Simulation and Penalized Least Squares

San-Teng Huang

National Dong Hwa University

2018/06/04

Outline

Model

- Consider the model:

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i \qquad , i = 1, 2, ..., n \tag{1}$$

where $f(x; \boldsymbol{\theta})$ is a known non-linear function, and the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$, and ε_i *iid* with mean 0, variance σ^2, for $i = 1, ..., n$.

- Assume only some parameters are non-zero. the true $\boldsymbol{\theta_0} = (\theta_{01}, \theta_{02}, ..., \theta_{0p}) = (\boldsymbol{\theta_{01}^T}, \boldsymbol{\theta_{02}^T}) = (\boldsymbol{\theta_{01}^T}, \mathbf{0})$

Goal

- Coefficient estimation
- Variable selection

  the true $\boldsymbol{\theta_0} = (\theta_{01}, \theta_{02}, ..., \theta_{0p}) = (\boldsymbol{\theta_{01}^T}, \boldsymbol{\theta_{02}^T}) = (\boldsymbol{\theta_{01}^T}, \mathbf{0})$

Penalized Least Squares

- To minimize

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f(x_i;\boldsymbol{\theta}))^2 + n\sum_{j=1}^{p} p_\lambda(|\theta_j|) \tag{2}$$

where $p_\lambda(|\theta|)$ is a penalty function.

- The penalized least squares estimator is

$$\hat{\boldsymbol{\theta}}_n = arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta})$$

Penalty function

- LASSO(L_1-penalty):

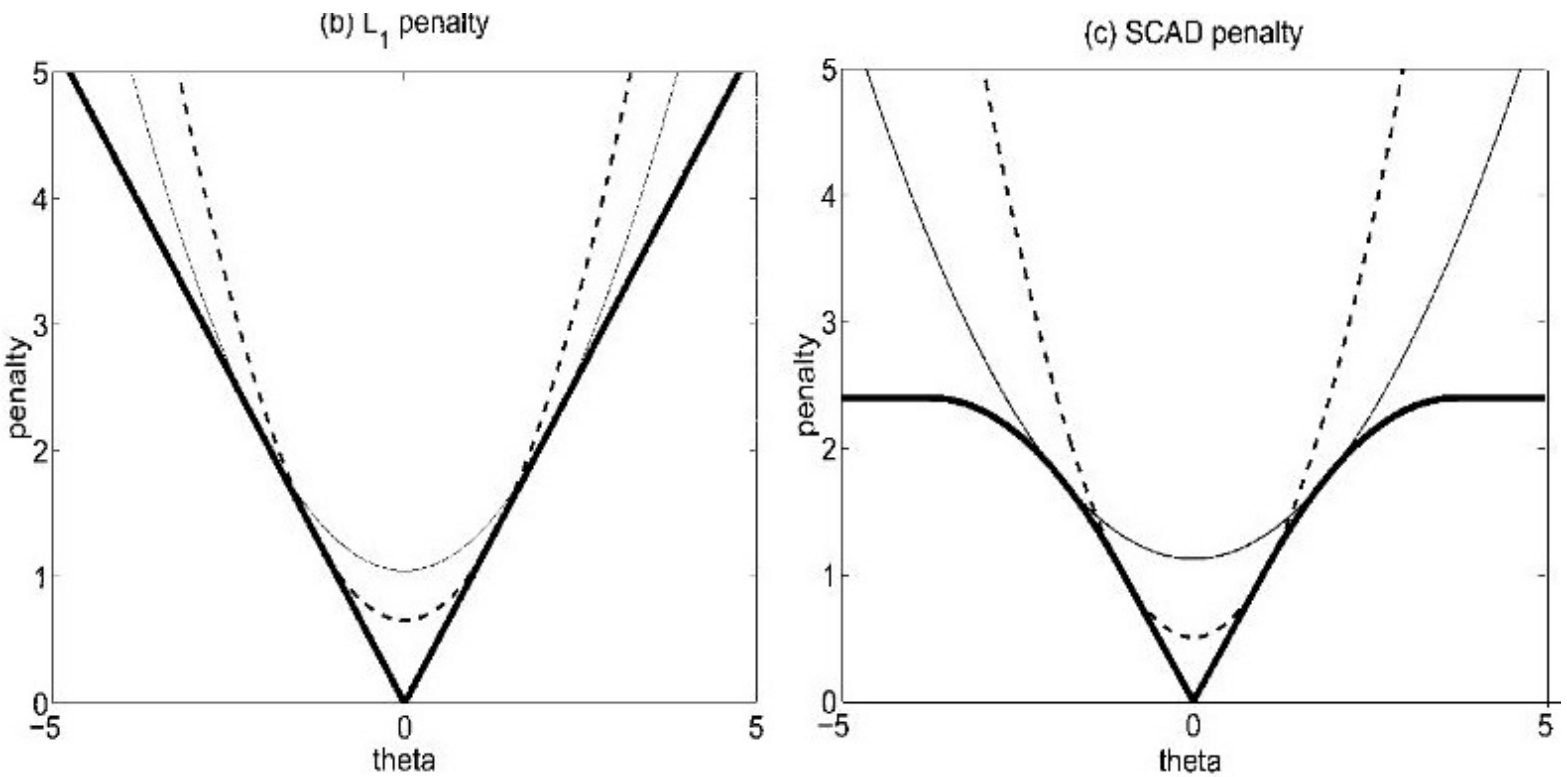
$$p_\lambda(|\theta|) = \lambda|\theta|$$

- Smoothly Clipped Absolute Deviation (SCAD):

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & , |\theta| \leq \lambda \\ -\left(\dfrac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}\right) & , \lambda < |\theta| \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & , |\theta| > a\lambda \end{cases}$$

where $a = 3.7$.

Plot of $p_\lambda(|\theta|)$:



	Unbiasedness	Sparsity	Continuity
Hard-thresholding	yes	yes	no
L_1-penalty(LASSO)	no	yes	yes
L_2-penalty(ridge regression)	no	no	yes
SCAD	yes	yes	yes

Case 1 (p=30,n=100)

Model: (4 additive + 22 linear)

$$y_i = \sum_{j=1}^{4} \frac{\alpha_j}{1 + exp(-\beta_j x_{ij})} + \beta_5 x_{i5} + ... + \beta_{26} x_{i26} + \varepsilon_i$$

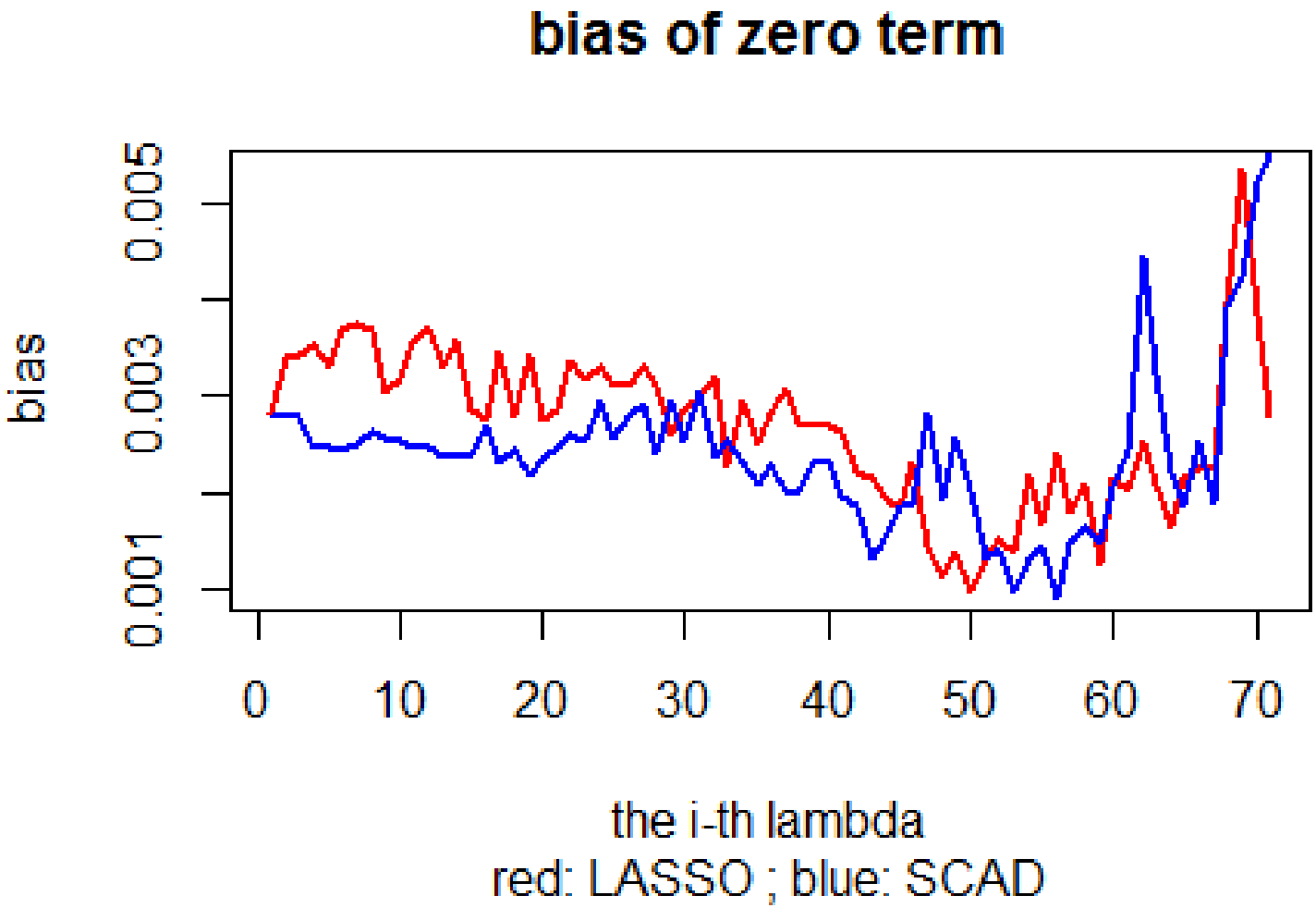
where $\varepsilon_i \sim N(0, 0.25)$, $x_i \sim U(-5, 5)$, $i = 1, ..., 100$
$\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\boldsymbol{T}}, \boldsymbol{\beta}^{\boldsymbol{T}}) = (5, 2, 8, 3, 1.5, 2.4, 3.3, 4.2, 0, ..., 0)$
With 100 data sets, initial value $(1, 1, ..., 1)$,
$\lambda = (0.001, 0.011023, 0.3303, 0.992275)$

	SCAD		LASSO	
$\lambda = 0.001$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 5$	-0.017(2.05)	1	0.056(1.32)	1
$\alpha_2 = 2$	-0.612(1.63)	1	0.317(2.23)	1
$\alpha_3 = 8$	0.019(0.11)	1	0.03(0.09)	1
$\alpha_4 = 3$	0.606(2.11)	1	-0.382(1.91)	1
$\beta_1 = 1.5$	-0.162(0.58)	1	-0.073(0.37)	1
$\beta_2 = 2.4$	-0.931(2.11)	1	-0.641(1.52)	1
$\beta_3 = 3.3$	-0.011(0.26)	1	-0.024(0.22)	1
$\beta_4 = 4.2$	-0.570(1.63)	1	-1.198(2.37)	1
$\max\{\beta_j = 0\}$	0.002796(0.019)	0.97	0.003390(0.015)	0.96

	SCAD		LASSO	
$\lambda = 0.011023$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 5$	-0.056(2.01)	1	-0.033(1.14)	1
$\alpha_2 = 2$	-0.450(1.48)	1	0.461(2.25)	1
$\alpha_3 = 8$	0.015(0.11)	1	-0.006(0.10)	1
$\alpha_4 = 3$	0.481(2.01)	1	-0.431(1.99)	1
$\beta_1 = 1.5$	-0.151(0.57)	1	-0.085(0.35)	1
$\beta_2 = 2.4$	-0.767(1.94)	1	-1.063(1.05)	1
$\beta_3 = 3.3$	-0.010(0.25)	1	-0.184(0.20)	1
$\beta_4 = 4.2$	-0.462(1.51)	1	-2.053(1.78)	1
$\max\{\beta_j = 0\}$	0.002804(0.018)	0.93	0.003123(0.014)	0.94

	SCAD		LASSO	
$\lambda = 0.3303$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 5$	-0.005(2.97)	1	-0.173(0.86)	1
$\alpha_2 = 2$	-1.622(4.49)	0.86	-0.226(0.91)	0.9
$\alpha_3 = 8$	-0.062(0.17)	1	0.021(0.23)	1
$\alpha_4 = 3$	1.751(4.47)	1	-0.050(0.76)	0.99
$\beta_1 = 1.5$	-0.109(0.76)	1	-0.496(0.26)	1
$\beta_2 = 2.4$	-1.689(0.95)	0.85	-1.878(0.20)	0.9
$\beta_3 = 3.3$	0.143(0.66)	1	-1.477(0.17)	1
$\beta_4 = 4.2$	-1.508(2.36)	1	-3.361(0.31)	0.99
$\max\{\beta_j = 0\}$	0.002048(0.014)	0.43	0.002139(0.023)	0.56

	SCAD		LASSO	
$\lambda = 0.992275$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 5$	0.978(4.14)	0.81	0.159(0.85)	0.99
$\alpha_2 = 2$	-0.984(3.51)	0.23	-1.683(0.82)	0.22
$\alpha_3 = 8$	0.836(1.31)	1	0.327(0.40)	1
$\alpha_4 = 3$	-0.583(3.66)	0.52	-0.561(1.27)	0.85
$\beta_1 = 1.5$	-0.369(3.28)	0.81	-0.795(0.12)	0.99
$\beta_2 = 2.4$	-2.254(1.28)	0.26	-2.354(0.13)	0.15
$\beta_3 = 3.3$	-0.414(3.47)	1	-2.061(0.14)	1
$\beta_4 = 4.2$	-3.683(1.81)	0.54	-3.774(0.22)	0.8
$\max\{\beta_j = 0\}$	0.005523(0.038)	0.5	0.002796(0.034)	0.41


bias of zero term
bias
0.001
0.003
0.005
0
10
20
30
40
50
60
70
the i-th lambda
red: LASSO ; blue: SCAD

Case 2 (p=60,n=100)

Model: (30 additive)

$$y_i = \sum_{j=1}^{30} \beta_j (1 + exp(-\alpha_j x_{ij}) - 0.5) + \varepsilon_i \quad , i = 1, ..., 100$$

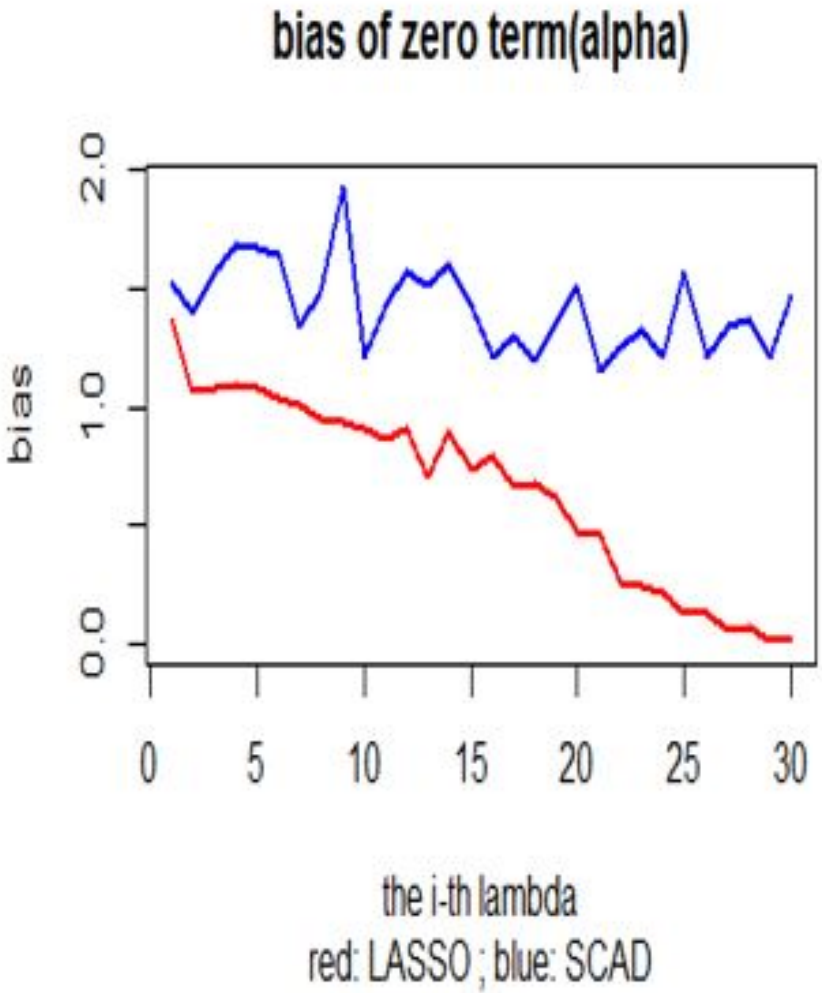
where $\varepsilon_i \sim N(0, 0.25)$, $x_i \sim U(-1, 1)$
$\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\boldsymbol{T}}, \boldsymbol{\beta}^{\boldsymbol{T}}) = (1.5, 0.4, 1.2, 2.1, 0, ..., 0, 3, 4, 1, 2, 0, ..., 0)$
With 20 data sets, initial value $(1, 1, ..., 1)$,
$\lambda = (0.001, 0.011023, 0.12151, 0.270426)$

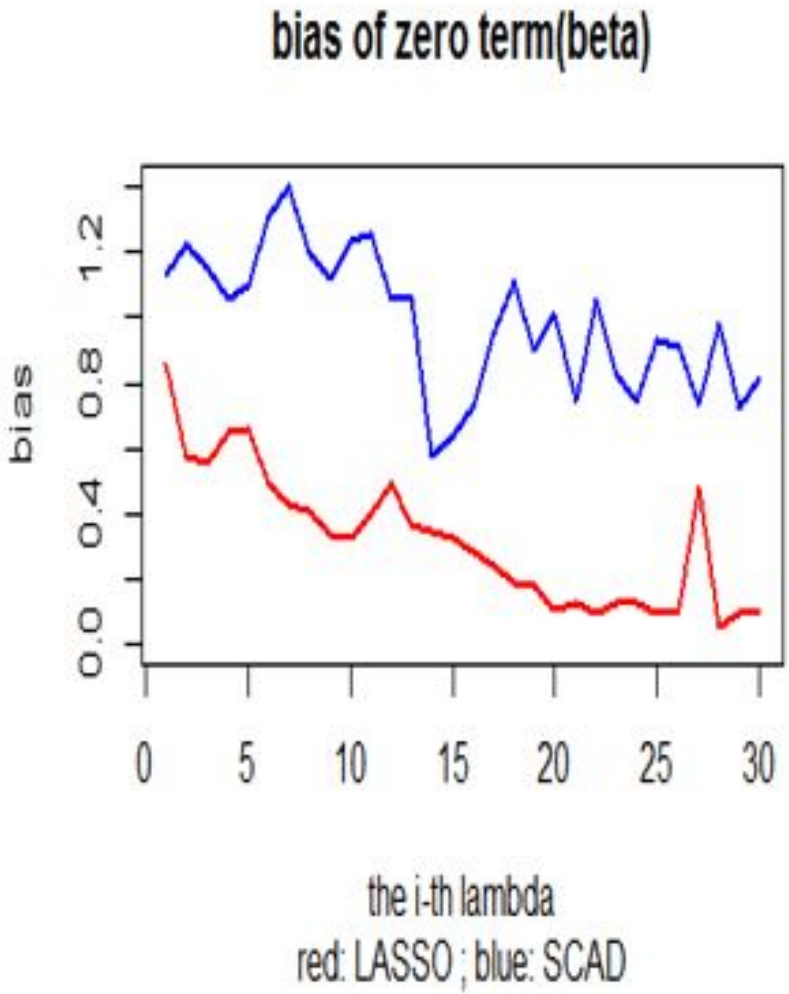
	SCAD		LASSO	
$\lambda = 0.001$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.006(0.04)	1	-0.005(0.04)	1
$\alpha_2 = 0.4$	0.023(0.12)	1	0.106 (0.14)	1
$\alpha_3 = 1.2$	-0.87(0.27)	1	-0.052(0.16)	1
$\alpha_4 = 2.1$	-0.009(0.04)	1	-0.015(0.03)	1
$\max\{\alpha_j = 0\}$	1.3897(2.66)	1	1.060(1.53)	1
$\beta_1 = 3$	-0.016(0.13)	1	0.025(0.13)	1
$\beta_2 = 4$	0.131(1.19)	1	-0.652(0.92)	1
$\beta_3 = 1$	0.246(0.66)	1	0.086(0.19)	1
$\beta_4 = 2$	0.020(0.07)	1	0.021(0.05)	1
$\max\{\beta_j = 0\}$	1.2196(3.12)	1	0.5730(2.97)	1

	SCAD		LASSO	
$\lambda = 0.011023$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.010(0.05)	1	-0.006(0.04)	1
$\alpha_2 = 0.4$	0.031(0.12)	1	0.230 (0.16)	1
$\alpha_3 = 1.2$	-0.154(0.33)	1	-0.104(0.26)	1
$\alpha_4 = 2.1$	-0.012(0.04)	1	-0.013(0.05)	1
$\max\{\alpha_j = 0\}$	1.6009(2.41)	1	0.8977(1.67)	1
$\beta_1 = 3$	-0.028(0.14)	1	0.027(0.13)	1
$\beta_2 = 4$	0.025 (1.11)	1	-1.430(0.59)	1
$\beta_3 = 1$	0.487(1.09)	1	0.197(0.38)	1
$\beta_4 = 2$	0.026(0.07)	1	0.017(0.09)	1
$\max\{\beta_j = 0\}$	0.5758(3.36)	1	0.3503(1.45)	1

	SCAD		LASSO	
$\lambda = 0.12151$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.0002(0.07)	1	-0.059(0.17)	1
$\alpha_2 = 0.4$	0.027(0.20)	1	0.276 (0.07)	1
$\alpha_3 = 1.2$	-0.368(0.45)	1	-0.422(0.09)	1
$\alpha_4 = 2.1$	-0.007(0.04)	1	-0.049(0.05)	1
$\max\{\alpha_j = 0\}$	1.2016(2.69)	0.8	0.1330(0.65)	0.95
$\beta_1 = 3$	0.008(0.24)	1	0.254(0.73)	1
$\beta_2 = 4$	0.480 (1.95)	1	-1.810(0.28)	1
$\beta_3 = 1$	1.455(2.07)	1	0.622(0.19)	1
$\beta_4 = 2$	0.015(0.09)	1	0.087(0.11)	1
$\max\{\beta_j = 0\}$	0.9150(3.91)	0.75	0.1038(0.38)	1

	SCAD		LASSO	
$\lambda = 0.270426$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	-0.131(0.19)	1	-0.110(0.10)	1
$\alpha_2 = 0.4$	0.192(0.69)	1	0.197 (0.15)	0.95
$\alpha_3 = 1.2$	-0.206(1.49)	1	-0.559(0.29)	0.95
$\alpha_4 = 2.1$	-0.160(0.19)	1	-0.122(0.08)	1
$\max\{\alpha_j = 0\}$	1.4611(2.60)	0.95	0.0366(0.18)	0.85
$\beta_1 = 3$	0.526(0.74)	1	0.390(0.38)	1
$\beta_2 = 4$	0.904(2.98)	1	-1.877(0.64)	1
$\beta_3 = 1$	2.352(2.09)	1	0.601(0.71)	1
$\beta_4 = 2$	0.345(0.51)	1	0.242(0.17)	1
$\max\{\beta_j = 0\}$	0.8183(4.46)	0.85	0.1036(0.35)	1


bias of zero term(alpha)
bias
0.0
1.0
2.0
0
5
10
15
20
25
30
the i-th lambda
red: LASSO ; blue: SCAD


bias of zero term(beta)
bias
0.0
0.4
0.8
1.2
0
5
10
15
20
25
30
the i-th lambda
red: LASSO ; blue: SCAD

Case 3 (p=100,n=100)

Model: (50 additive)

$$y_i = \sum_{j=1}^{50} \beta_j (1 + exp(-\alpha_j x_{ij}) - 0.5) + \varepsilon_i \quad , i = 1, ..., 100$$

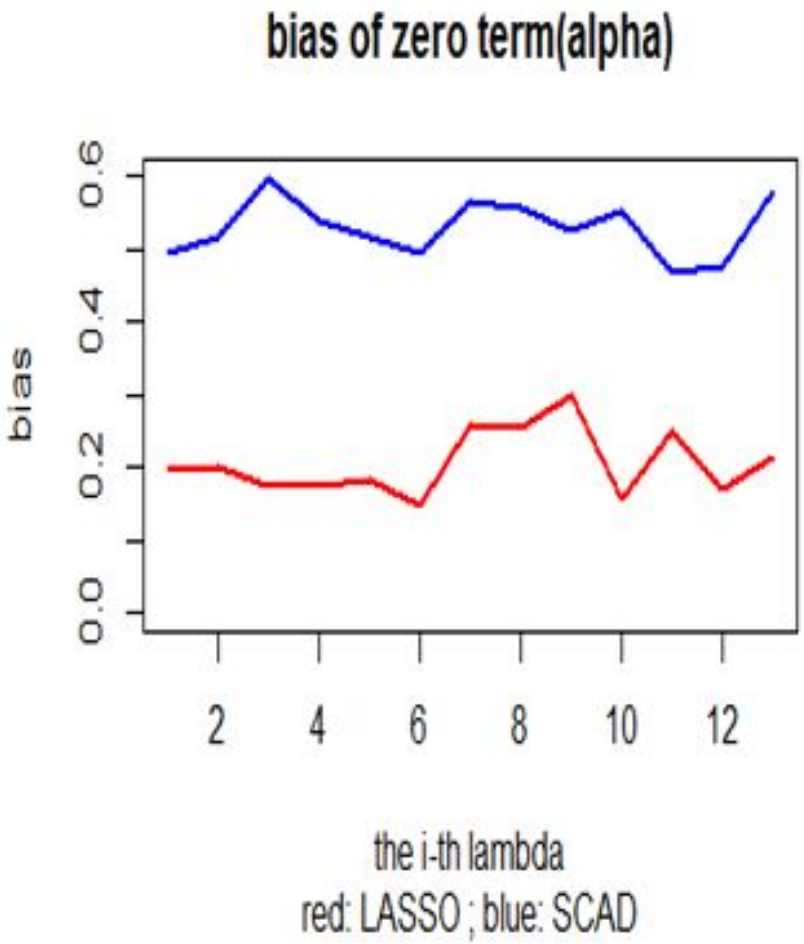
where $\varepsilon_i \sim N(0, 0.25)$, $x_i \sim U(-5, 5)$
$\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\boldsymbol{T}}, \boldsymbol{\beta}^{\boldsymbol{T}}) = (1.5, 0.4, 1.2, 2.1, 0, ..., 0, 3, 4, 1, 2, 0, ..., 0)$
With 50 data sets, initial value $(0, 0, ..., 0)$,
$\lambda = (0.2, 0.4, 0.5, 0.6)$

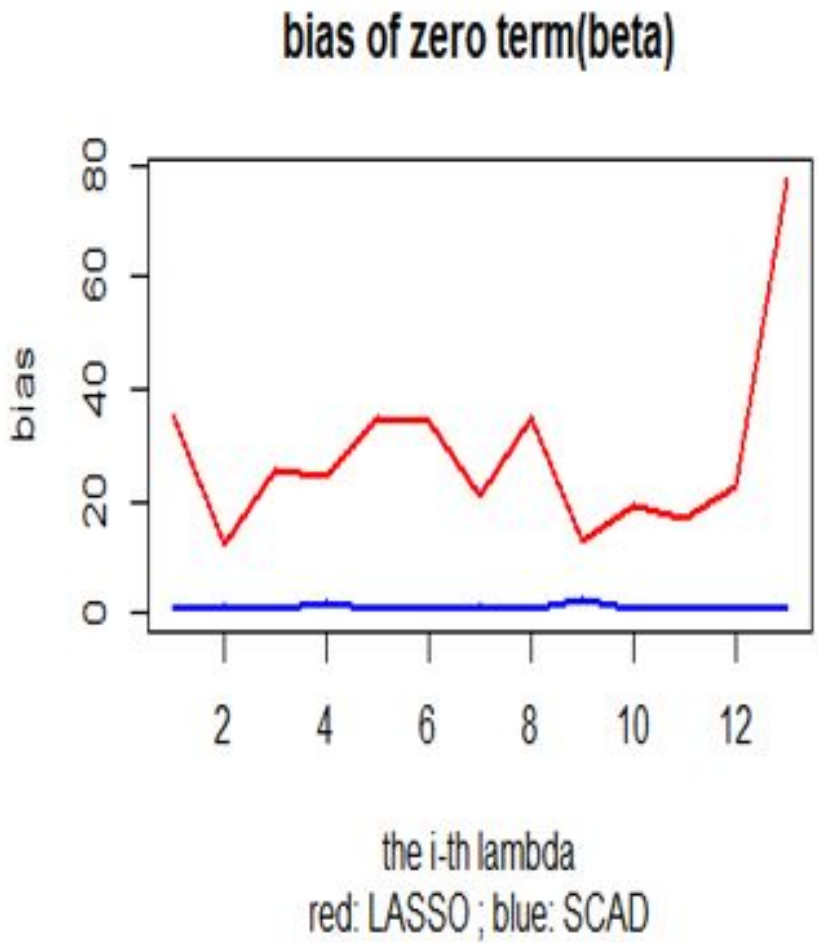
	SCAD		LASSO	
$\lambda = 0.2$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.150(0.12)	1	-1.150(1.06)	1
$\alpha_2 = 0.4$	-0.005(0.87)	1	-0.195 (0.60)	1
$\alpha_3 = 1.2$	0.002(0.63)	1	-1.118(0.64)	1
$\alpha_4 = 2.1$	0.041(0.05)	1	-0.499(0.16)	1
$\max\{\alpha_j = 0\}$	0.4938(1.08)	1	0.1959(0.82)	1
$\beta_1 = 3$	-1.365(0.76)	1	9.273(49.44)	1
$\beta_2 = 4$	-3.860(1.89)	0.98	3.081(64.85)	1
$\beta_3 = 1$	-0.451(1.15)	1	9.711(72.10)	1
$\beta_4 = 2$	-0.322(0.33)	1	25.662(32.57)	1
$\max\{\beta_j = 0\}$	0.5989(2.35)	1	35.444(118.7)	1

	SCAD		LASSO	
$\lambda = 0.4$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.119(0.11)	1	-0.947(0.94)	1
$\alpha_2 = 0.4$	-0.061(0.89)	1	-0.268 (0.63)	1
$\alpha_3 = 1.2$	-0.098(0.60)	1	-1.176(0.65)	1
$\alpha_4 = 2.1$	0.040(0.06)	1	-0.499(0.16)	1
$\max\{\alpha_j = 0\}$	0.5638(1.08)	1	0.2564(0.75)	1
$\beta_1 = 3$	-1.054(1.27)	1	17.073(68.99)	1
$\beta_2 = 4$	-3.591(1.77)	1	4.020(38.89)	1
$\beta_3 = 1$	-0.446(1.68)	1	4.035(52.77)	1
$\beta_4 = 2$	-0.305(0.36)	1	23.915(19.50)	1
$\max\{\beta_j = 0\}$	1.4542(9.04)	1	21.0934(126.6)	1

	SCAD		LASSO	
$\lambda = 0.5$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.134(0.12)	1	-0.782(1.02)	1
$\alpha_2 = 0.4$	0.012(0.77)	1	-0.321 (0.69)	1
$\alpha_3 = 1.2$	-0.031(0.61)	1	-1.331(0.70)	1
$\alpha_4 = 2.1$	0.042(0.05)	1	-0.483(0.15)	1
$\max\{\alpha_j = 0\}$	0.5513(1.05)	1	0.1570(0.81)	1
$\beta_1 = 3$	-1.160(1.19)	1	6.231(44.54)	1
$\beta_2 = 4$	-3.689(1.97)	1	8.418(42.37)	1
$\beta_3 = 1$	-0.566(1.34)	1	8.934(63.44)	1
$\beta_4 = 2$	-0.322(0.34)	1	22.444(19.11)	1
$\max\{\beta_j = 0\}$	0.8678(3.63)	1	18.97(95.15)	1

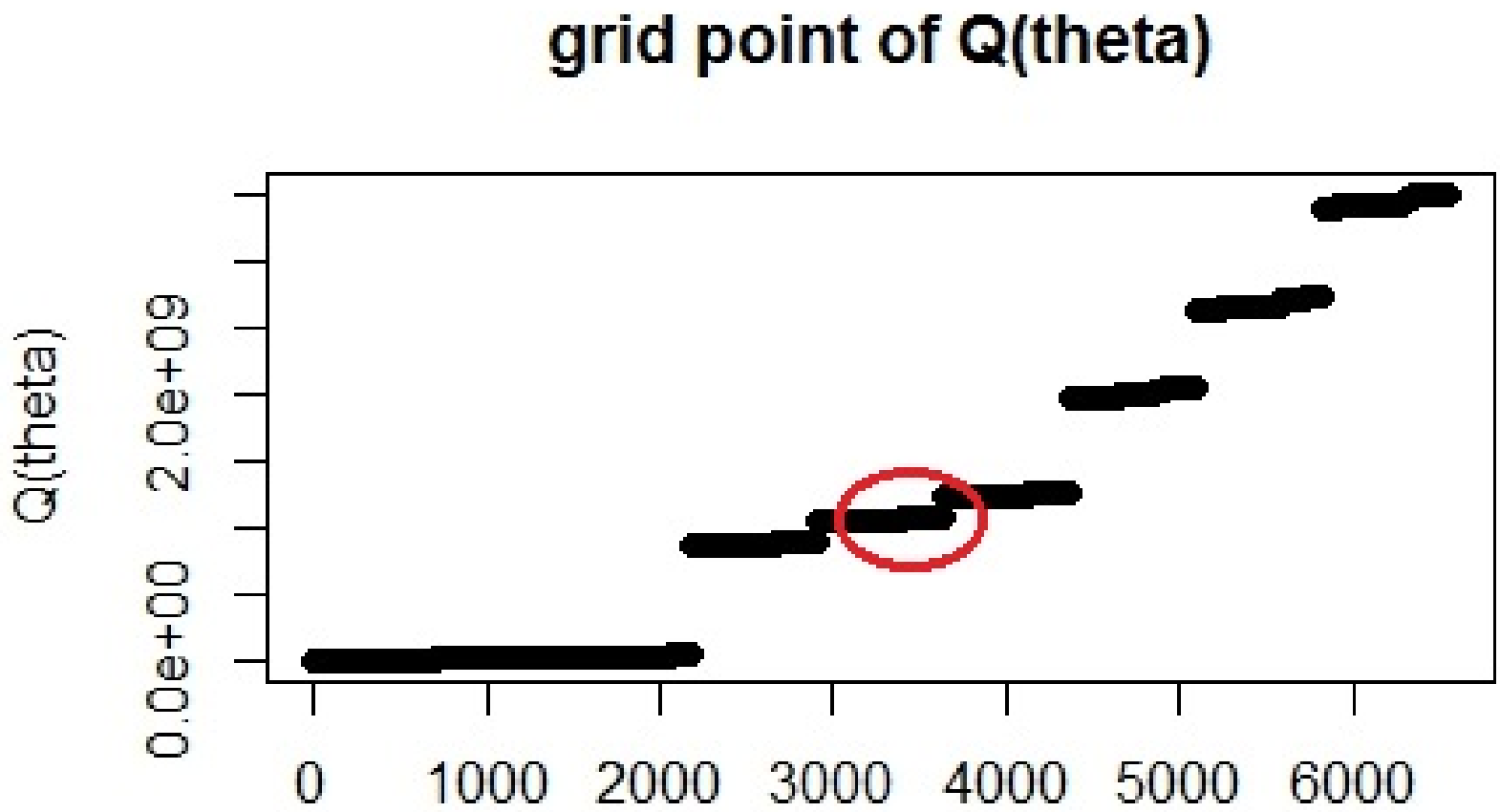
	SCAD		LASSO	
$\lambda = 0.6$	bias(sd)	count	bias(sd)	count
$\alpha_1 = 1.5$	0.146(0.13)	1	-0.939(1.05)	1
$\alpha_2 = 0.4$	-0.078(0.94)	1	-0.387 (0.67)	1
$\alpha_3 = 1.2$	-0.100(0.80)	1	-1.137(0.70)	0.96
$\alpha_4 = 2.1$	0.043(0.06)	1	-0.483(0.17)	1
$\max\{\alpha_j = 0\}$	0.5771(1.134)	1	0.2119(0.78)	1
$\beta_1 = 3$	-1.275(0.97)	1	-0.935(76.48)	1
$\beta_2 = 4$	-4.157(1.82)	1	0.131(31.63)	1
$\beta_3 = 1$	-0.555(0.94)	1	16.812(80.38)	1
$\beta_4 = 2$	-0.322(0.39)	1	23.127(18.96)	1
$\max\{\beta_j = 0\}$	0.8098(2.95)	1	77.7354(520.8)	1


bias of zero term(alpha)
bias
0.0
0.2
0.4
0.6
2
4
6
8
10
12
the i-th lambda
red: LASSO ; blue: SCAD


bias of zero term(beta)
bias
0
20
40
60
80
2
4
6
8
10
12
the i-th lambda
red: LASSO ; blue: SCAD

Problems

- How to choose λ ?
  Exhaustive method $\rightarrow$ BIC criterion.
- Optimization problem in high dimension:
  quasi-Newton method approach a local minimum.


grid point of Q(theta)
Q(theta)
0.0e+00
2.0e+09
0
1000
2000
3000
4000
5000
6000

Reference :

[1]Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association 96.456 (2001): 1348-1360.
[2]TIBSHIRANI, R.(1996)Regression Shrinkage and Selection via the Lasso. J.R.Stat.Soc.B.58, No.1, 267-288.