymptotic normal.

Theorem 13.2.2: Further Properties of the Global Minimizer $\hat{\boldsymbol{\beta}}$

Under the same regularity conditions, with $\epsilon_1, \cdots, \epsilon_n$ independent and identically distributed as ϵ_0, the global minimizer $\hat{\boldsymbol{\beta}}$ in Theorem 13.2.1 satisfies that $\forall \tau > 0$

$$\mathbb{E}\left\{\frac{1}{n}\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right\|_2^2\right\} = O\left(\kappa^{-2}\lambda_0^2 s + m_{2,\tau} + \gamma\lambda_0 p^{-c_0}\right) \tag{13.9}$$

$$\mathbb{E}\left\{\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_q^q\right\} = O\left[\kappa^{-2q}\lambda_0^q s + (2-q)\lambda_0^{-1}m_{2,\tau} + (q-1)\lambda_0^{-2}m_{4,\tau} + \left((2-q)\gamma + (q-1)\gamma^2\right)p^{-c_0}\right] \tag{13.10}$$

$$\mathbb{E}\left\{\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right)\right\} = O\left[\kappa^{-4}\left(\frac{\lambda_0}{\lambda}\right)^s s + \lambda^{-2}m_{2,\tau} + \left(\frac{\gamma\lambda_0}{\lambda^2} + s\right)p^{-c_0}\right] \tag{13.11}$$

where $m_{q,\tau} = \mathbb{E}\left(|\epsilon_0|^q \mathbf{1}_{\{|\epsilon_0|>\tau\}}\right)$ denotes the tail moment and $\gamma = \|\boldsymbol{\beta}_0\|_1 + s\lambda_0^{-1}p_\lambda(\infty) + \tau^2\lambda_0^{-1}$. If in addition $\lambda \ge 56(1-c_1)^{-1}\kappa^{-2}\lambda_0\sqrt{s}$, then we have

$$\mathbb{E}\left\{\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right)\right\} = O\left\{\lambda^{-2}m_{2,\tau} + \left(\gamma\frac{\lambda_0}{\lambda^2} + s\right)p^{-c_0}\right\}$$

$$\mathbb{E}\left\{\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right\|_\infty\right\} = O\left\{\lambda_0\|(n^{-1}\mathbf{X}_1'\mathbf{X}_1)^{-1}\|_\infty + \lambda_0^{-1}m_{2,\tau} + \gamma p^{-c_0}\right\}$$

Again, λ_0 enteres all bounds for the oracle risk inequalities, λ only enters the risk bound for the variable selection loss. This reflects the different roles played by the L_1 penalty and concave penalty in prediction and variable selection.

How to understand Theorem 13.2.2?

- The 3 bounds can have leading orders given in the **first terms** since they are independent of the τ and p^{-c_0}, and the remainders in each bound can be made sufficiently small since τ and c_0 can be chosen arbitrarily large:
  - for bounded error $\epsilon_i \in [-b, b]$, take $\tau = b$ makes the tail moments $m_{q,\tau}$ vanish
  - for Gaussian error $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $m_{q,\tau} = O\left[\tau^{q-1}\exp(-\frac{\tau^2}{2\sigma^2})\right]$ for positive integer q
- the new oracle risk inequalities complement the common results: the inclusion of L_1−component $\lambda_0 t$ stabilizes prediction and variable selection, and leads to oracle risk bounds.
- It's **unclear** whether the concave method alone can enjoy similar risk bounds.

13.3 Computable Solutionss

The global minimizer established so far has nice properties, but due to the non-convexity, there might be computational difficulties in finding such global minimizer. Here, with the coordinate optimization algorithm, one can obtain a path of sparse computable solutions that are global minimizers in each coordinate, as shown in the following Theorem.

Theorem 13.3.1: Asymptotic Properties of the Computable Solutions

Let $\hat{\boldsymbol{\beta}}$ be a computable local minimizer of 13.1 that is global minimizer in **each coordinate** produced by any algorithm satisfying

- $\|\hat{\boldsymbol{\beta}}\|_0 \leq c_2 s$
- $\|\frac{1}{n}\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|_\infty = O(\lambda_0)$, $\lambda \geq c_3\lambda_0$
- $\min_{\|\boldsymbol{\delta}\|_2=1, \|\boldsymbol{\delta}\|_0 \leq c_4 s} \frac{1}{\sqrt{n}}\|\mathbf{X}\boldsymbol{\delta}\|_2 \geq \kappa_0$ for some positive constant c_2, c_3, κ_0
- sufficiently large positive constant c_4

Then under Thm. 13.2.1, $\hat{\boldsymbol{\beta}}$ has the **same** asymptotic properties as for the global minimizer in Thm. 13.2.1.

The implementation is

- employ **path-following** coordinate optimization algorithm (Fan and Lv, 2011; Mazumder et al., 2011) for implementation and choose initial estimate as Lasso estimator
- A convergence analysis of such an algorithm was presented in Lin and Lv (2013)
- One can obtain a path of sparse computable solutions that are global minimizers in each coordinate
- Thm. 13.3.1 suggests that a sufficiently sparse computable solution with small correlation between residual vector and all covariates can enjoy desirable properties

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.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

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

Wei Lin and Jinchi Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

Jinchi Lv and Yingying Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Journal of the American Statistical Association*, 2009.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Cun-Hui Zhang and T Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, page 576, 2012.