Econometrics **May 10, 2023**

Topic 14: Regularization Methods in Thresholded Parameter Space

by Sai Zhang

Key points: The connections and differences of all regularization methods and some interesting phase transition phenomena.

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.*

14.1 Model Setup

Now, consider a generalized linear model (GLM) linking a p-dimensional predictor $\mathbf{x}$ to a scalar response Y. With canonical link, the conditional distribution of Y given $\mathbf{x}$ has density

$$f(y;\theta,\phi) = \exp\left[y\theta - b(\theta) + c(y,\phi)\right]$$

where $\theta = \mathbf{x}'\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ a $p-$dimensional regression coefficient vector, $b(\dot{})$ and $c(\cdot,\cdot)$ are know functions and ϕ is dispersion parameter. Again, $\boldsymbol{\beta} = \left(\beta_{0,1},\cdots,\beta_{0,p}\right)'$ is sparse with many zero components, and $\log p = O(n^a)$ for some $0 < a < 1$.

The penalized negative log-likelihood is

$$Q_n(\boldsymbol{\beta}) = -n^{-1}\left[\mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{1}'\mathbf{b}(\mathbf{X}\boldsymbol{\beta})\right] + \|p_\lambda(\boldsymbol{\beta})\|_1$$

where

- $\mathbf{y} = \left(y_1,\cdots,y_n\right)'$, $\mathbf{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)'$, each column of $\mathbf{X}$ is rescaled to have L_2-norm $\sqrt{n}$
- $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1),\cdots,b(\theta_n))'$ with $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_n)'$
- $\|p_\lambda(\boldsymbol{\beta})\|_1 = \sum_{j=1}^p p_\lambda(|\beta_j|)$

Next, define **robust spark** κ_c

Definition 14.1.1: Robust spark κ_c

The robust spark κ_c of the $n\times p$ design matrix $\mathbf{X}$ is defined as the smallest possible positive integer s.t. there exists an $n\times\kappa_c$ submatrix of $\frac{1}{\sqrt{n}}\mathbf{X}$ having a singular value less than a given positive constant c (Zheng et al., 2014), and

$$\kappa_c \le n+1$$

Bounding sparse model size can control collinearity and ensure model identifiability and stability, and as $c \to 0+$, κ_c approaches the spark. Robust spark can be some large number diverging with n:

Proposition 14.1.2: Order of κ_c

Assume $\log p = o(n)$ and that the rows of the $n\times p$ random design matrix $\mathbf{X}$ are i.i.d. as $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has smallest eigenvalue bounded from below by some positive constant. Then there exist positive constants c and $\tilde{c}$ s.t. with asymptotic probability one, $\kappa_c \geq \frac{\tilde{c}n}{\log p}$

Next, we define a thresholded parameter space

Definition 14.1.3: Thresholded parameter space

$$\mathcal{B}_{\tau,c} = \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 < \frac{\kappa_c}{2}, \text{ and for each } j, \beta_j = 0 \text{ or } |\beta_j| \geq \tau\right\}$$

where $\boldsymbol{\beta} = \left(\beta_1, \cdots, \beta_p\right)'$. τ is some positive threshold on parameter magnitude:

Here, τ is very important:

- τ is key to distinguishing between important covariates and noise covariates for the purpose of variable selection
- τ typically needs to satisfy $\tau\sqrt{n/\log p} \xrightarrow{n\to\infty} \infty$

It turns out that the solution to the regularizaiton problem has the (very natural) hard-thresholding property:

Proposition 14.1.4: Hard-thresholding property

or the L_0-penalty $p_\lambda(t) = \lambda \mathbf{1}_{t\neq 0}$, the global minimizer $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)'$ of the regularization problem over $\mathbb{R}^p$ satisfies that each component $\hat{\beta}_j$ is either 0 or has magnitude larger than some positive threshold

This hard-thresholding property is shared by many other penalties such as SICA penalties. This property guarantees sparcity of the model: weak signals are generally difficult to stand out comparing to noise variables due to impact of high dimensionality

14.2 Asymptotic Equivalence of Regularization Methods

For a universal $\lambda = c_0\sqrt{\log p/n}$ with $c_0 > 0$ and p implicitly as $n \vee p$, consider 2 key events:

$$\mathcal{E} = \left\{\|n^{-1}\mathbf{X}'\boldsymbol{\epsilon}\|_\infty \leq \lambda/2\right\} \qquad \mathcal{E}_0 = \left\{\|n^{-1}\mathbf{X}'_{\alpha_0}\boldsymbol{\epsilon}\|_\infty \leq c_0\sqrt{\log n/n}\right\}$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbb{E}\mathbf{y}$, $\mathbf{X}_\alpha$ is a submatrix of $\mathbf{X}$ consisting of columns in α. Here, let $\alpha_0 = \operatorname{supp}\left(\boldsymbol{\beta}_0\right)$ (non-zero variables in the true model).

For this setting, consider the following technical conditions:

C1 **Error tail distribution**: $\Pr(\mathcal{E}^c) = O(p^{-c_1})$ and $\Pr(\mathcal{E}_0^c) = O(n^{-c_1})$ for some positive constant c_1 that can be sufficiently large for large enough c_0

C2 **Bounded variance**: $b(\theta)$ satisfies that $c_2 \leq b''(\theta) \leq c_2^{-1}$ in its domain, where c_2 is some positive constant

C3 **Concave penalty function**: $p_\lambda(t)$ is increasing and concave in $t \in [0,\infty)$ with $p_\lambda(0) = 0$, and is differentiable with $p'_\lambda(0+) = c_3\lambda$ for some positive constant c_3[1]

C4 **Ultra-high dimensionality**: $\log p = O(n^a)$ for some constant $a \in (0,1)$

[1]A wide class of penalties, including L_1-penalty in Lasso, SCAD, MCP and SICA, satisfy this condition.

C5 **True parameter vector**: $s = o(n^{1-a})$ and $\exists c > 0$ s.t. the **robust spark** $\kappa_c > 2s$. Moreover, $\min_{1\le j\le s}|\beta_{0,j}| \gg \sqrt{\log p/n}$

Given these 5 conditions, we have that the global minimizer $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}_\tau} Q_n(\boldsymbol{\beta})$ exists and satisfies oracle inequalities:

Theorem 14.2.1: Oracle Inequalities

Assume that Condition 1-5 hold and τ is chosen s.t. $\tau < \min_{1\le j\le s}|\beta_{0,j}|$ and $\lambda = c_0\sqrt{\log p/n} = o(\tau)$, then the global minimizer exists, and any such global minimizer satisfies that with probability at least $1 - O(p^{-c_1})$, it holds simultaneously that

- **False sign**:

$$FS(\hat{\boldsymbol{\beta}}) \le \frac{Cs\lambda^2\tau^{-2}}{1 - C\lambda^2\tau^{-2}}$$

- **Estimation losses**:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q \le C\lambda s^{1/q}(1 - C\lambda^2\tau^{-2})^{-1/q} \qquad \forall q \in [1,2]$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \le C\lambda s^{1/2}(1 - C\lambda^2\tau^{-2})^{-1/2}$$

- **Prediction loss**:

$$\frac{1}{\sqrt{n}}\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\|_2 \le C\lambda s^{1/2}(1 - C\lambda^2\tau^{-2})^{-1/2}$$

where C is some positive constant.

How to understand Thm.14.2.1

- These results hold uniformly over the set of all possible global minimizers
- c_1 in probability bound can be chosen arbitrarily large, affecting **only** C
- $FS\left(\hat{\boldsymbol{\beta}}\right) = o(s)$ since $\lambda = o(\tau)$, while $\|\hat{\boldsymbol{\beta}}\|_0 = O(\phi_{\max}s)$ where $\phi_{\max}$ is the largest eigenvalue of $\frac{1}{n}\mathbf{X}'\mathbf{X}$
- $\forall q \in [1,2]$, the convergence rates of estimation losses

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q = O\left\{s^{1/q}\sqrt{\frac{\log p}{n}}\right\}$$
$$\frac{1}{\sqrt{n}}\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\|_2 = O\left(\sqrt{\frac{s\log p}{n}}\right)$$

  are consistent with Lasso.

We also have a sign consistency result:

Theorem 14.2.2: Sign Consistency and Oracle Inequalities

Assume the same conditions of Thm.14.2.1, further assume $\min_{1\le j\le s}|\beta_{0,j}| \ge 2\tau$ and $\lambda = c_0\sqrt{\log p/n} = o(s^{-1/2}\tau)$, and $\gamma_n = o\left(\tau\sqrt{\frac{n}{s\log n}}\right)$, then any global minimizer $\hat{\boldsymbol{\beta}}$ defined satisfies that with probability at least $1 - O(n^{-c_1})$, it holds simultaneously that

- **Sign consistency**: $\operatorname{sgn}\left(\hat{\boldsymbol{\beta}}\right) = \operatorname{sgn}\left(\boldsymbol{\beta}_0\right)$
- **Estimation and prediction losses**: If the penalty function further satisfies $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$, then $\forall q \in [1,2]$,

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_q \le C s^{1/q} \sqrt{\frac{\log n}{n}} \qquad \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_\infty \le C \gamma_n^* \sqrt{\frac{\log n}{n}} \qquad n^{-1} D\left(\hat{\boldsymbol{\beta}}\right) \le C \frac{s \log n}{n}$$

where γ_n^* is a constant showing the behavior of $\|\left[\frac{1}{n}\mathbf{X}'_{\alpha_0}\mathbf{H}\left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n\right)\mathbf{X}_{\alpha_0}\right]^{-1}\|_\infty$ in a small neighborhood of $\boldsymbol{\beta}_0$, $D(\hat{\boldsymbol{\beta}})$ is the Kullback-Leibler divergence, and C is some positive constant

How to understand Thm.14.2.2 Consider a linear model, where

$$\gamma_n^* = \left\|\left(\frac{1}{n}\mathbf{X}'_{\alpha_0}\mathbf{X}_{\alpha_0}\right)^{-1}\right\|_\infty \le \sqrt{s}\left\|\left(\frac{1}{n}\mathbf{X}'_{\alpha_0}\mathbf{X}_{\alpha_0}\right)^{-1}\right\|_2 \le \frac{\sqrt{s}}{c} \qquad \gamma_n = \sup_{\alpha \subset \{s+1,\cdots,p\}, |\alpha| \le s} \left\|\frac{1}{n}\mathbf{X}'_{\alpha_0}\mathbf{X}_\alpha\right\|_\infty$$

when all ture covariates are orthogonal to each other, $\gamma_n^* = 1$ and

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \le C\sqrt{\frac{\log n}{n}}$$

within a logarithmic factor $\log n$ or oracle rate. Meanwhile, the penalty function condition $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$ can be easily satisfied by concave penalties such as SCAD and SICA, having convergence rates improved with $\log n$ in place of $\log p$.

Phase transition phenomenon Combining Thm.14.2.1 and 14.2.2, it's shown that

- for $p = O(n^a)$, Lasso and concave regularization methods are **asymptotically equivalent**, having the same convergence rates in the oracle inequalities, with a logarithmic factor of $\log n$
- for $\log p = O(n^a)$, concave regularization methods are **asymptotically equivalent** and still enjoy the same convergence rates in the oracle inequalities, with a logarithmic factor of $\log n$[2].

A phase diagram on how the performance of regularization methods, in the thresholded parameter space, evolves with dimensionality and penalty function.

Further, we have the following **oracle risk inequalities** of the global minimizer

Theorem 14.2.3: Oracle Risk Inequalities

Assume that conditions of Thm.14.2.2 hold and **the fourth moments of errors $\mathbb{E}\epsilon_i^4$ are uniformly bounded** . Then any global minimizer $\hat{\boldsymbol{\beta}}$ defined satisfies that

[2]For Lasso, the condition $p'_\lambda(\tau) = O\left(\frac{\log n}{n}\right)$ and the choice of $\lambda = c_0\sqrt{\frac{\log p}{n}}$ are **incompatible** with each other in this ultra-high dimensional case, and the convergence rates for Lasso (of $\log p$) are slower than those for concave regularization methods.

- **Sign risk**

$$\mathbb{E}\left[\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right)\right] = \frac{1}{p_\lambda(\tau)}\left[\left(\left\|p_\lambda\left(\boldsymbol{\beta}_0\right)\right\|_1 + s\lambda^2\right)O(n^{-c_1}) + O(p^{-c_1/2})\kappa_c\right]$$

- **Estimation and prediction risks**: If the pnnalty function further satisfies $p'_\lambda(\tau) = O\left(\sqrt{\frac{\log n}{n}}\right)$, then $\forall q \in [1, 2]$

$$\mathbb{E}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_q^q \le Cs\left(\frac{logn}{n}\right)^{q/2} \qquad \mathbb{E}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_\infty \le C\gamma_n^*\sqrt{\frac{\log n}{n}} \qquad \mathbb{E}\left[\frac{1}{n}D\left(\hat{\boldsymbol{\beta}}\right)\right] \le Cs\frac{\log n}{n}$$

where C is some positive constant.

How to understand Thm.14.2.3

- $\mathbb{E}\left[\mathrm{FS}\left(\hat{\boldsymbol{\beta}}\right)\right]$ converges to 0 at a polynomial rate of n
- Consistent with the risk bounds $O\left(\frac{s\log n}{n}\right)$ of the regularized estimators under the L_2-loss in wavelets setting with orthogonal design
- No additional cost in risk bounds for generalizing to the ultra-high dimensional nonlinear model setting of GLM

14.3 Computability and Implementation

These properties are quite nice, but what about the computability? Specifically, what if computable solutions produced by an algorithm are not actually the **global** minimizer?

Theorem 14.3.1: Asymptotic Properties of Computable solutions

Let $\hat{\boldsymbol{\beta}} \ = \in \mathcal{B}_\tau$ be a computable solution to the minimization problem produced by any algorithm that is the global minimizer when constrained on the subspace given by $\mathrm{supp}\left(\hat{\boldsymbol{\beta}}\right)$ and $\eta_n = \left\|\frac{1}{n}\mathbf{X}'\left[\mathbf{y} - \boldsymbol{\mu}\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)\right]\right\|_\infty$

References

Zemin Zheng, Yingying Fan, and Jinchi Lv. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 627–649, 2014.