Exercise 5: Sparsity **Cenying(Tracy) Yang**
SDS 385 Due Date: 10/12/2016
Prof. James Scott

Penalized likelihood and soft thresholding

(A) **Solution**
Suppose an IID sequence Y_n is normally distributed with mean θ and variance 1. The probability density function of a generic term of the sequence is

$$f_Y(y_j) = (2\pi)^{-1/2} \exp(-\frac{1}{2}(y_j - \theta)^2) \tag{1}$$

Its likelihood function is

$$\begin{aligned} L(y;\theta) &= \prod_{j=1}^{n} f_Y(y_j) \\ &= (2\pi)^{-n/2} \exp(-\frac{1}{2}\sum_{j=1}^{n}(y_j - \theta)^2)) \end{aligned} \tag{2}$$

Its negative log likelihood function is

$$L(y;\theta) = \frac{n}{2}\ln(x\pi) + \frac{1}{2}\sum_{j=1}^{n}(y_j - \theta) \tag{3}$$

Taking derivative with respect to θ, we get $\frac{1}{2}(y-\theta)$.
Now we derive proof for soft thresholding.

$$S_\lambda(y) = \arg\min_\theta \frac{1}{2}(y-\theta)^2 + \lambda|\theta| = \arg\min_\theta \begin{cases} \frac{1}{2}(y-\theta)^2 + \lambda\theta, & \text{if } \theta > 0 \\ \frac{1}{2}y^2, & \text{if } \theta = 0 \\ \frac{1}{2}(y-\theta)^2 - \lambda\theta, & \text{if } \theta < 0 \end{cases} \tag{4}$$

Taking derivative when $\theta \neq 0$ and set it to zero, we get θ as a function of λ and y,

$$S_\lambda(y) = \begin{cases} y - \lambda, & \text{where } y > \lambda \\ y + \lambda, & \text{where } y < -\lambda \end{cases} = \text{sgn}(y)(|y| - \lambda)_+ \tag{5}$$

(B) **Solution**
I present plots of varying sparsity and varying λ (see Figure1 - Figure4). The takeaway is that given sparsity, bigger $lambda$ forces more estimated θ to 0, and as sparsity grows, the soft thresholding method works better in terms of smaller MSE.

Lasso

(A) **Solution**
We plot the solution path of $\hat{\beta_\lambda}$ as a function of λ (actually $log(\lambda)$ in glm package in R)(see Figure 5). The values across the top of the plot represent degrees of freedom,i.e.,non-zero coefficients, at each λ value.
(B) **Solution**

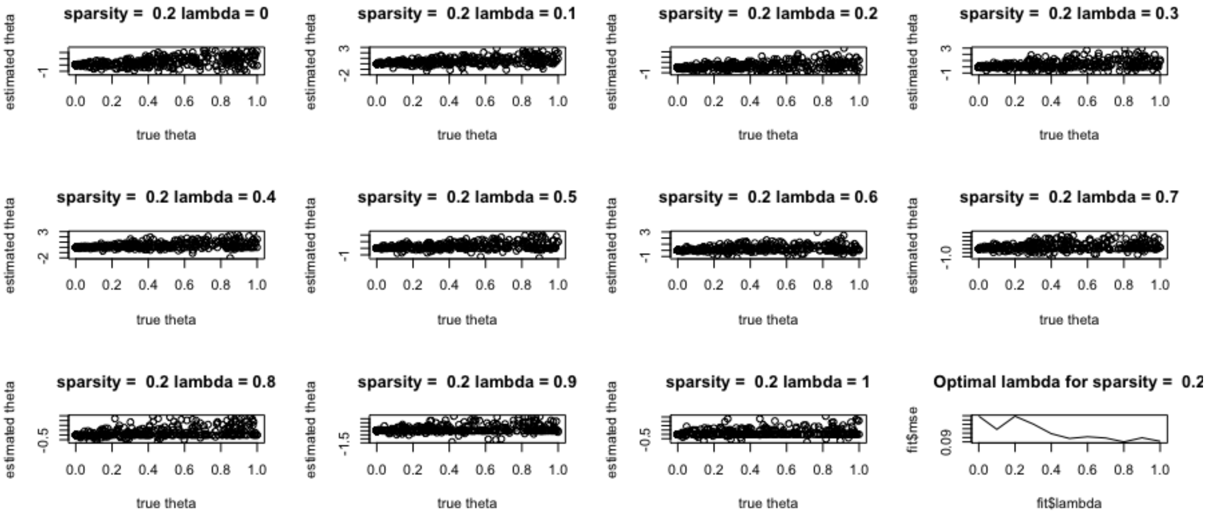

Figure 1: θ v.s. $\hat{\theta}$ at sparsity = 0.2

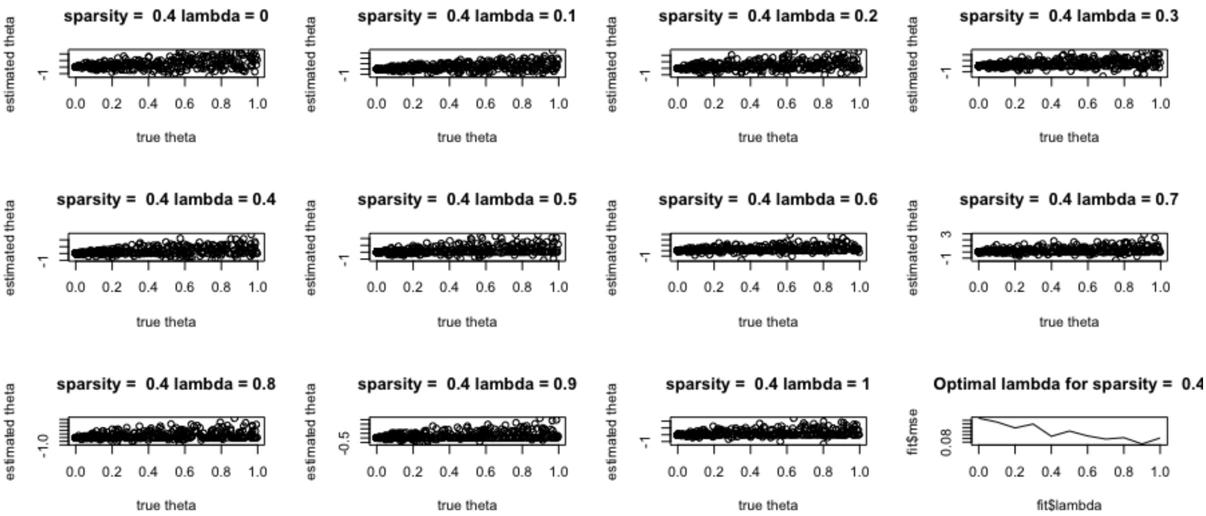

Figure 2: θ v.s. $\hat{\theta}$ at sparsity = 0.4

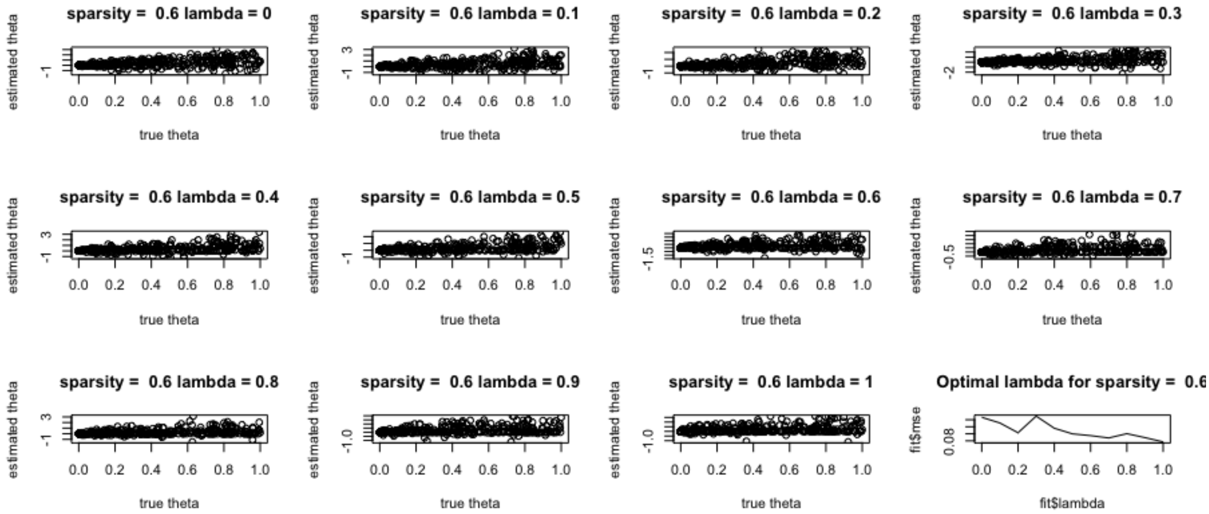

Figure 3: θ v.s. $\hat{\theta}$ at sparsity = 0.6

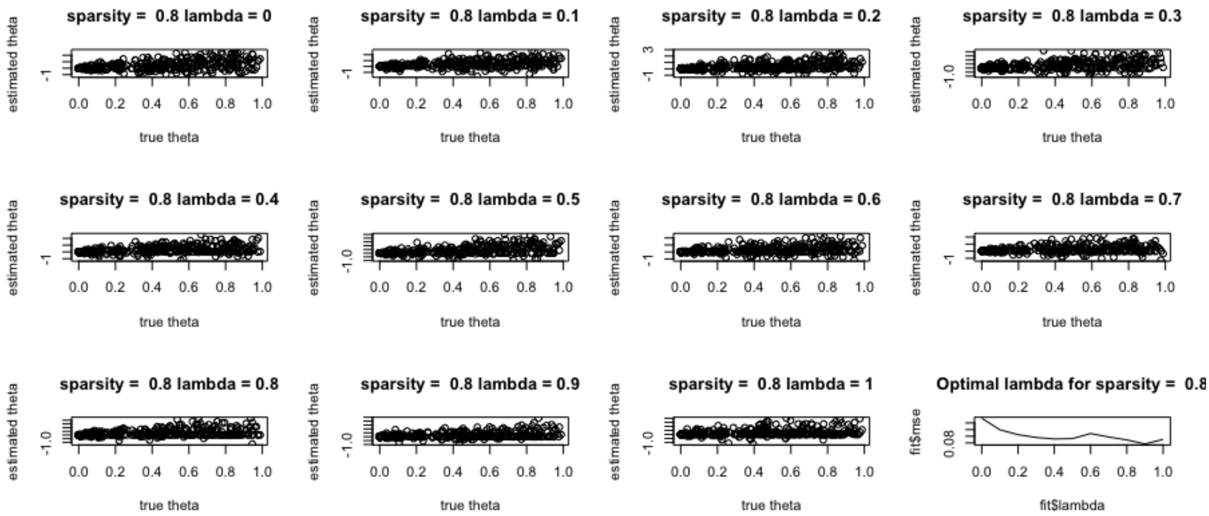

Figure 4: θ v.s. $\hat{\theta}$ at sparsity = 0.8

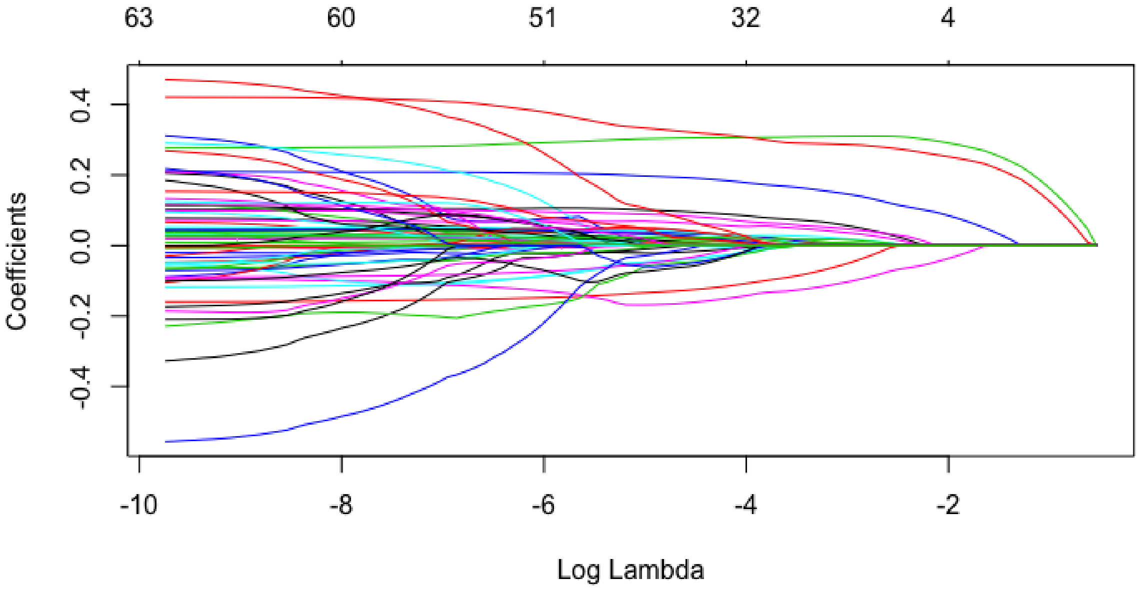

Figure 5: solution path of $\hat{\beta}_\lambda$ as a function of λ

I split data into 10 folds, and train the model on 9 folds of data and leave 1 fold of data for validation each time. And finally take mean error of 10 iterations as the out-of-sample mean square error. The best λ is 0.03944774. While the best λ is 0.03594331 if using glm package to do 10-fold cross validation.

(C) **Solution**

Based on C_p statistics, the best λ is 0.0327502.See Figure 6 for comparison different methods in terms of error.

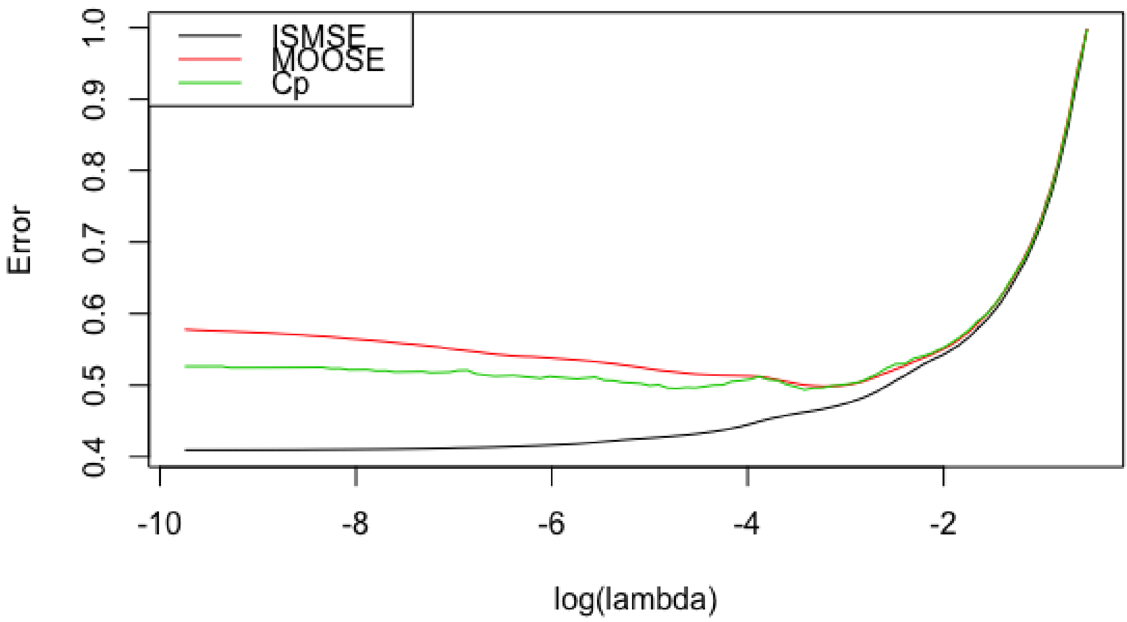

Figure 6: comparison of methods in terms of error

Class Note

Lasso is useful when the number of sample, N, is much smaller than the number of features/characteristics, d.
True model: $y = X\beta_0 + \epsilon$, where $\epsilon N(0, \sigma^2)$
We estimate β by $\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|$
We would like to recover β in $support(\beta_0) = i : (\beta_0)_i \neq 0$, i.e., choosing features/characteristics that matter.
TEO: assume that $\|y\| \leq B$, $\|X\|_{\inf} \leq B$, then

$$\beta^* = \arg\min_{\|\beta\| \leq L} E[(y^* - X^*\beta)^2] \tag{6}$$

where y^* is a scalar and X^* is a vector. Then

$$P[E(y - X\beta)^2 \leq (y^* - X^*\beta)^2 + \sqrt[2]{\frac{L}{N} + \log(\frac{d}{\delta})}] \geq 1 - \delta \tag{7}$$