Econometrics **March 19, 2023**

Topic 12: Non-convex Learning

by Sai Zhang

Key points: L-0 penalty is the best choice, but mostly computationally infeasible. Concave penalty (such as SCAD) works well with high dimensional problems.

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.*

12.1 L0 Penalized Likelihood

Consider the model selection problem of choosing a parameter vector $\boldsymbol{\theta}$ that maximizes the penalized likelihood

$$\mathcal{L}_n(\boldsymbol{\theta}) - \lambda \|\boldsymbol{\theta}\|_0 \tag{12.1}$$

where the L_0-norm $\|c\|_0$ denotes the **the number of nonzero components**, and $\lambda \geq 0$ is still the regularization parameter.

The L_0-penalized likelihood method is equivalent to **the best subset selection**

- given $\|\boldsymbol{\theta}_0\|_0 = m$, the solution to Problem 12.1 is the **best subset** that has the **largest** maximum likelihood among all subsets of size m
- then, choose the model size m among the p size-m best subsets ($1 \leq m \leq p$) by maximizing 12.1

hence it's a combinatorial problem, computationally complex.

L_0-Penalized Empirical Risk Minimization More generally, consider a unified approach of L_0-penalized empirical risk minimization for variable selection:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \hat{R}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 \right\} \tag{12.2}$$

where $\hat{R}(\boldsymbol{\theta})$ is the empirical risk function, which could be of different forms

- **negative log-likelihood loss**: equivalent t L_0-penalized likelihood
- **squared error (quadratic) loss**: L_0-penalized least squares
- selection via **RSS** (residual sum of squares): for the adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-d} \frac{RSS_d}{TSS}$$

it's clear that $\max R^2_{\text{adj}} \Leftrightarrow \min \log\left(\frac{RSS_d}{n-d}\right)$, and since $\frac{RSS_d}{n} \simeq \sigma^2$, then

$$n \log \frac{RSS_d}{n-d} \simeq \frac{RSS_d}{\sigma^2} + d + n(\log \sigma^2 - 1)$$

which shows that adjusted R^2 method is approximately equivalent to 12.2 with $\lambda = 1/2$

- **generalized corss-validation** (GCV), **corss-validation** (CV)
- **risk inflation factor** (RIC): use $\lambda = \log p$, adjusting for the inflation of prediction risk caused by searching p variables[1]
- **AIC** ($\lambda = 1$), **BIC** ($\lambda = \frac{\log n}{2}$)

12.1.1 Properties of L0-Regularization Methods

risk bounds for model selection (Barron et al., 1999): for a family of models $\{S_m : m \in \mathcal{M}_p\}$, The penalty term generally takes the form of

$$\frac{\kappa L_m D_m}{n}$$

where

- κ: a positive constant
- $D_m = |S_m|$: the model dimension, account for the difficulty to estimate **within** the model S_m
- $L_m \geq 1$: a weight that satisfies: $\sum_{m \in \mathcal{M}_p} \exp(-L_m D_m) \leq 1$, accounting for the noise due to **the size** of the list of models

hence, in the linear model, the L_0-regularized estimator $\hat{\boldsymbol{\beta}}$ satisfies that

$$\mathbb{E}\left[n^{-1}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2\right] \leq C \inf_{m \in \mathcal{M}_p}\left\{\min_{\boldsymbol{\beta} \in \text{model } S_m}\left[n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2\right] + \frac{\kappa L_m D_m}{n}\right\}$$

where ***the tradeoff***: approximation error $n^{-1}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2$, and the cost of searching $\frac{\kappa L_m D_m}{n}$

computational complexity L_0-regularization methods are appealing w.r.t. risk properties, but in high-dimensional settings, the computation is infeasible (combinatorial), and discontinuous, non-convex penalty function $\lambda\|\boldsymbol{\beta}\|_0$

12.1.2 Generalizations of L0-Regularization Methods

Consider continuous or convex relaxation of the L_0-regularization method

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p}\left\{\hat{R}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda\left(|\beta_j|\right)\right\} \tag{12.3}$$

where, as in Problem 12.2

- $\hat{R}(\boldsymbol{\beta})$: the empirical risk function
- $p_\lambda(t), t \geq 0$: the nonnegative penalty function indexed by the regularization parameter $\lambda \geq 0$ with $p_\lambda(0) = 0$

[1]The $\log p$ is, once again, from the fact that for Gaussian random variables

$$\max_{1 \leq j \leq p} |Z_i| \simeq \sqrt{2 \log p}$$

for $(Z_1, \cdots, Z_p)' \sim \mathcal{N}(0, \mathbf{I}_p)$

Choices of penalty function In general, the choices of penalty function can be up for the researchers to decide. Fan and Li (2001) proposed 3 criteria for the selection of penalty function $p_\lambda(t)$

- **Sparsity**: $p'_\lambda(0+) > 0$, sets small coefficients to 0, for *variable selection* and *reducing model complexity*
- **Approximate unbiasedness**: nearly unbiased, especially when the true coefficient β_j is large
- **Continuity**: continuous in data to reduce instability in model selection

To elaborate the 3 criterion, consider a class of penalty function, L_q-penalty

$$p_\lambda(t) = \lambda t^q, t \geq 0 \Rightarrow p'_\lambda(t) = \lambda q t^{q-1}$$

then we can compare

	Sparsity	Approx. unbiasedness	Continuity
$0 < q < 1$	Y	Y	N
$q = 1$	Y	N	Y
$1 < q \leq 2$	N	N	Y

this class of penalty functions includes:

- $q = 0$: L_0- regression (best subset selection)
- $q = 1$: Lasso
- $q = 2$: Ridge
- $0 < q < 2$: Bridge estimator

12.2 High Dimensional Variable Selection

For a generalized linear model

$$f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ c(y_i) \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi} \right) \right\}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)' = \mathbf{X}\boldsymbol{\beta}$ is the **natural parameter vector**, which can a very challenging problem. Instead of the penalized least squares, now we examine **penalized likelihood**

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) - \sum_{j=1}^{p} p_{\lambda_n}\left(|\beta_j|\right) \tag{12.4}$$

where $\mathcal{L}_n(\boldsymbol{\beta}) = n^{-1}\left[\mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{1}'\mathbf{b}(\mathbf{X}\boldsymbol{\beta})\right]$ is the *affine transformation* of log-likelihood,

$$\mathbf{b}(\boldsymbol{\theta}) = \mathbf{b}(\mathbf{X}\boldsymbol{\beta}) = (b(\theta_1), \cdots, b(\theta_n))'$$

So the natural question is, when can we find the solution to Problem 12.4, s.t. $\text{supp}(\hat{\boldsymbol{\beta}}) = \text{supp}(\boldsymbol{\beta}_0)$, that is, covering exactly the ture underlying sparse model?

12.3 Penalized Likelihood with Concave Penalties

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{L}_n(\boldsymbol{\beta}) - \sum_{j=1}^{p} p_{\lambda_n}\left(|\beta_j|\right)$$

where $\mathcal{L}_n(\boldsymbol{\beta}) = n^{-1}\left[\mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{1}'\mathbf{b}(\mathbf{X}\boldsymbol{\beta})\right]$, and $p_\lambda(\cdot)$ is a concave penalty function. Let $\rho(t;\lambda) = \lambda^{-1}p_\lambda(t), t \geq 0$, we aim for penalty functions that satisfy

- $\rho(t)$ is **increasing and concave** in t
- $\rho'(t)$ is **continuous** with $\rho'(0+) > 0$
- if $\rho(t)$ depends on λ, $\rho'(t;\lambda)$ is **increasing** in λ and $\rho'(0_+;\lambda)$ is **independent** of λ

Here are some notations

- Moment property: $k-$th component-wise derivative corresponds to $k-$th moment
  - $\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \cdots, b'(\theta_n))'$: $\mathbb{E}(\mathbf{y})$
  - $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}\{b''(\theta_1), \cdots, b''(\theta_n)\}$
- local concavity of ρ at $\mathbf{v} = (v_1, \cdots, v_q)' \in \mathbb{R}^q$, with $\|\mathbf{v}\|_0 = q$, that is

$$\kappa(\rho;\mathbf{v}) = \lim_{\epsilon\to 0_+} \max_{1\leq j\leq q} \sup_{t_1<t_2\in(|v_j|-\epsilon,|v_j|+\epsilon)} -\frac{\rho'(t_2)-\rho'(t_1)}{t_2-t_1}$$

  if $\rho''(t)$ is continuous, this becomes

$$\max_{1\leq j\leq q} -\rho''(|v_j|)$$

And the solution is given by the following theorem

Theorem 12.3.1: Penalized Likelihood estimator

$\hat{\boldsymbol{\beta}}$ is **strict local** maximizer of penalized likelihood if

$$\begin{aligned}\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}) - n\lambda_n \text{sign}(\hat{\boldsymbol{\beta}}_1)\circ\rho'(|\hat{\boldsymbol{\beta}}_1|) &= \mathbf{0}\\ \|(n\lambda_n)^{-1}\mathbf{X}_2'[\mathbf{y}-\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})]\|_\infty &< \rho'(0_+)\\ \lambda_{\min}\left[\mathbf{X}_1'\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\mathbf{X}_1\right] &> n\lambda_n\kappa(\rho;\hat{\boldsymbol{\beta}}_1)\end{aligned}$$

where $\circ$ is the component-wise multiplication, $\lambda_{\min}(\cdot)$ is the smallest eigenvalue.

12.3.1 Global Optimality

Theorem 12.3.1 gives the rule to find local maximizers, but what about global optimality?

Proposition 12.3.2: Global Optimality of Penalized Likelihood Estimator

Assume that $\mathbf{X}$ has rank p, and satisfies

$$\min_{\boldsymbol{\beta}\in\mathcal{L}_c} \lambda_{\min}\left[n^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}\right] \geq \kappa(p_{\lambda_n})$$

where

- NOT high-dimensional: $p \leq n$
- for some $c < \mathcal{L}_n(\mathbf{0})$,

$$\mathcal{L}_c = \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \mathcal{L}_n(\boldsymbol{\beta}) \geq c\right\}$$

  is a sublevel set of $-\mathcal{L}_n(\boldsymbol{\beta})$
- maximum concavity

$$\kappa(p_\lambda) = \sup_{t_1 < t_2 \in (0,\infty)} -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1}$$

12.3.2 SCAD penalty

Now, consider a penalized likelihood model: **SCAD** (Fan and Li, 2001, smoothly clipped absolute deviation)

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + p(\boldsymbol{\beta})$$

where the derivative of the penalty function

$$p_\lambda^{\text{SCAD}}(\beta_j) = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda \\ -\left(\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \lambda < |\beta_j| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda \end{cases}$$

and its derivative

$$p'(\boldsymbol{\beta}) = \lambda\left[I(\boldsymbol{\beta} \leq \lambda) + \frac{(a\lambda - \boldsymbol{\beta})_+}{(a-1)\lambda} I(\boldsymbol{\beta} > \lambda)\right]$$

the solution to SCAD penalty model is

$$\hat{\beta}_j^{\text{SCAD}} = \begin{cases} (|\hat{\beta}_j|)_+ \text{sign}(\hat{\beta}_j) & |\hat{\beta}_j| < 2\lambda \\ \frac{(a-1)\hat{\beta}_j - \text{sign}(\hat{\beta}_j)a\lambda}{a-2} & 2\lambda < |\hat{\beta}_j| \leq a\lambda \\ \hat{\beta}_j & |\hat{\beta}_j| > a\lambda \end{cases}$$

the SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, singular at 0. SCAD has some great properties, one of them is robustness.

Proposition 12.3.3: Robustness of SCAD

Assume that $\mathbf{X}$ has rank $p = s$, and $\exists c < \mathcal{L}_n(\mathbf{0})$ s.t. for some $c_0 > 0$

$$\min_{\boldsymbol{\beta} \in \mathcal{L}_c} \lambda_{\min}\left[n^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}\right] \geq c_0$$

then the SCAD penalized likelihood estimator $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ is the **global** maximizer and equals the oracle MLE $\boldsymbol{\beta}^*$, if $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ and

$$\min_{j=1}^{p} |\hat{\beta}_j^{\text{SCAD}}| > \left(a + \frac{1}{2c_0}\right)\lambda_n$$

Next, we extend this global optimality result to high-dimensional cases, where $p > n$

Proposition 12.3.4: Global Optimality, $p > n$

On the union of all $s-$dimensional coordinate subspaces of $\mathbb{R}^p$

- Under Proposition 12.3.2 for each $n \times 2s$ submatrix of $\mathbf{X}$, then the NCPMLE $\hat{\boldsymbol{\beta}}$ is a global maximizer on $\mathbb{S}_s$
- Under Proposition 12.3.3 for $n \times s$ submatrix of $\mathbf{X}$ formed by columns in $\text{supp}(\boldsymbol{\beta}_0)$, the true model is δ-identifiable for some $\delta > \frac{(a+1)s\lambda_n^2}{2}$, and $\text{supp}(\hat{\boldsymbol{\beta}}) = \text{supp}(\boldsymbol{\beta}_0)$. Then the SCAD penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ is the global maximizer on $\mathbb{S}_s$ and **equals** to the oracle MLE $\boldsymbol{\beta}^*$

12.3.3 Regularity Conditions for Concave Penalties

The regularity conditions for concave penalty are

- the true sub design matrix $\mathbf{X}_1$ should be well conditioned

$$\left\| \left[\mathbf{X}_1' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty = O(b_s n^{-1})$$

- A generalized version of the irrepresentable condition

$$\left\| \mathbf{X}_2' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1 \left[\mathbf{X}_1' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty \leq \min \left\{ C \frac{\rho'(0+)}{\rho'(d_n)}, O(n^{\alpha_1}) \right\} \tag{12.5}$$

- also

$$\max_{\boldsymbol{\delta} \in \mathcal{N}_0} \max_{j=1}^{p} \lambda_{\max} \left[\mathbf{X}_1' \text{diag} \left\{ |\mathbf{x}_j| \circ |\boldsymbol{\mu}''(\mathbf{X}_1 \boldsymbol{\delta})| \right\} \mathbf{X}_1 \right] = O(n)$$

Here, $b_s \to \infty$ with $s = \|\boldsymbol{\beta}_0\|_0 = O(n^{\alpha_0})$, $\alpha_1 \in [0, 1/2]$, $C \in (0,1)$, $\mathcal{N}_0 = \left\{ \boldsymbol{\delta} \in \mathbb{R}^s : \|\boldsymbol{\delta} - \boldsymbol{\beta}_1\|_\infty \leq d_n \right\}$; $\alpha = \min(\frac{1}{2}, 2\gamma - \alpha_0) - \alpha_1$, $d_n \geq n^{-\gamma} \log n$ for some $\gamma \in (0, 1/2]$.

Notice that in a linear model, Condition 12.5 becomes

$$\left\| \mathbf{X}_2' \mathbf{X}_1 \left(\mathbf{X}_1' \mathbf{X}_1 \right)^{-1} \right\|_\infty \leq \min \left\{ C \frac{\rho'(0+)}{\rho'(d_n)}, O(n^{\alpha_1}) \right\}$$

- For L_1 penalty, this becomes ($\rho'(0+) = \rho'(d_n) = 1$) a **stronger** form of the irrepresentable condition

$$\left\| \mathbf{X}_2' \mathbf{X}_1 \left(\mathbf{X}_1' \mathbf{X}_1 \right)^{-1} \right\|_\infty \leq C < 1$$

  , this speaks about the restrictive nature of L_1 penalty in higher dimensions
- For concave penalty, $\frac{\rho'(0+)}{\rho'(d_n)}$ **can grow** to ∞, hence, it is a much weaker condition: **the flexibility** of concave penalty.

12.3.4 Properties of Concave Penalty

Next, we establish the nonasymptotic weak oracle property for estimator with concave penalties.

Theorem 12.3.5: Nonasymptotic Weak Oracle Property

Under some regularity conditions, $s = o(n)$ and $\log p = O(n^{1-2\alpha})$, there exists a penalized likelihood estimator $\boldsymbol{\beta}$ s.t. for sufficiently large n, with probability of at least

$$1 - 2\left[sn^{-1} + (p-s)e^{-n^{1-2\alpha}\log n}\right]$$

$\hat{\boldsymbol{\beta}}$ satisfies

- Sparsity: $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$
- L_∞ loss: $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_\infty = O(n^{-\gamma} log n)$

This theorem shows that concave penalties can reduce biases of estimates. The L_∞ estimation loss can de decomposed into $L_\infty \leq h_1 + h_2 + h_3$, $h_2 \sim b_s \lambda_s \frac{\rho'(d_n)}{\rho'(0+)}$. Theorem 12.3.5 establishes nonasymptotic weak oracle property of penalized likelihood estimator with penalties, where dimensionality p can grow non-polynomially with sample size n.

Theorem 12.3.6: Non-Concave Penalized Likelihood Estimator

Under some regularity conditions, $s \ll n$ and $\log p = O(n^\alpha)$ for some $\alpha \in (0, 1/2)$, there exists a strict local maximizer $\hat{\boldsymbol{\beta}}$ of penalized likelihood such that $that\boldsymbol{\beta}_2 = \mathbf{0}$ with probability tending to 1 as $n \to \infty$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_P(\sqrt{s}n^{-1/2})$.

These conditions are incompatible for L_1 penalty, suggesting that L_1 penalized likelihood estimator generally **cannot** achieve consistency rate $O_P(\sqrt{s}n^{-1/2})$ and **does not** have oracle property, when dimensionality p is diverging with sample size n.

Theorem 12.3.7: Oracle Property of Non-Concave Penalty

Under some regularity conditions and $s = o(n^{1/3})$, then with probability tending to 1 as $n \to \infty$, then non-concave penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ in Theorem 12.3.6 must satisfies

- **Sparsity**
$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$$
- **Asymptotic normality**
$$\mathbf{A}_n \left[\mathbf{X}_1' \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{X}_1\right]^{1/2} (\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_1) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \phi \mathbf{G})$$
where $\mathbf{A}_n$ is a $q \times s$ matrix s.t. $\mathbf{A}_n \mathbf{A}_n' \to \mathbf{G}$, and $\mathbf{G}$ is a $q \times q$ symmetric positive definite matrix.

A simulation of SCAD versus Lasso is presented in Figure 12.1 ($n = 100, p = 1000$). It's clear that SCAD gives a more consistent estimation of the model. In this simulation, all covariates (1000) are independent from each other. If there are some covariates correlated with each other, SCAD would be robustly consistent while Lasso performs even worse.

References

Andrew Barron, Birgé Lucien, and Massart Pascal. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Figure 12.1: Numercial Comparison: SCAD vs Lasso