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 <u>local minimizer</u> 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 <u>understand</u> this proposition? Let $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$, 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

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

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 $\hat{\beta}$ on an event with significant probability (Fan and Lv, 2014)

$$\|\delta_2\|_1 \leq 7\|\delta_1\|_1 \tag{13.2}$$

where $\delta = \hat{\beta} - \beta_0 = \left(\hat{\beta}'_1, \hat{\beta}'_2\right)' - \left(\beta'_{0,1}, \beta'_{0,2}\right)' = (\delta'_1, \delta'_2)'$, with $\delta_1 \in \mathbb{R}^s$. Where does this constraint come from? For the penalized least square quesion 13.1

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

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

$$\begin{aligned}
(2n)^{-1}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \lambda_0\|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1 &= (2n)^{-1}\|\mathbf{X}\beta_0 + \epsilon - \mathbf{X}\hat{\beta}\|_2^2 + \lambda_0\|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1\\
&= (2n)^{-1}\|\epsilon - \mathbf{X}(\hat{\beta} - \beta_0)\|_2^2 + \lambda_0\|\hat{\beta}\|_1 + \|p_\lambda(\hat{\beta})\|_1\\
&\leq (2n)^{-1}\|\mathbf{y} - \mathbf{X}\beta_0\|_2^2 + \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1\\
&= (2n)^{-1}\|\epsilon\|_2^2 + \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1
\end{aligned}$$

then, plug in $\delta = \hat{\beta} - \beta_0$, we get

$$(2n)^{-1}\|\epsilon - \mathbf{X}\delta\|_2^2 + \lambda_0\|\beta_0 + \delta\|_1 + \|p_\lambda(\beta_0 + \delta)\|_1 \leq (2n)^{-1}\|\epsilon\|_2^2 + \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1$$
$$(2n)^{-1}\|\mathbf{X}\delta\|_2^2 - n^{-1}\epsilon'\mathbf{X}\delta + \lambda_0\|\beta_0 + \delta\|_1 + \|p_\lambda(\beta_0 + \delta)\|_1 \leq \lambda_0\|\beta_0\|_1 + \|p_\lambda(\beta_0)\|_1$$

since $\beta_{0,2} = 0$, $\delta_2 = \beta_{0,2} + \delta_2$, we have

$$\|\beta_0 + \delta\|_1 = \|\beta_{0,1} + \beta_{0,2} + \delta_1 + \delta_2\|_1 = \|\beta_{0,1} + \delta_1 + \delta_2\|_1 \leq \|\beta_{0,1} + \delta_1\|_1 + \|\delta_2\|_1$$

hence

$$(2n)^{-1}\|\mathbf{X}\delta\|_2^2 - n^{-1}\epsilon'\mathbf{X}\delta + \lambda_0\|\delta_2\|_1 \leq \lambda_0\|\beta_{0,1}\|_1 - \lambda_0\|\beta_{0,1} + \delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$$

and by the reverse triangle inequality $\|\beta_{0,1}\|_1 - \|\beta_{0,1} + \delta_1\|_1 \leq \|\delta_1\|_1$, we get

$$(2n)^{-1}\|\mathbf{X}\delta\|_2^2 - n^{-1}\epsilon'\mathbf{X}\delta + \lambda_0\|\delta_2\|_1 \leq \lambda_0\|\delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1$$

If assume the distribution of the model error ϵ as

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

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

$$-n^{-1}\epsilon'\mathbf{X}\delta + \lambda_0\|\delta_2\|_1 - \lambda_0\|\delta_1\|_1 \geq -\frac{\lambda_0}{2}\|\delta\|_1 + \lambda_0\|\delta_2\|_1 - \lambda_0\|\delta_1\|_1 = \frac{\lambda_0}{2}\|\delta_2\|_1 - \frac{3\lambda_0}{2}\|\delta_1\|_1$$

plug this result back, get

$$\frac{1}{2n}\|\mathbf{X}\delta\|_2^2 + \frac{\lambda_0}{2}\|\delta_2\|_1 \leq \frac{3\lambda_0}{2}\|\delta_1\|_1 + \|p_\lambda(\beta_0)\|_1 - \|p_\lambda(\beta_0 + \delta)\|_1 \tag{13.3}$$

Now, if we further impose 2 conditions:

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

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

$$\kappa = \kappa(s, 7) = \min_{\delta \neq 0, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left\{\frac{1}{\sqrt{n}}\frac{\|\mathbf{X}\delta\|_2}{\|\delta_1\|_2 \vee \|\tilde{\delta}_2\|_2}\right\} > 0 \tag{B}$$

  where $\tilde{\delta}_2$ is the subvector of δ_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

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

  however, when $p > n$, this condition **never** holds, hence we replace $\|\delta\|_2$ with the L_2-norm of $\|\delta_1\|_2$, a subvector of δ

$$\kappa = \kappa(s, 7) \min_{\delta \neq 0, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left\{\frac{1}{\sqrt{n}}\frac{\|\mathbf{X}\delta\|_2}{\|\delta\|_2}\right\} > 0$$

  and for L_q loss with $q \in (1, 2]$, we further bound $\|\tilde{\delta}_2\|_2$, 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 $\hat{\beta}_j \ne 0$, we must have $|\hat{\beta}_j| > (1 - c_1)\lambda$. And by the <u>mean-value theorem</u>, 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. $\hat{\beta}_{j'} \ne 0$, similarly

$$|p_\lambda(|\beta_{0,j}|) - p_\lambda(|\hat{\beta}_{j'}|)| \le |p_\lambda(|\beta_{0,j}|) - p_\lambda((1 - c_1)\lambda)| + |p_\lambda(|\hat{\beta}_{j'}|) - p_\lambda((1 - c_1)\lambda)|$$

$$= p'_\lambda(b_1)(|\beta_{0,j}| - (1 - c_1)\lambda) + p'_\lambda(b_2)(|\hat{\beta}_{j'}| - (1 - c_1)\lambda)$$

$$\le p'_\lambda(b_1)\left(|\beta_{0,j}| - \underbrace{|\hat{\beta}_j|}_{=0}\right) + p'_\lambda(b_2)\left(|\hat{\beta}_{j'}| - \underbrace{|\beta_{0,j'}|}_{=0}\right)$$

$$= p'_\lambda(b_1)|\delta_j| + p'_\lambda(b_2)|\delta_{j'}| \le \frac{\lambda_0}{4}\left(|\delta_j| + |\delta_{j'}|\right)$$

  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

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

  similar to Case 1, we have the desired upper bound

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

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

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

$$\le 3\lambda_0\|\boldsymbol{\delta}_1\|_1 + \frac{1}{2n}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \frac{1}{2}\lambda_0 \underbrace{\|\boldsymbol{\delta}\|_1}_{=\|\boldsymbol{\delta}_1\|_1 + \|\boldsymbol{\delta}_2\|_1}$$

$$\le 7\lambda_0\|\boldsymbol{\delta}_1\|_1$$

which leads to the constraint in 13.2 and $\frac{1}{n}\|\mathbf{X}\delta\|_2^2 \leq 7\lambda_0\|\delta_1\|_1$.

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

$$\kappa = \kappa(s,7) = \min_{\delta \neq 0, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left\{ \frac{1}{\sqrt{n}} \frac{\|\mathbf{X}\delta\|_2}{\|\delta_1\|_2 \vee \|\tilde{\delta}_2\|_2} \right\} > 0$$

we have

$$\frac{1}{4}\kappa^2(s,7)\|\delta_1\|_2^2 \leq \frac{1}{4}\kappa^2(s,7)\left(\|\delta_1\|_2^2 \vee \|\tilde{\delta}_2\|_2^2\right) \leq \frac{1}{4n}\|\mathbf{X}\delta\|_2^2 \leq \underbrace{\frac{7}{4}\lambda_0\|\delta_1\|_1 \leq \frac{7}{4}\lambda_0\sqrt{s}\|\delta_1\|_2}_{\text{Cauchy-Schwartz inequality}}$$

hence

$$\|\delta_1\|_2 \leq \frac{7\lambda_0\sqrt{s}}{\kappa^2(s,7)} \qquad \|\delta_1\|_1 \leq \sqrt{s}\|\delta_1\|_1 \leq \frac{7\lambda_0 s}{\kappa^2(s,7)} \qquad \|\delta_2'\|_2 \leq \frac{\sqrt{7\lambda_0\sqrt{s}\|\delta_1\|_2}}{\kappa(s,7)} \qquad (13.4)$$

Notice that the $k-$th largest absolute component of δ_2 is bounded from above by $\|\delta_2\|_1/k$, then for δ_{2_s}, the subvector of δ_2 consisting of components **excluding** those with the s largest magnitudes, we have

$$\|\delta_{2_s}\|_2^2 \leq \sum_{k=s+1}^{p-s} \frac{1}{k^2}\|\delta_2\|_1^2 \leq s^{-1}\|\delta_2\|_1^2 \Rightarrow \|\delta_{2_s}\|_2 \leq \frac{1}{\sqrt{s}}\|\delta_2\|_1 \overset{13.2}{\leq} \frac{7}{\sqrt{s}}\|\delta_1\|_1 \overset{\text{C-S}}{\leq} 7\|\delta_1\|_2$$

since δ_{2_s} and δ_2' are a partition of δ, we have

$$\|\delta_2\|_2 \leq \|\delta_{2_s}\|_2 + \|\delta_2'\|_2 \leq 7\|\delta_1\|_2 + \frac{\sqrt{7\lambda_0\sqrt{s}\|\delta_1\|_2}}{\kappa(s,7)} \leq \frac{56\lambda_0\sqrt{s}}{\kappa^2(s,7)} \qquad (13.5)$$

Together, for the estimation loss $\delta = \hat{\beta} - \beta_0$, we have

- L_2-**covar-loss-correlation**: $\frac{1}{n}\|\mathbf{X}\delta\|_2^2 \leq 7\lambda_0\|\delta_1\|_1 \leq \frac{(7\lambda_0)^2 s}{\kappa^2(s,7)} \Rightarrow \frac{1}{\sqrt{n}}\|\mathbf{X}\delta\|_2 \leq \frac{7\lambda_0\sqrt{s}}{\kappa(s,7)}$

- L_2-**loss**: $\|\delta\|_2 \leq \|\delta_1\|_2 + \|\delta_2\|_2 \leq \frac{63\lambda_0\sqrt{s}}{\kappa^2(s,7)}$

- L_q-**loss**: $\|\delta\|_q \leq \left(s^{(2-q)/2}\|\delta_1\|_2^q\right)^{1/q} = s^{(2-q)/2q}\|\delta_1\|_2 \leq s^{(2-q)/2q}\frac{7\lambda_0\sqrt{s}}{\kappa^2(s,7)} = \frac{7\lambda_0 s^{1/q}}{\kappa^2(s,7)}$ ←by Holder's inequality

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

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

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 $\text{sgn}(\hat{\beta}_j) \neq \text{sgn}(\beta_{0,j})$, we must have $|\delta_j| = |\hat{\beta}_j - \beta_{0,j}| \geq (1-c_1)\lambda$. Therefore, it follows that

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

hence

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

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

Altogether, we have the following theorem:

---

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

> ### Theorem 13.1.2: Properties of the Global Minimizer $\hat{\beta}$
>
> Assume that Condition 1 and 2 and the model error bound $\Pr\left(\left\|n^{-1}\mathbf{X}'\epsilon\right\|_\infty > \frac{\lambda_0}{2}\right) = O\left(p^{-c_0}\right)$, and $p_\lambda(t)$ is continuously differentiable. Then the global minimizer $\hat{\beta}$ 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
>
> $$\frac{1}{\sqrt{n}}\left\|\mathbf{X}\left(\hat{\beta} - \beta_0\right)\right\|_2 = O(\kappa^{-1}\lambda_0 s^{1/2}) \tag{13.6}$$
>
> $$\left\|\hat{\beta} - \beta_0\right\|_q = O(\kappa^{-2}\lambda_0 s^{1/q}), \qquad\qquad q \in [1,2] \tag{13.7}$$
>
> $$\mathrm{FS}\left(\hat{\beta}\right) = O\left(\kappa^{-4}\left(\frac{\lambda_0}{\lambda}\right)^2 s\right) \tag{13.8}$$
>
> 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
>
> $$\mathrm{sgn}\left(\hat{\beta}\right) = \mathrm{sgn}\left(\beta_0\right) \qquad\qquad \left\|\hat{\beta} - \beta_0\right\|_\infty = O\left(\lambda_0\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\right)$$
>
> 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 $\hat{\beta} = \left(\hat{\beta}_{0,1}', \hat{\beta}_{0,2}'\right)'$, with $\hat{\beta}_{0,1} = \left(\hat{\beta}_1,\cdots,\hat{\beta}_s\right)'$, a similar contradiction argument leads to $\hat{\beta}_{0,2} = \mathbf{0}$. Together, we have the sign consistency: $\mathrm{sgn}\left(\hat{\beta}\right) = \mathrm{sgn}\left(\beta_0\right)$. Under this result, applying Theorem 1 of Lv and Fan (2009), the estimation $\hat{\beta}_{0,1}$ solves the following equation for $\gamma \in \mathbb{R}^s$

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

where

- $\mathbf{X}_1$ is the $n \times s$ submatrix of $\mathbf{X}$ corresponding to the s non-zero regression coefficients $\beta_{0,j}$
- $\mathbf{b} = \left\{\lambda_0\mathbf{1}_s + p_\lambda'(|\gamma|)\right\} \circ \mathrm{sgn}(\tilde{\beta}_{0,1}) - n^{-1}\mathbf{X}_1'\epsilon$, 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 $\hat{\beta}_{0,1}$ has magnitude larger than $(1-c_1)\lambda$. Since $\|n^{-1}\mathbf{X}_1'\epsilon\|_\infty \leq \|n^{-1}\mathbf{X}'\epsilon\|_\infty \leq \frac{\lambda_0}{2}$ on the event $\mathcal{E}$, hence we have

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

which completes the proof for Theorem 13.1.2.

How to understand Theorem 13.1.2?

- False sign rate $\text{FS}\left(\hat{\beta}\right) = O\left(\kappa^{-4}\left(\frac{\lambda_0}{\lambda}\right)^2 s\right)$ 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 $\frac{1}{\sqrt{n}}\left\|\mathbf{X}\left(\hat{\beta}-\beta_0\right)\right\|_2$ and $\left\|\hat{\beta}-\beta_0\right\|_q$ 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 $\left\|\hat{\beta}-\beta_0\right\|_\infty = O\left(\lambda_0\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty\right)$ involves $\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_\infty$, which is bounded from above by $\sqrt{s}\left\|\left(\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\right\|_2 \leq \sqrt{s}\kappa_0^{-2}$ and can be **dimension-free** in certain scenarios.

- **Oracle property**: Under all conditions of Theorem 13.1.2 hold, and let $\tilde{\beta}$ be the refitted least-squares estimator given by covariates in $\text{supp}\left(\hat{\beta}\right)$, with $\hat{\beta}$ being the estimator in Theorem 13.1.2. The with probability $1 - O(p^{-c_0})$, $\tilde{\beta}$ equals the oracle estimator, and has the oracle property if the oracle estimimaor is asymptotic normal.

Theorem 13.1.3: Further Properties of the Global Minimizer $\hat{\beta}$

Under the same regularity conditions, the global minimizer $\hat{\beta}$ in Theorem 13.1.2 satisfies that $\forall \tau > 0$

$$\mathbb{E}\left\{\frac{1}{n}\left\|\mathbf{X}\left(\hat{\beta}-\beta_0\right)\right\|_2^2\right\} = O\left(\kappa^{-2}\lambda_0^2 s + m_{2,\tau} + \gamma\lambda_0 p^{-c_0}\right) \tag{13.9}$$

$$\mathbb{E}\left\{\left\|\hat{\beta}-\beta_0\right\|_q^q\right\} = O\left[\kappa^{-2q}\lambda_0^q s + (2-q)\lambda_0^{-1}m_{2,\tau} + (q-1)\lambda_0^{-2}m_{4,\tau} + \left((2-q)\gamma + (q-1)\gamma^2\right)p^{-c_0}\right] \tag{13.10}$$

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

where $m_{q,\tau} = \mathbb{E}\left(|\epsilon_0|^q \mathbf{1}_{\{|\epsilon_0|>\tau\}}\right)$ denotes the tail moment and $\gamma = \|\beta_0\|_1 + s\lambda_0^{-1}p_\lambda(\infty) + \tau^2\lambda_0^{-1}$. If in addition $\lambda \geq 56(1-c_1)^{-1}\kappa^{-2}\lambda_0\sqrt{s}$, then we have

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

$$\mathbb{E}\left\{\left\|\hat{\beta}-\beta_0\right\|_\infty\right\} = O\left\{\lambda_0\|(n^{-1}\mathbf{X}_1'\mathbf{X}_1)^{-1}\|_\infty + \lambda_0^{-1}m_{2,\tau} + \gamma p^{-c_0}\right\}$$

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 ???

-

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.

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.

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