ST509 Computational Statistics

Lecture 7: Nonconvex Penalty

Seung Jun Shin

Department of Statistics
Korea University

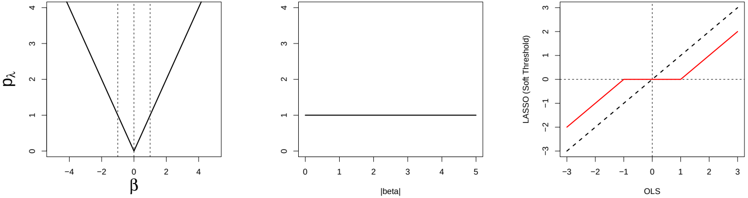
E-mail: sjshin@korea.ac.kr

Analysis of Penalty Function I

- For orthonormal predictor x_i and centered y_i, LASSO solves

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda |\beta| = S_\lambda(\hat{\beta}_{\text{ols}})$$

with $\hat{\beta}_{\text{ols}} = \sum x_i y_i$.



(a) Penalty

(b) Derivative

(c) Comparison to OLS

Analysis of Penalty Function II

- Fan & Li (2001) claim that a good penalty should possess:
  - Unbiasedness: unbiased when $|\beta|$ is large.
  - Sparsity: shrink small estimates to zero.
  - Continuity: continuous with respect to the data.

Analysis of Penalty Function III

- Consider

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|) \tag{1}$$

  where $p_j(\cdot)$ denotes a penalty function.

- Let $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{X}\mathbf{X}^T\mathbf{y}$, then (1) becomes

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|) \\
=&\frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \frac{1}{2}\sum_{j=1}^{p}(z_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|)
\end{aligned}$$

Analysis of Penalty Function IV

- Equivalent to consider the following componentwise problem:

$$\frac{1}{2}(z-\beta)^2 + p_\lambda(|\beta|) \tag{2}$$

where $p_\lambda(\beta) = \lambda \cdot p_j(\beta)$ with j being suppressed.

- Subgradient equation of (2) is

$$\mathrm{sign}(\beta)\left\{|\beta| + p'_\lambda(|\beta|)\right\} - z = 0$$

where p'_λ denotes the derivative of p_λ.

Analysis of Penalty Function V

- Conditions:
  1. Unbiasedness: $p'_\lambda(|\beta|) = 0$ for large $|\beta|$.

     $\Rightarrow \hat{\beta} = z$ for large β.

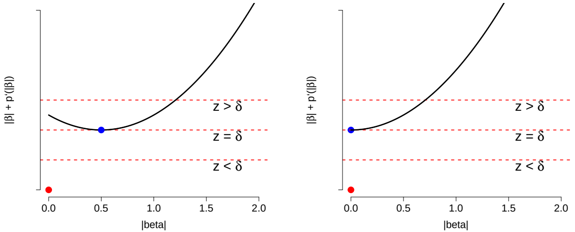
  2. Sparsity: $\min_{\beta \neq 0}\{|\beta| + p'_\lambda(|\beta|)\} = \delta > 0$.
     When $|z| < \delta$, the derivative of (1) is positive (negative) for a positive (negative) β.

     $\Rightarrow \hat{\beta} = 0$ if $|z| < \min_{\beta \neq 0}\{|\beta| + p'_\lambda(|\beta|)\}$.

  3. Continuity: $\operatorname{argmin}_{|\beta|}\{|\beta| + p'_\lambda(|\beta|)\} = 0$.

     $\Leftrightarrow \hat{\beta}$ is continuous function of z. (See Figure 2 in the following)

Analysis of Penalty Function VI



(d) Not Continuous　　　(e) Continuous

Figure: Red circles represent the solution when $z < \delta$, and blue circles the solution when $z > \delta$. It the minimum of $\{|\beta| + p'(|\beta|)\}$ is not attained at $|\beta| = 0$, then $\hat{\beta}$ is not a continuous function of z.

- Ridge with $p_\lambda(|\beta|) = \lambda|\beta|^2$ and $p'(|\beta|) = 2\lambda\beta$ violates 1 and 2.
- LASSO with $p_\lambda(|\beta|) = \lambda|\beta|$ and $p'(|\beta|) = \lambda \text{sign}(\beta)$ violates 1.

Analysis of Penalty Function VII

- In order to mitigate the bias of LASSO, Zou (2006) proposed the adaptive LASSO that solves

$$\sum_{i=1}^{n} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|$$

  where

$$\hat{w}_j \propto |\hat{\beta}_{\mathrm{ols},j}|^{\gamma}$$

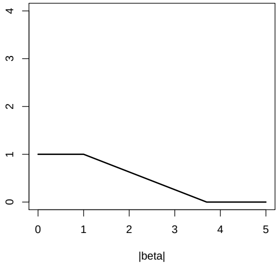
- Larger $|\beta_j|$ penalized less.
- Computation is exactly the same as that of LASSO!
- Zhang & Lu (2007) independently proposed the same idea to Cox-proportional hazard model for survival data.

SAD Penalty I

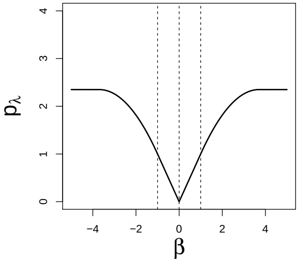
- Fan & Li (2001) propose *Smoothly Clipped Absolution Deviation (SCAD)* penalty defined through its derivative

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

  for some $a > 2$ and $\beta > 0$.



(a) Derivative ($p'_\lambda(|\beta|)$)



(b) SCAD Penalty ($p_\lambda(\beta)$)

SAD Penalty II

- One-dimensonal orthogonal solution is

$$\hat{\beta}_{\text{scad}} = \begin{cases} \text{sign}(\hat{\beta}_{\text{ols}})(|\hat{\beta}_{\text{ols}}| - \lambda)_+, & \text{when } |\hat{\beta}_{\text{ols}}| \le 2\lambda; \\ \{(a-1)\hat{\beta}_{\text{ols}} - \text{sign}(\hat{\beta}_{\text{ols}})a\lambda\}/(a-2), & \text{when } 2\lambda < |\hat{\beta}_{\text{ols}}| \le a\lambda; \\ \hat{\beta}_{\text{ols}}, & \text{when } |\hat{\beta}_{\text{ols}}| > a\lambda. \end{cases}$$

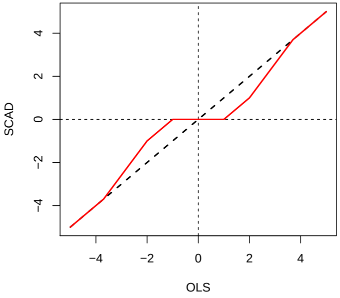


Figure: Comparison to $\hat{\beta}_{\text{ols}}$ in the one-dimensional orthonormal regression.

SAD Penalty III

- Let S denote the index set of true informative variables, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_S, \boldsymbol{\beta}_{S^c})^T$ without loss of generality.
- The SCAD-penalized estimator, $\hat{\boldsymbol{\beta}}_{\text{scad}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^c})^T$ achieves the Oracle property:
  1. (Selection Consistency)
  $$P(\hat{\boldsymbol{\beta}}_{S^c} = 0) \to 1, \qquad \text{as } n \to \infty.$$
  2. (Asymptotic Normality)
  $$\sqrt{n}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) \sim N(\mathbf{0}, \mathbf{I}_S^{-1})$$
  where $\mathbf{I}_S^{-1}$ denotes the information matrix for X_S.
- Oracle property states that $\boldsymbol{\beta}_{\text{scad}}$ asymptotically behaves as if we know S in advance.

Computation of SAD Penalty I

- Let $\beta^{(t)}$ as the current value at tth iteration.
- We can let $\beta^{(t+1)}$ as 0 if $\beta^{(t)}$ is sufficiently closed to 0.
- For $\beta^{(t)} \neq 0$,

$$[p_\lambda(|\beta|)]' = p'_\lambda(|\beta|)\text{sign}(\beta) = \frac{p'_\lambda(|\beta|)}{|\beta|}\beta \approx \frac{p'_\lambda(|\beta^{(t)}|)}{|\beta^{(t)}|}\beta,$$

- The 2nd order Taylor expansion of $p_\lambda(|\beta|)$ at $\beta = \beta^{(t)}$ gives

$$\begin{aligned} p_\lambda(|\beta|) &\approx p_\lambda(|\beta^{(t)}|) + \left[p_\lambda(|\beta^{(t)}|)\right]'(\beta - \beta^{(t)}) + \frac{1}{2}\left[p_\lambda(|\beta^{(t)}|)\right]''(\beta - \beta^{(t)})^2 \\ &= p_\lambda(|\beta^{(t)}|) + \frac{p'_\lambda(|\beta^{(t)}|)}{2|\beta^{(t)}|}(\beta^2 - \beta^{(t)2}) \end{aligned}$$

Computation of SAD Penalty II

- (LQA) Updating equation is

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} w_j \beta_j^2 \right\}$$

  where

$$w_j = \frac{p'_\lambda(|\beta_j^{(t)}|)}{2|\beta_j^{(t)}|}.$$

- Set $\beta_j^{(t)} = 0$ if $|\beta_j^{(t)}| < \epsilon_0$ for pre-specified small ϵ_0. This is equivalent to remove $\mathbf{x}_j$.
- Once $\mathbf{x}_j$ is removed then it never can be come back to the model (just like backward deletion).

Computation of SAD Penalty III

- Applying 1st order Taylor Expansion,

$$p_\lambda(|\beta|) \approx p_\lambda(|\beta^{(t)}|) + p'_\lambda(|\beta^{(t)}|)(|\beta| - |\beta^{(t)}|)$$

- (LLA) Updating equation is

$$\boldsymbol{\beta}^{(t+1)} = \operatorname{argmin}\left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} w_j|\beta_j|\right\}$$

where

$$w_j = p'_\lambda(|\beta_j^{(t)}|).$$

Computation of SAD Penalty IV

- LQA and LLA are MM algorithm.
- LLA produces the best convex majoring function of the SCAD penalty.

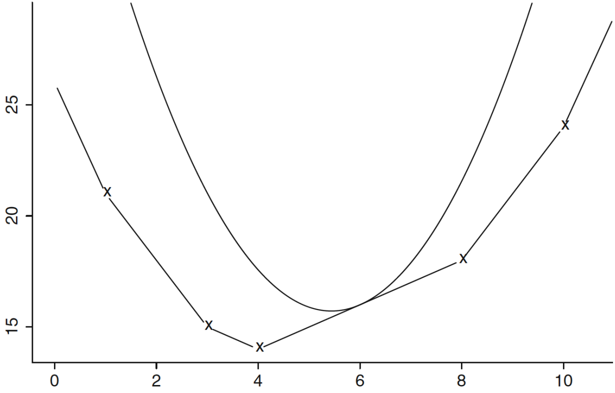


Figure: Plot of local quadratic approximation (thin dotted lines) and local linear approximation (thick broken lines) at $\beta = 4$ and 1

Computation of SAD Penalty V

- Due to such optimality of LLA, Zou & Li (2008) showed that

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^0|)|\beta_j|\right\}$$

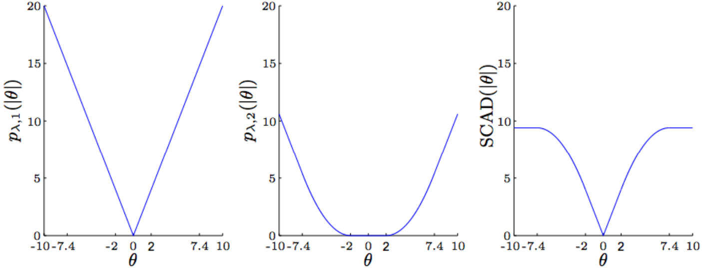
  achieves the Oracle property, where $\hat{\boldsymbol{\beta}}^0$ is any consistent estimator of $\boldsymbol{\beta}$ (eg. MLE, ...).

- This is known as the One-step Estimator.

Computation of SAD Penalty VI

- SCAD penalty can be decomposed as a difference of two convex functions $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$, where

$$p'_{\lambda,1}(\theta) = \lambda \qquad \text{and} \qquad p'_{\lambda,2}(\theta) = \lambda \left\{ 1 - \frac{(a\lambda - \theta)_+}{(a-1)\lambda} \right\} \mathbb{1}(\theta > \lambda)$$



Computation of SAD Penalty VII

- The objective function of the SCAD penalized regression

$$Q(\boldsymbol{\beta}) = \underbrace{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_{\lambda,1}(|\beta_j|)}_{Q_{\text{vex}}(\boldsymbol{\beta})} + \underbrace{\left\{-\sum_{j=1}^{p} p_{\lambda,2}(|\beta_j|)\right\}}_{Q_{\text{cav}}(\boldsymbol{\beta})}$$

- Difference Convex (DC) algorithm iteratively solves

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{Q_{\text{vex}}(\boldsymbol{\beta}) + \left\langle Q'_{\text{cav}}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right\rangle\right\}$$

- (DC) Updating equation for SCAD-penalized regression is

$$\boldsymbol{\beta}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p}|\beta_j| - \sum_{j=1}^{p} w_j\beta_j\right\} \tag{3}$$

where

$$w_j = p'_{\lambda,2}(|\beta_j^{(t)}|)\text{sign}\{\beta_j^{(t)}\}.$$

Minmax Concave Penalty I

- Zhang (2010) proposed the minimax concave penalty (MCP) defined as

$$p'_\lambda(|\beta|) = \begin{cases} \lambda - |\beta|/a, & \text{if } |\beta| \le a\lambda \\ 0, & \text{if } |\beta| > a\lambda \end{cases}$$

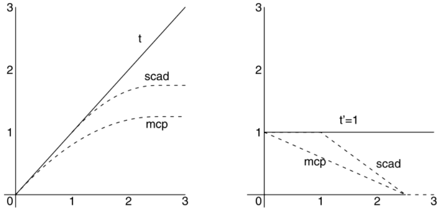


FIG. 1. *The ℓ_1 penalty $\rho_1(t) = t$ for the LASSO along with the MCP $\rho_2(t)$ and the SCAD penalty $\rho_3(t)$, $t > 0$, $\gamma = 5/2$. Left: penalties $\rho_m(t)$. Right: their derivatives $\dot{\rho}_m(t)$.*

Minmax Concave Penalty II

- For one-dimensional orthogonal regression, $\hat{\beta}_{\text{mcp}}$ that minimizes

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta x_i)^2 + p_\lambda(|\beta|)$$

is

$$\hat{\beta}_{\text{ols}} = \begin{cases} S_\lambda(\hat{\beta}_{\text{ols}})/(1-1/a), & \text{if } |\beta| \le a\lambda; \\ \hat{\beta}_{\text{ols}}, & \text{if } |\beta| > a\lambda. \end{cases}$$

Coordinate Decent Algorithm I

- Breheny & Huang (2011) proposed the CD algorithm for non-convex penalties such as MCP and SCAD.
- Due to the non-convexity, convergence of CD algorithm is not trivial.
- Pathwise coordinate decent can be adopted.
- `ncvreg` package in `R`

Coordinate Decent Algorithm II

1. After the marginal standardization ($\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n$), initialize $\boldsymbol{\beta}^{(t)} = \mathbf{0}$ and $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)}$.
2. Calculate partial slope of $\mathbf{x}_j$ (unpenalized LS solution)

$$z_j = \frac{1}{n}\mathbf{x}_j^T \mathbf{r}_{-j} = \frac{1}{n}\mathbf{x}_j^T \mathbf{r} + \beta^{(t)}$$

3. Repeat until convergence:

   3.1 Update $\beta_j^{(t+1)}, j = 1, \cdots, p$

$$\text{(SCAD)} \quad \beta_j^{(t+1)} \leftarrow \begin{cases} S(z_j, \lambda), & \text{if } |z_j| < 2\lambda \\ S(z_j, a\lambda/(a-1))/(1 - 1/(a-1)), & \text{if } 2\lambda < |z_j| \le a\lambda \\ z_j, & \text{if } |z_j| > a\lambda \end{cases}$$

$$\text{(MCP)} \quad \beta_j^{(t+1)} \leftarrow \begin{cases} S(z_j, \lambda)/(1 - 1/\gamma), & \text{if } |z_j| \le a\lambda \\ z_j, & \text{if } |z_j| > a\lambda \end{cases}$$

   3.2 Update residuals

$$\mathbf{r} \leftarrow \mathbf{r} - \left(\beta_j^{(t+1)} - \beta_j^{(t)}\right)\mathbf{x}_j.$$

Algorithm 1: Coordinate Decent Algorithm for the non-convex penalty.

Reference

- Fan & Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties JASA, 96(456), 1348-1360.
- Zou (2006) The adaptive lasso and its oracle properties, JASA, 101(476), 1418-1429.
- Zhang & Lu (2007) Adaptive Lasso for Cox's proportional hazards model Biometrika, 94(3), 691-703.
- Zou & Li (2008) One-step sparse estimates in nonconcave penalized likelihood models AOS, 36(4), 1509.
- Zhang (2010) Nearly unbiased variable selection under minimax concave penalty AOS, 38(2),894-942.
- Breheny & Huang (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection AOAS, 5(1), 232.