Econometrics **April 10, 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)

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

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

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

References

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