

Adaptive Linear Regression Selection

Hung Chen

Department of Mathematics
Joint work with Mr. Chiuan-Fa Tang
Hsu Centennial Memorial Conference at Peking University

7/07/2010

- 1 Introduction
 - Objective
 - Nested Linear Regression Models
- 2 Adaptive Penalty
 - Unbiased Risk Estimate
 - Generalized degrees of freedom
- 3 Shen and Ye's proposal
- 4 Proof
- 5 Conclusion

Outline

Introduction

●○○○

Adaptive Penalty

○○○○○○○○

Linear Regression Models

Consider a linear regression model with normal error,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is an $n \times p$ matrix,
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$,
- $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T = \mathbf{X}\boldsymbol{\beta}$,
- $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, and σ^2 is known.

Nested Models

We only consider the nested linear competing model

$$\{M_k, k = 0, \dots, p\}.$$

- Lasso leads to a data-driven nested models.
- For model M_k , $\beta_j \neq 0$ for $j \leq k$ and $\beta_j = 0$ for $j > k$.
- β 's are estimated by the **least square method** and
- μ is estimated by

$$\hat{\mu}_{M_k} = P_{M_k} \mathbf{Y},$$

where P_{M_k} is the projection matrix corresponding to model M_k .

- Its residual sum of squares is defined as

$$RSS(M_k) = (\mathbf{Y} - \hat{\mu}_{M_k})^T (\mathbf{Y} - \hat{\mu}_{M_k}).$$

Model Selection

If AIC (Mallows' C_p) is used to score models, we choose the model \hat{M} by minimizing

$$RSS(M_k) + 2|M_k|\sigma^2$$

with respect to all competing models $\{M_k, k = 0, \dots, p\}$, where $|M_k|$ is the size of M_k .

Note that

- It does not include the random error introduced in model selection procedure.
- What can be done?
 - Refer to the proposal in Shen and Ye (2002).



Shen and Ye's proposal (2002, *JASA*)

Shen and Ye (2002) proposed to choose $\lambda > 0$ to minimize the unbiased risk estimator

$$\hat{\lambda} = \operatorname{argmin}_{\lambda > 0} \operatorname{RSS}(\hat{M}(\lambda)) + g_0(\lambda)\sigma^2 .$$

The resulting selected model is $\hat{M}(\hat{\lambda})$.

As an attempt to understand their proposal, consider the situation

- BIC is consistent (no underfitting).
- nested competing models
- $\lambda \in [0, \log n]$

Is

$$\hat{M}(\hat{\lambda}) = \hat{M}(\log n) = M_{k_0}$$

or $\hat{\lambda} = \log n$?

Assumptions: BIC is consistent

Recall that p_0 is the number of covariates in the true model.
Assume that

Assumption B1. There exists a constant $c > 0$ such that

$$\boldsymbol{\mu}^T (\mathbf{I} - \mathbf{P}_{M_k}) \boldsymbol{\mu} \geq cn \text{ for all } k < p_0, \text{ where}$$

$$\boldsymbol{\mu} = \mathbf{X}_{p_0} (\beta_1, \dots, \beta_{p_0})^T$$

is the mean vector of the true model.

Assumption B2. The sample size n is large enough such that
 $cn > 2p_0 \log n$.

Assumption N. $\log n > 2 \log(p - p_0)$.

Determine $g_0(\lambda)$.

It follows from the results of Spitzer (1956), Woodroffe (1982) and Zhang (1992) that, for all $\lambda \in [0, \log n]$,

$$g_0(\lambda) = 2 \sum_{j=1}^{p-p_0} [P(\chi_{j+2}^2 > j\lambda)] + 2p_0.$$

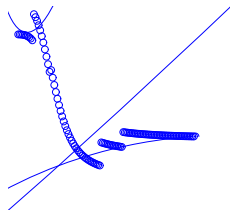
Note that

- $g_0(\lambda)$ is strictly decreasing.
- $g_0(0) = 2p$.
- $g_0(\log n) \rightarrow 2p_0$ as $n \rightarrow \infty$.

AMS improves.

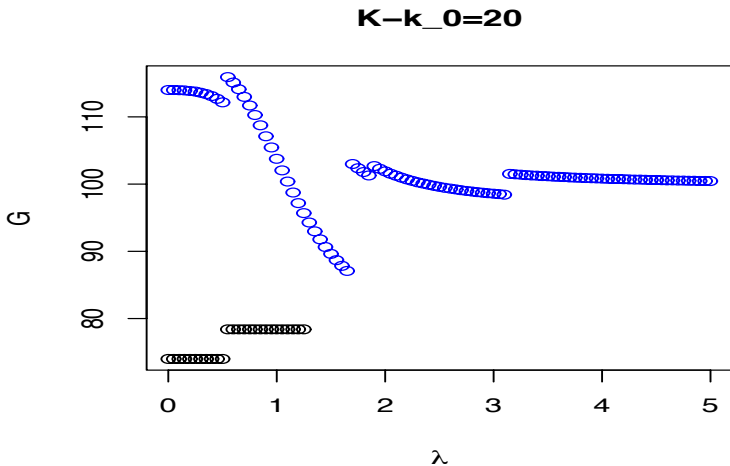
Consider a simulation study with $p_0 = 0$, $p - p_0 = 20$, $n = 404$ ($\log n = 6$), and $\sigma^2 = 1$.

The black points are $RSS(\hat{M}(\lambda)) - RSS(M_{p_0})$ and the blue points are $RSS(\hat{M}(\lambda)) + g_0(\lambda) - RSS(M_{p_0})$.



Generalized degrees of freedom

AMS may not work but how often?



Generalized degrees of freedom

Probability of correct selection:

$\hat{M}(\hat{\lambda}) = M_{p_0+}$	$[0, \log n]$	$[0.5, \log n]$	$[1, \log n]$	$[1.5, \log n]$	$[2, \log n]$
0	0.5457	0.5457	0.5457	0.6483	0.7539
1	0.0565	0.0565	0.0565	0.0681	0.0807
2	0.0312	0.0312	0.0312	0.0386	0.0474
3	0.0262	0.0262	0.0262	0.0320	0.0348
4	0.0239	0.0239	0.0239	0.0283	0.0249
5	0.0188	0.0188	0.0188	0.0227	0.0166
6	0.0156	0.0156	0.0156	0.0190	0.0103
7	0.0134	0.0134	0.0134	0.0169	0.0071
8	0.0136	0.0136	0.0136	0.0157	0.0051
9	0.0140	0.0140	0.0140	0.0151	0.0041
10	0.0155	0.0155	0.0155	0.0132	0.0039
11	0.0155	0.0155	0.0155	0.0107	0.0022
12	0.0153	0.0153	0.0153	0.0106	0.0018
13	0.0163	0.0163	0.0163	0.0097	0.0018
14	0.0177	0.0177	0.0177	0.0080	0.0015
15	0.0185	0.0185	0.0185	0.0074	0.0012
16	0.0210	0.0210	0.0210	0.0070	0.0008
17	0.0242	0.0242	0.0242	0.0074	0.0005
18	0.0212	0.0212	0.0212	0.0069	0.0006
19	0.0307	0.0307	0.0307	0.0065	0.0005
20	0.0452	0.0452	0.0452	0.0079	0.0003

Need a detailed description of $g_0(\lambda)$

Recall

$$\hat{\lambda} = \min_{\lambda > 0} \{ \lambda : RSS(\hat{M}(\lambda)) + g_0(\lambda) \}$$

and choose model $\hat{M}(\hat{\lambda})$ which retains the first $\hat{j}(\hat{\lambda})$ predictors.

- When $\lambda = 0$, $|\hat{M}(0)| = p$ for all realizations and $RSS(\hat{M}(0)) = \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_p)\mathbf{Y}$. Then $g_0(0) = 2p$.
- When $\lambda = \ln n$, $|\hat{M}(\ln n)| = p_0$ for almost all realizations and $RSS(\hat{M}(\ln n)) = \mathbf{Y}^T(\mathbf{I} - \mathbf{P}_{p_0})\mathbf{Y}$. Then $g_0(\ln n) = 2p_0$.

Note that

$$RSS(\hat{M}(0)) + 2p\sigma^2 - RSS(\hat{M}(\ln n)) + 2p_0\sigma^2 = \sigma^2 \sum_{k=1}^{p-p_0} (2 - V_k)$$

which is greater than 0 with probability close to 1 when $p - p_0$ is large.

Adaptive selection over $\lambda \in [0, 0.5] \cup \{\log n\}$

Show that $\hat{\lambda} = \log n$ with probability close to 1 by finding a bound on the following probability.

$$P \left(RSS(\hat{j}(\lambda)) + g_0(\lambda) < RSS(\hat{j}(\ln n)) + g_0(\ln n) \text{ for all } \lambda \in [0, 0.5] \right).$$

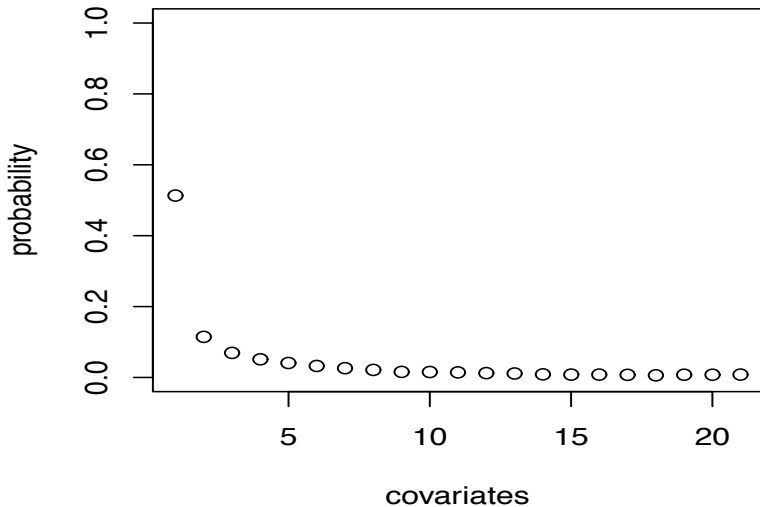
Note that

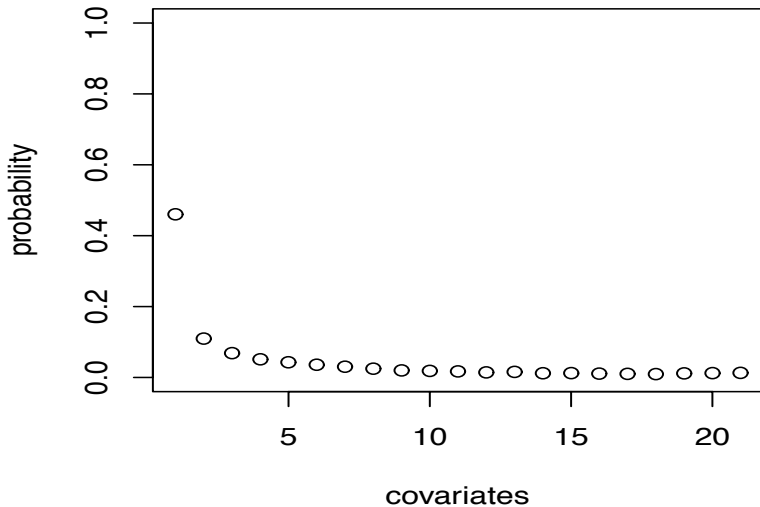
$$\begin{aligned} & P \left(V_1 + \cdots + V_{\hat{j}(\lambda)} < g_0(\lambda) \text{ for all } \lambda \in [0, 0.5] \right) \\ & \geq P(V_1 + \cdots + V_{p-p_0} < g_0(0) - 4) \\ & = P(V_1 + \cdots + V_{p-p_0} < 2(p - p_0) - 4). \end{aligned}$$

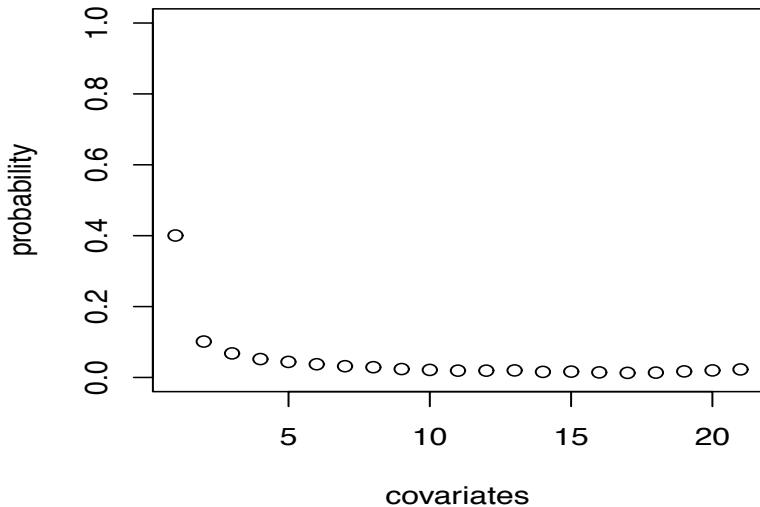
Note that

- $g_0(\lambda)$ is strictly decreasing and continuous on $\lambda \in [0, \ln n]$.
- For all $g_0(\ln n) < \delta \leq g_0(0)$, there exists a unique λ_δ such that $g_0(\lambda_\delta) = g_0(0) - \delta$.
- Claim: When $\delta = 4$, $0.5 \leq \lambda_\delta$.



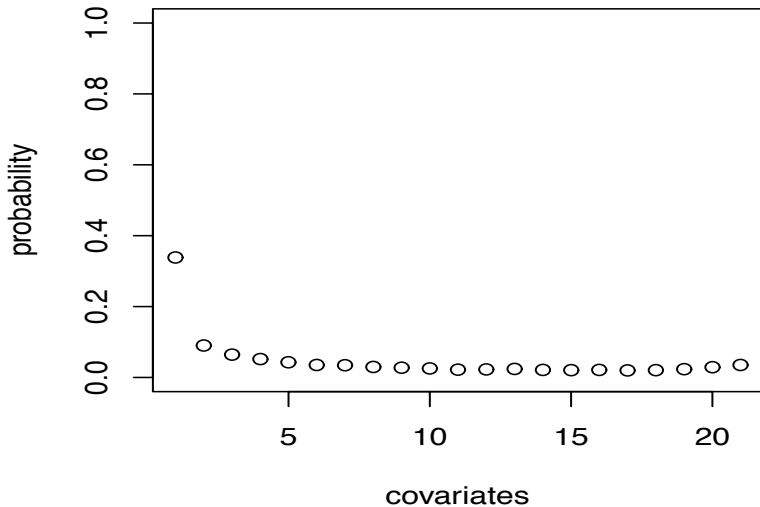
Simulation of $\{S_k(1.5)\}$ $\lambda = 1.5$ 

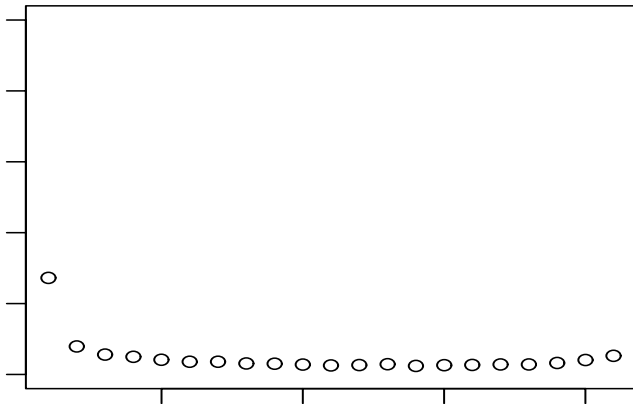
Simulation of $\{S_k(1.4)\}$ $\lambda = 1.4$ 

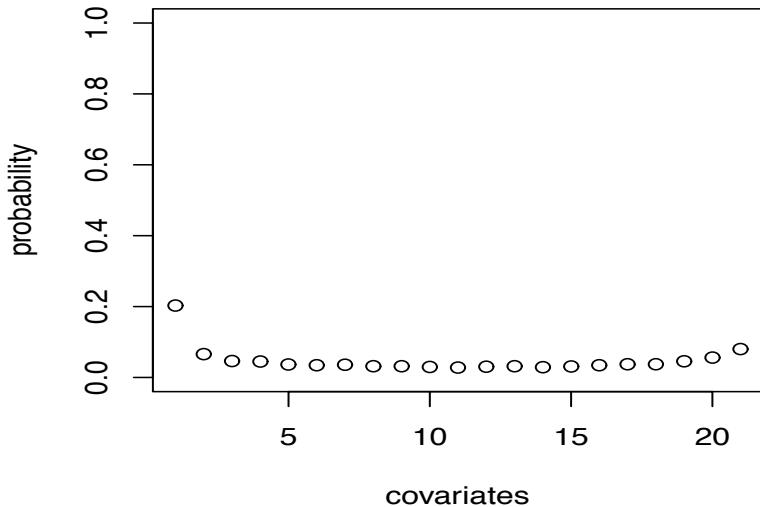
Simulation of $\{S_k(1.3)\}$ $= 1.3$ 

Simulation of $\{S_k(1.2)\}$

$$\lambda = 1.2$$

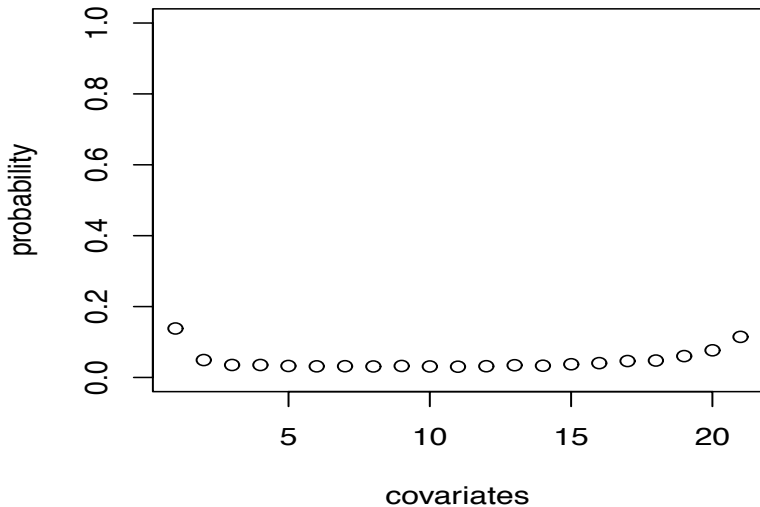


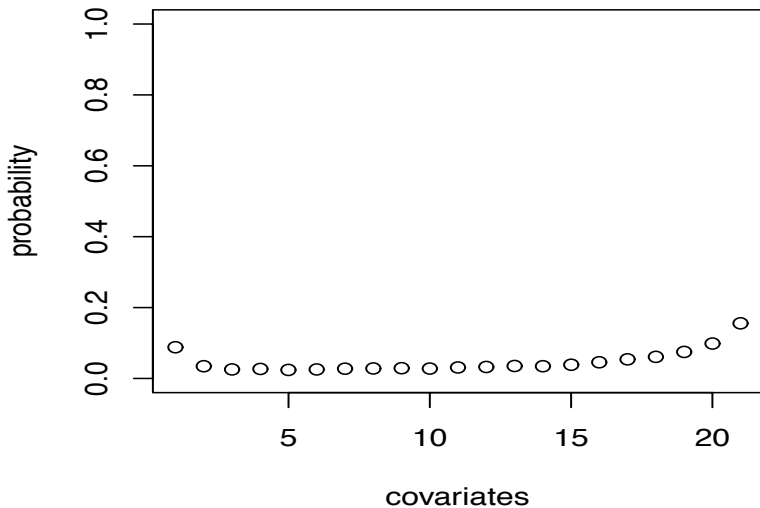
Simulation of $\{S_k(1.1)\}$ λ 

Simulation of $\{S_k(1.0)\}$ $\lambda = 1.0$ 

Simulation of $\{S_k(0.9)\}$

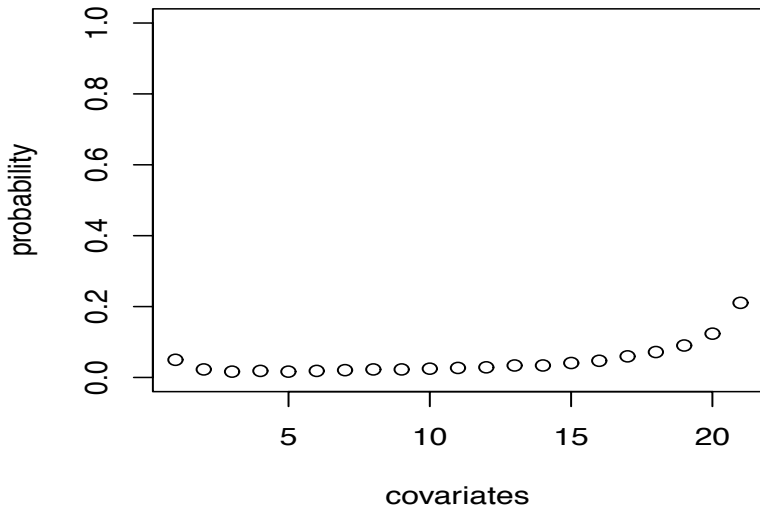
$$\lambda = 0.9$$



Simulation of $\{S_k(0.8)\}$ $= 0.8$ 

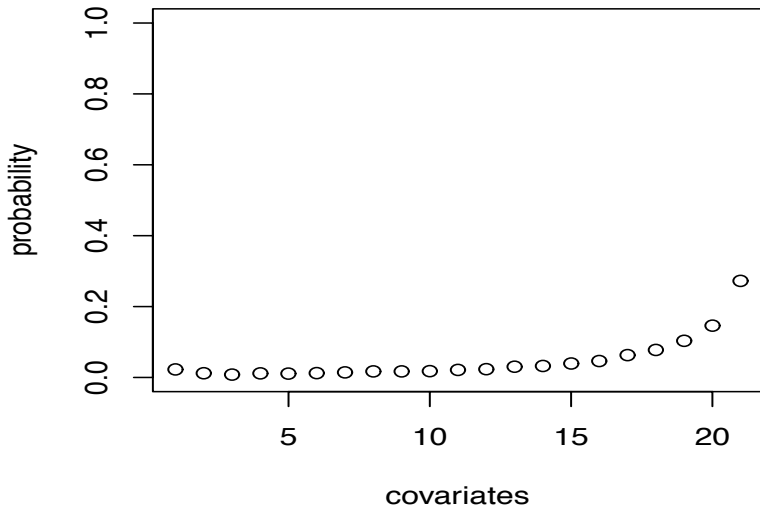
Simulation of $\{S_k(0.7)\}$

$$\lambda = 0.7$$



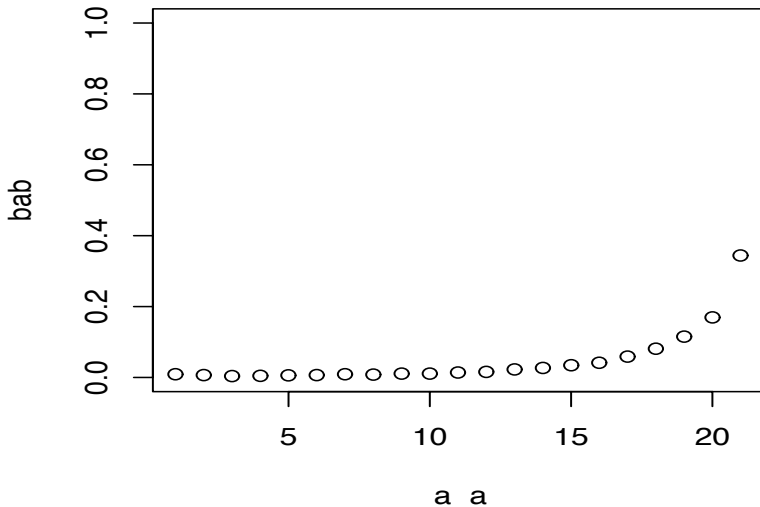
Simulation of $\{S_k(0.6)\}$

$$\lambda = 0.6$$



Simulation of $\{S_k(0.5)\}$

$$\lambda = 0.5$$



Conclusion

- When $\lambda \in (2, \log n]$, there are about 75% to choose the true model.
- The probability of selecting correct model decreases to 55% if $\lambda \in [1, 2) \cup [2, \log n]$.
- For the region of λ are $[0, \log n]$, $\in [0.5, \log n]$, or $n[1, \log n]$, there are no differences in the probability of correct selection.
 - We still cannot provide a good interpretation.