

SCAD: Smoothly Clipped Absolute Deviation Penalty

Ganchao Wei

November 3, 2021

Overview

- 1 Penalized Least Squares and Variable Selection
- 2 Variable Selection via Penalized Likelihood
- 3 Simulation
- 4 Simulation
- 5 Simulation
- 6 Simulation
- 7 Simulation
- 8 Application

Penalties

Consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Denote $\mathbf{z} = \mathbf{X}^T \mathbf{y}$ and let $\hat{\mathbf{y}} = \mathbf{X}\mathbf{X}^T \mathbf{y}$. Then a form of the penalized least square squares is

$$\begin{aligned} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^d p_j(|\beta_j|) &= \frac{1}{2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \frac{1}{2} \sum_{j=1}^d (z_j - \beta_j)^2 \\ &\quad + \lambda \sum_{j=1}^d p_j(|\beta_j|). \quad (2.2) \end{aligned}$$

Further, denote $p_\lambda(|\cdot|) = \lambda p(|\cdot|)$. Minimizing (2.2) leads us to consider the penalized least squares problem

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

Penalties

A good penalty function should have 3 properties:

- **Unbiasedness:** The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.
- **Sparsity:** The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.
- **Continuity:** The resulting estimator is continuous in data z to avoid instability in model prediction.

What are conditions to satisfy these properties?

Penalties

First order derivative for $\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|)$ w.r.t. θ is $\text{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z$

- sufficient condition for **Unbiasedness**: $p'_\lambda(|\theta|) = 0$ for large $|\theta|$
- sufficient condition for **Sparsity**: $\min(|\theta| + p'_\lambda(|\theta|)) > 0$
- iff condition for **Continuity**: $\min(|\theta| + p'_\lambda(|\theta|))$ is attained at 0

Let use these criteria to evaluate different penalties: (1) Hard thresholding penalty; (2) L_q penalty $p_\lambda(|\theta|) = \lambda|\theta|^q$ (LASSO when $q = 1$); (3) SCAD penalty

Penalties

Three penalties & corresponding thresholding functions.

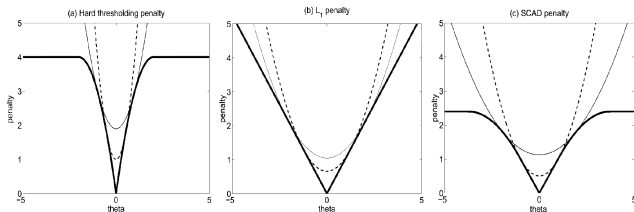


Figure 1. Three Penalty Functions $p_\lambda(\theta)$ and Their Quadratic Approximations. The values of λ are the same as those in Figure 5(c).

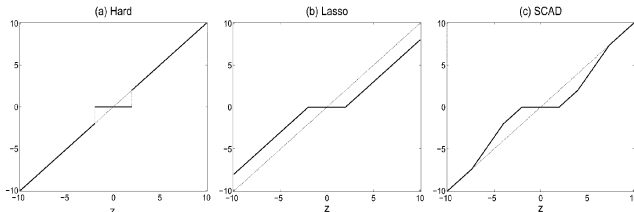


Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

Hard thresholding: not continuous; LASSO: bias

Discussion on L_q penalty

Continuous only when $q \geq 1$. But when $q > 1$, $\min(|\theta| + p'_\lambda(|\theta|)) = 0 \rightarrow$ only when $q = 1$, sparse and continuous

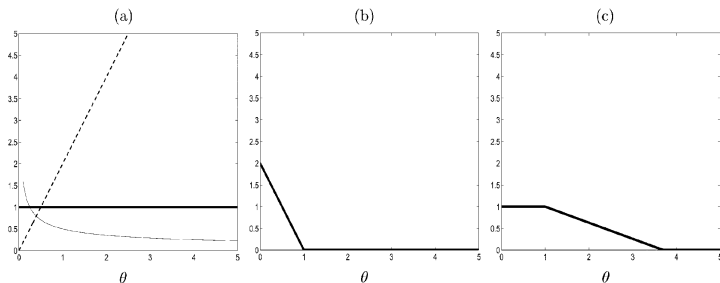


Figure 4. Plot of $p'_\lambda(\theta)$ Functions Over $\theta > 0$ (a) for L_q Penalties, (b) the Hard Thresholding Penalty, and (c) the SCAD Penalty. In (a), the heavy line corresponds to L_1 , the dash-dot line corresponds to L_2 , and the thin line corresponds to L_3 penalties.

However, as shown previously, it's bias...

Smoothly Clipped Absolute Deviation Penalty

The SCAD is defined by the following first order derivative:

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

for some $a > 2$ and $\theta > 0$, (2.7)

The resulting solution is given by

$$\hat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda, \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda, \\ z, & \text{when } |z| > a\lambda \end{cases}$$

Tuning Parameters for SCAD

Can be chosen by cross-validation, but computationally expensive. Can implement tools in Bayesian risk analysis. Assume the prior is $\theta \sim N(0, a\lambda)$. Calculate the Bayes risk for $\lambda = \sqrt{2 \log(d)}$ for $d = 512, 1024, 2048, 4096$. It seems $a \approx 3.7$ is fine for all cases.

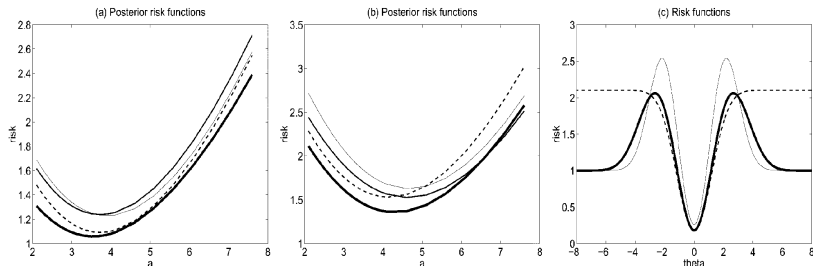


Figure 5. Risk Functions of Proposed Procedures Under the Quadratic Loss. (a) Posterior risk functions of the SCAD under the prior $\theta \sim N(0, a\lambda)$ using the universal thresholding $\lambda = \sqrt{2 \log(d)}$ for four different values d : heavy line, $d = 20$; dashed line, $d = 40$; medium line, $d = 60$; thin line, $d = 100$. (b) Risk functions similar to those for (a): heavy line, $d = 572$; dashed line, $d = 1,024$; medium line, $d = 2,048$; thin line, $d = 4,096$. (c) Risk functions of the four different thresholding rules. The heavy, dashed, and solid lines denote minimum SCAD, hard, and soft thresholding rules, respectively.

Penalized Least Square and Likelihood

- Linear Regression:

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^d p_{\lambda}(|\beta_j|)$$

- Robust Regression:

$$\sum_{i=1}^n \psi(|y_i - \mathbf{x}_i \boldsymbol{\beta}|) + n \sum_{j=1}^d p_{\lambda}(|\beta_j|)$$

- GLM:

$$\sum_{i=1}^n \ell_i(g(\mathbf{x}_i^T \boldsymbol{\beta}), y_i) - n \sum_{j=1}^d p_{\lambda}(|\beta_j|)$$

Sampling Properties and Oracle Properties

Firstly, some notations:

- Let $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$, where $\beta_{20} = 0$. Let $I(\beta_0)$ be Fisher information and let $I(\beta_{10}, 0)$ be the Fisher information knowing $\beta_{20} = 0$.
- Set $\mathbf{V}_i = (\mathbf{X}_i, Y_i)$. Let $L(\beta)$ be the log-likelihood and $Q(\beta)$ be the penalized likelihood.

Then we can first show the existence of a penalized likelihood estimator converging at rate $O_p(n^{-1/2} + a_n)$

Theorem 1. Let $\mathbf{V}_1, \dots, \mathbf{V}_n$ be independent and identically distributed, each with a density $f(\mathbf{V}, \beta)$ (with respect to a measure μ) that satisfies conditions (A)–(C) in the Appendix. If $\max\{|p''_{\lambda_n}(|\beta_{j0}|)|: \beta_{j0} \neq 0\} \rightarrow 0$, then there exists a local maximizer $\hat{\beta}$ of $Q(\beta)$ such that $\|\hat{\beta} - \beta_0\| = O_p(n^{-1/2} + a_n)$, where a_n is given by (3.4).

Sampling Properties and Oracle Properties

The oracle property:

Theorem 2 (Oracle Property). Let $\mathbf{V}_1, \dots, \mathbf{V}_n$ be independent and identically distributed, each with a density $f(\mathbf{V}, \boldsymbol{\beta})$ satisfying conditions (A)–(C) in Appendix. Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.5). If $\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n \rightarrow \infty$ as $n \rightarrow \infty$, then with probability tending to 1, the root- n consistent local maximizers $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ in Theorem 1 must satisfy:

- (a) Sparsity: $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.
- (b) Asymptotic normality:

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\mathbf{b}\} \rightarrow N\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\}$$

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

As a consequence, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is

$$\frac{1}{n}\{I_1(\boldsymbol{\beta}_{10}) + \Sigma\}^{-1}I_1(\boldsymbol{\beta}_{10})\{I_1(\boldsymbol{\beta}_{10}) + \Sigma\}^{-1},$$

which approximately equals $(1/n)I_1^{-1}(\boldsymbol{\beta}_{10})$ for the thresholding penalties discussed in Section 2 if λ_n tends to 0.

A New Unified Algorithm

Use quadratic approximation and update by Newton-Raphson.

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2} \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2),$$

for $\beta_j \approx \beta_{j0}$.

$$\begin{aligned} \ell(\beta_0) + \nabla \ell(\beta_0)^T (\beta - \beta_0) + \frac{1}{2} (\beta - \beta_0)^T \nabla^2 \ell(\beta_0) (\beta - \beta_0) \\ + \frac{1}{2} n \beta^T \Sigma_\lambda(\beta_0) \beta, \quad (3.8) \end{aligned}$$

For linear regression:

$$\beta_1 = \{\mathbf{X}^T \mathbf{X} + n \Sigma_\lambda(\beta_0)\}^{-1} \mathbf{X}^T \mathbf{y}.$$

For robust regression:

$$\beta_1 = \{\mathbf{X}^T \mathbf{W} \mathbf{X} + \frac{1}{2} n \Sigma_\lambda(\beta_0)\}^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y},$$

where $\mathbf{W} = \text{diag}\{\psi(|y_1 - \mathbf{x}_1^T \beta_0|)/(y_1 - \mathbf{x}_1^T \beta_0)^2, \dots, \psi(|y_n - \mathbf{x}_n^T \beta_0|)/(y_n - \mathbf{x}_n^T \beta_0)^2\}$.

Standard Error Formula

Use sandwich formula.

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}_1) = \left\{ \nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) + n \Sigma_{\lambda}(\hat{\boldsymbol{\beta}}_1) \right\}^{-1} \widehat{\text{cov}}\{ \nabla \ell(\hat{\boldsymbol{\beta}}_1) \} \\ \times \left\{ \nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) + n \Sigma_{\lambda}(\hat{\boldsymbol{\beta}}_1) \right\}^{-1}. \quad (3.10)$$

Simulation 1: Linear Regression

$Y = \mathbf{x}^T \boldsymbol{\beta} + \sigma \epsilon$, where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and the components of \mathbf{x} and ϵ are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = .5$.

when sample size is small and noise level is high, LASSO is best.

when noise level reduce/ sample size increase, SCAD pops out.

Simulation 1: Linear Regression

Table 1. Simulation Results for the Linear Regression Model

		Avg. No. of 0 Coefficients	
Method	MRME (%)	Correct	Incorrect
$n = 40, \sigma = 3$			
SCAD ¹	72.90	4.20	.21
SCAD ²	69.03	4.31	.27
LASSO	63.19	3.53	.07
Hard	73.82	4.09	.19
Ridge	83.28	0	0
Best subset	68.26	4.50	.35
Garrote	76.90	2.80	.09
Oracle	33.31	5	0
$n = 40, \sigma = 1$			
SCAD ¹	54.81	4.29	0
SCAD ²	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n = 60, \sigma = 1$			
SCAD ¹	47.54	4.37	0
SCAD ²	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0
Oracle	29.82	5	0

NOTE: The value of a in SCAD¹ is obtained by generalized cross-validation, whereas the value of a in SCAD² is 3.7.

Simulation 1: Linear Regression

Table 2. Standard Deviations of Estimators for the Linear Regression Model ($n = 60$)

Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
SCAD ¹	.166	.161 (.021)	.170	.160 (.024)	.148	.145 (.022)
SCAD ²	.161	.161 (.021)	.164	.161 (.024)	.151	.143 (.023)
LASSO	.164	.154 (.019)	.173	.150 (.022)	.153	.142 (.021)
Hard	.169	.161 (.022)	.174	.162 (.025)	.178	.148 (.021)
Best subset	.163	.155 (.020)	.152	.154 (.026)	.152	.139 (.020)
Oracle	.155	.154 (.020)	.147	.153 (.024)	.146	.137 (.019)

Simulation 2: Robust Regression

Table 3. Simulation Results for the Robust Linear Model

Method	MRME (%)	Avg. No. of 0 Coefficients	
		Correct	Incorrect
SCAD ($a = 3.7$)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	.18
Oracle	23.33	5	0

Table 4. Standard Deviations of Estimators for the Robust Regression Model

Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
SCAD	.167	.171 (.018)	.185	.176 (.022)	.165	.155 (.020)
LASSO	.158	.165 (.022)	.159	.167 (.020)	.182	.154 (.019)
Hard	.179	.168 (.018)	.176	.176 (.025)	.157	.154 (.020)
Best subset	.198	.172 (.023)	.185	.175 (.024)	.199	.152 (.023)
Oracle	.163	.199 (.040)	.156	.202 (.043)	.166	.177 (.037)

Simulation 3: Logistic Regression

Table 5. Simulation Results for the Logistic Regression

Method	MRME (%)	Avg. No. of 0 Coefficients	
		Correct	Incorrect
SCAD ($a = 3.7$)	26.48	4.98	.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	.01
Oracle	25.71	5	0

Table 6. Standard Deviations of Estimators for the Logistic Regression

Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
SCAD ($a = 3.7$)	.571	.538 (.107)	.383	.372 (.061)	.432	.398 (.065)
LASSO	.310	.379 (.037)	.285	.284 (.019)	.244	.287 (.019)
Hard	.675	.561 (.126)	.428	.400 (.062)	.467	.421 (.079)
Best subset	.624	.547 (.121)	.398	.383 (.067)	.468	.412 (.077)
Oracle	.553	.538 (.103)	.374	.373 (.060)	.432	.398 (.064)

Burns Data

Table 7. Estimated Coefficients and Standard Errors for Example 4.4

Method	MLE	Best Subset (AIC)	Best Subset (BIC)	SCAD	LASSO	Hard
Intercept	5.51 (.75)	4.81 (.45)	6.12 (.57)	6.09 (.29)	3.70 (.25)	5.88 (.41)
X_1	-8.83 (2.97)	-6.49 (1.75)	-12.15 (1.81)	-12.24 (.08)	0 (—)	-11.32 (1.1)
X_2	2.30 (2.00)	0 (—)	0 (—)	0 (—)	0 (—)	2.21 (1.41)
X_3	-2.77 (3.43)	0 (—)	-6.93 (.79)	-7.00 (.21)	0 (—)	-4.23 (.64)
X_4	-1.74 (1.41)	.30 (.11)	-.29 (.11)	0 (—)	-.28 (.09)	-1.16 (1.04)
X_1^2	-.75 (.61)	-1.04 (.54)	0 (—)	0 (—)	-1.71 (.24)	0 (—)
X_3^2	-2.70 (2.45)	-4.55 (.55)	0 (—)	0 (—)	-2.67 (.22)	-1.92 (.95)
X_1X_2	.03 (.34)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_1X_3	7.46 (2.34)	5.69 (1.29)	9.83 (1.63)	9.84 (.14)	.36 (.22)	9.06 (.96)
X_1X_4	.24 (.32)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_2X_3	-2.15 (1.61)	0 (—)	0 (—)	0 (—)	-0.10 (.10)	-2.13 (1.27)
X_2X_4	-.12 (.16)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_3X_4	1.23 (1.21)	0 (—)	0 (—)	0 (—)	0 (—)	.82 (1.01)