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

My Own Curiosity

- How do we get an unbiaed risk estimate (prediction error) with model selection?
  - C_p is derived to give an unbiased prediction error when a particular model M_k is used.
  - The prediction error of a linear model M_k is

    $$PE(\hat{\boldsymbol{\beta}}_k) = E\|\mathbf{Y}^* - \mathbf{X}_k\hat{\boldsymbol{\beta}}_k\|^2$$

    where $\mathbf{Y}^*$ comes from same distribution as $\mathbf{Y}$ in the training data.
- The first local minimum Lasso coupled with C_p sets almost all $\hat{\beta}_j$ ($\beta_j = 0$) to zero except those $\hat{\beta}_j$ exceeding the threshold $|\hat{\beta}|_{(p-\hat{p}_0+1)}$ when the regressors are orthogonal.
  - Note that

    $$\|\mathbf{y} - \hat{\mu}_k^{LS}\|^2 = \|\mathbf{y} - \hat{\mu}_k^{Lasso}\|^2 - k\frac{n}{p}\|\hat{\beta}|^2_{(p-k+1)}.$$

  Will the proposal made in Shen and Ye (2002, *JASA*) lead to Lasso estimate though least-squares estimate?

Linear Regression Models

Consider a linear regression model with normal error,

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

where

- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ matrix,
- $\boldsymbol{\beta} = (\beta_1 \ \ldots, \beta_p)^T$,
- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T = \mathbf{X}\boldsymbol{\beta}$,
- $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, and σ^2 is known.

Nested Models

We only consider the nested linear competing model

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

- 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
- $\boldsymbol{\mu}$ is estimated by

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

  where P_{M_k} is the projection matrix corresponding to model M_k.
- Its residual sum of squares is defined as

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

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 $\{M_k, k = 0, \ldots, p\}$, 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).

Unbiased risk estimate

Define the **loss function**

$$\ell\left(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\hat{M}}\right) = \frac{1}{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{M}})^T(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{M}}) + \sigma^2$$

and the **risk** is

$$E\left[\ell(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\hat{M}})\right] = E\left[\frac{1}{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{M}})^T(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{M}}) + \sigma^2\right],$$

where

$$\hat{\boldsymbol{\mu}}_{\hat{M}} = \sum_{k=0}^{p} \hat{\boldsymbol{\mu}}_{M_k} \cdot 1_{\{\hat{M}=k\}} = \sum_{k=0}^{p} P_{M_k}\mathbf{Y} \cdot 1_{\{\hat{M}=k\}}.$$

Generalized degrees of freedom

Define $\hat{M}(\lambda)$ to be the minimizer of

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

with respect to all competing models $\{M_k, k = 0, \ldots, p\}$. Note that

$$\frac{1}{n}\left\{RSS(\hat{M}(\lambda)) + 2E[\varepsilon^T(\hat{\boldsymbol{\mu}}_{\hat{M}(\lambda)} - \boldsymbol{\mu})]\right\}$$

are **unbiased risk estimator** for each $\lambda > 0$. Define

$$g_0(\lambda) = \frac{2}{\sigma^2} E\left[\boldsymbol{\epsilon}^T(\hat{\boldsymbol{\mu}}_{\hat{M}(\lambda)} - \boldsymbol{\mu})\right].$$

- $g_0(\lambda)/2$ is defined as the generalized degrees of freedom (GDF) by Ye (1998, *JASA*).

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

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

$$\hat{\lambda} = argmin_{\lambda > 0} \left\{ RSS(\hat{M}(\lambda)) + g_0(\lambda)\sigma^2 \right\}.$$

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

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

is the mean vector of the true model.

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

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

Set-up

Assume $\epsilon \sim N(\mathbf{0}, \mathbf{I})$.

- Note that $RSS(p_0) - RSS(p_0+1)$, $RSS(p_0+1)$ $-RSS(p_0+2)$, ..., $RSS(p-1) - RSS(p)$ consists of a sequence of iid random variables with χ_1^2 distribution.
- Write $RSS(p_0+j-1) - RSS(p_0+j)$ as V_j where $V_j \sim \chi_1^2$ and

$$C(k,\lambda) = \epsilon^T\epsilon - \delta_k(\lambda) = RSS(M_k) + \lambda k, \quad k = p_0, \ldots, p,$$

  where $\delta_k(\lambda) = \epsilon^T P_k \epsilon - \lambda k$.
- Consider the minimizer of $C(M_{p_0+j}, \lambda)$ over $0 \le j \le p - p_0$.
  - Define a partial sum process with drift $\lambda - 1$

$$S_j(\lambda) = \sum_{k=1}^{j}(-V_k + \lambda) \quad \text{and} \quad S_0(\lambda) = 0$$

  Find $\hat{j}$ to achieve the minimum of $\{S_j(\lambda), 0 \le j \le p - p_0\}$.
- Where the minimum should occur when $\lambda = 2$?
  at the very beginning or at the end

Determine $g_0(\lambda)$.

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

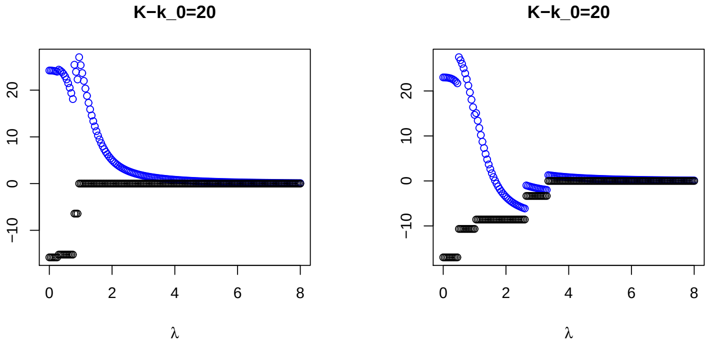
$$g_0(\lambda) = 2 \sum_{j=1}^{p-p_0} \left[P(\chi^2_{j+2} > j\lambda) \right] + 2p_0.$$

Note that

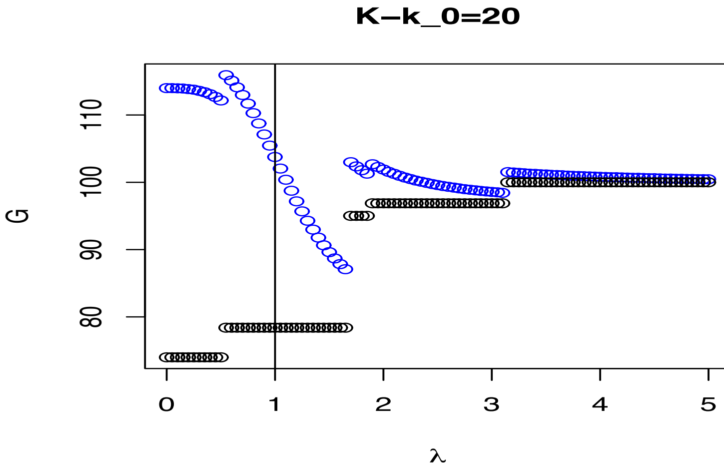
- $g_0(\lambda)$ is strictly decreasing.
- $g_0(0) = 2p$.
- $g_0(\log n) \to 2p_0$ as $n \to \infty$.

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 $RSS(\hat{M}(\lambda)) + g_0(\lambda) - RSS(M_{p_0})$.



AMS may not work but how often?



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

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

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

Estimate $g_0(\lambda)$ when $\lambda = 2$

Consider the case that $p - p_0 = 20$.

- For one realization, we have 2 observations 4.7 and 7.2 which are greater than 2. (i.e. $V_1 = 4.7$ and $V_{14} = 7.2$.)
- Minimum of random process $\{S_j(2), 0 \leq j \leq 20\}$ occurs at $\hat{j}(2) = 1$ for this realization.
  - Include one extra predictor x_{p_0+1}. (Note that $S_0(2) = 0$.)
- Let $N(\lambda)$ denote the number of V_j which are greater than λ.
  - Note that $N(2) \sim Bin(20, 0.1573)$
- $S_j(2)$: positive drift
  - $\hat{j}(2)$ cannot be large.

AMS improves when $\lambda \geq 2$.

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

$$\begin{aligned} & P\left(V_1 + \cdots + V_{\hat{j}(\lambda)} < g_0(\lambda) \text{ for all } \lambda \in [0.0.5]\right) \\ \geq \quad & P\left(V_1 + \cdots + V_{p-p_0} < g_0(0) - 4\right) \\ = \quad & P\left(V_1 + \cdots + V_{p-p_0} < 2(p - p_0) - 4\right). \end{aligned}$$

Note that

- $g_0(\lambda)$ is strickly 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 $\delta = 4$, $0.5 \leq \lambda_\delta$.

When $\delta = 4$, $0.5 \leq \lambda_\delta$.

Need to prove that, for given $\lambda < 1$,

$$P\left(\sum_{j=1}^{i+2} V_j > i\lambda\right) \to 1 \quad \text{for } i \text{ large enough.}$$

Then

$$g_0(0.5) \approx \sum_{j=1}^{20} P\left(\sum_{j=1}^{i+2} V_j > i\lambda\right) + ((p - p_0) - 20).$$

Cont.

Theorem 1 in Teicher(1984)

- Let Y_j be independent random variables with $E[Y_j] = 0$, $E[Y_j^2] = \sigma_j^2$ and $E|Y_j|^k \leq k! c_2^{k-2} \sigma_j^2 / 2$, for all $k \geq 3$ and some $c_2 > 0$.
  - Define $S_n = \sum_{j=1}^n a_{nj} Y_j$ where a_{nj} are arbitarary constants.
  - Set $v_n^2 = \sum_{j=1}^n a_{nj}^2 \sigma_j^2$ and $c_n = c_2 \max_{1 \leq j \leq n} |a_{nj}|$.

  Then, for $x > 0$,

$$P(S_n > x v_n) \leq \exp\left\{ \frac{-x^2}{2} \left(1 + \frac{c_n x}{v_n} \right)^{-1} \right\}.$$

In our case, $Y_j = V_j - 1$, $E[Y_j] = 0$, and $E[Y_j]^2 = \sigma_j^2 = 2$.
It follows from Lemma 5 in Henry Teicher(1984) that
$E|Y_j|^k = E|V_j - 1|^k \leq k! 2^{k-2}$ for all $k \geq 3$

Cont. $p - p_0 > 20$

For $\lambda = 0.5$, $c(0.5) = 0.9207$,

$$2\left(\sum_{i=1}^{20} P\left(\sum_{j=1}^{i+2} V_j > i\lambda\right) + ((p-p_0) - 20)\right) - g_0(\lambda)$$

$$\leq 2\left(\sum_{i=21}^{p-p_0} P\left(\sum_{j=1}^{i+2} V_j \leq i\lambda\right)\right) \leq 2\left(\sum_{i=21}^{\infty} P\left(\sum_{j=1}^{i+2} V_j \leq i\lambda\right)\right)$$

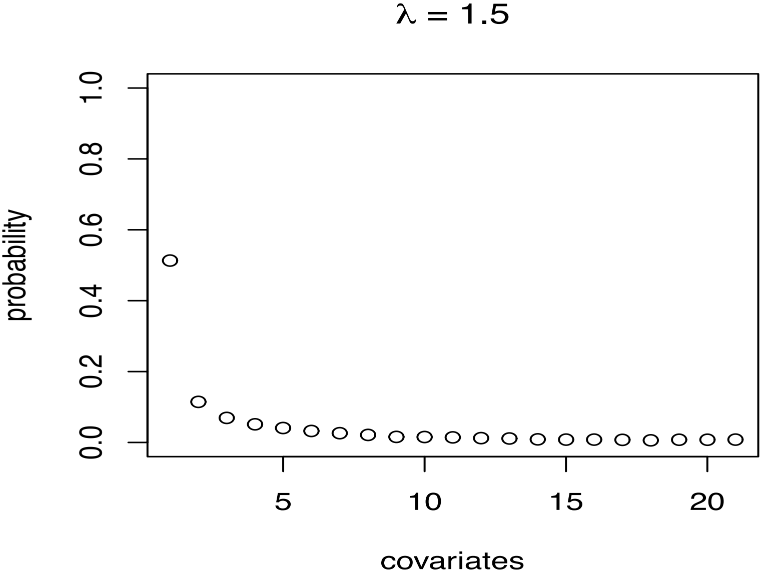
$$\leq 2 \cdot c(0.5)\frac{\exp\{-(21+2)(1-\lambda)^2/12\}}{1-\exp\{-(1-\lambda)^2/12\}} = 1.2186.$$

Moreover,

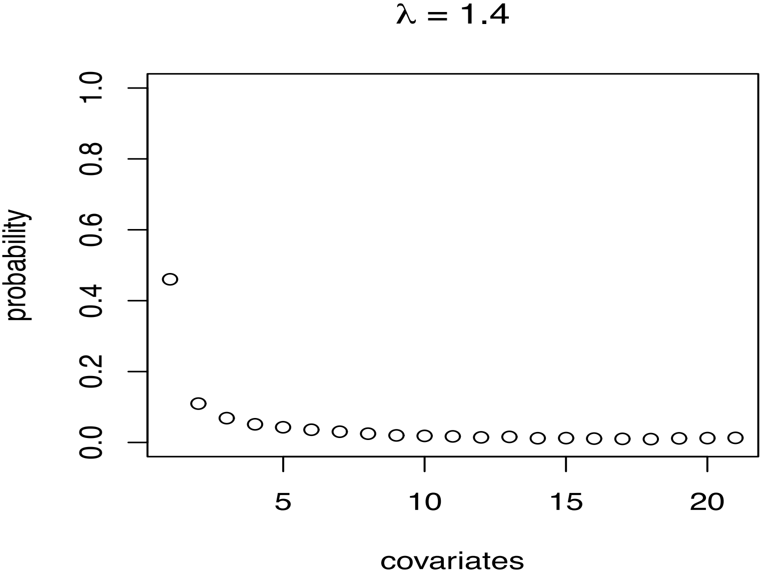
$$2\sum_{i=1}^{20} P\left(\sum_{j=1}^{i+2} V_j \leq i\lambda\right) = 38.1684 = 40 - 1.8316.$$

We conclude that $1.8316 + 1.2186 = 3.0502 < 4$ and
$g_0(0.5) > 2(p - p_0) - 4$ for $p - p_0 > 20$. ($P(\chi^2_{20} > 40) = 0.0050$

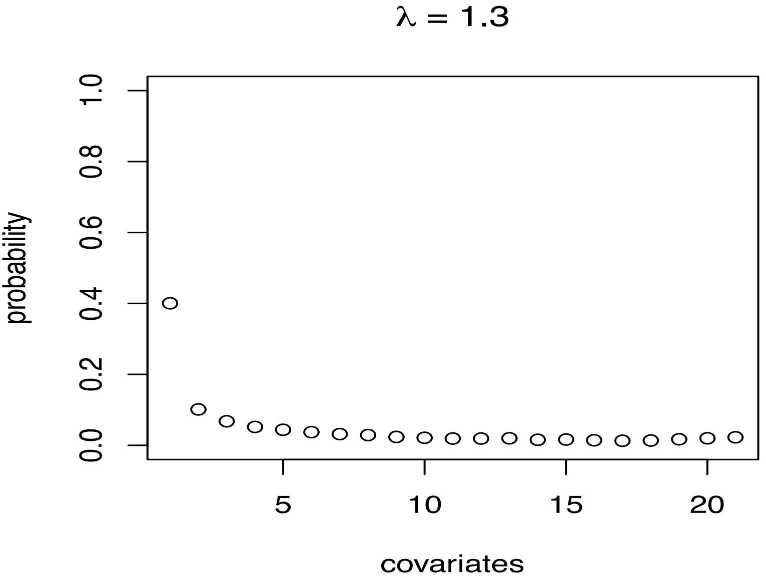
Simulation of $\{S_k(1.5)\}$



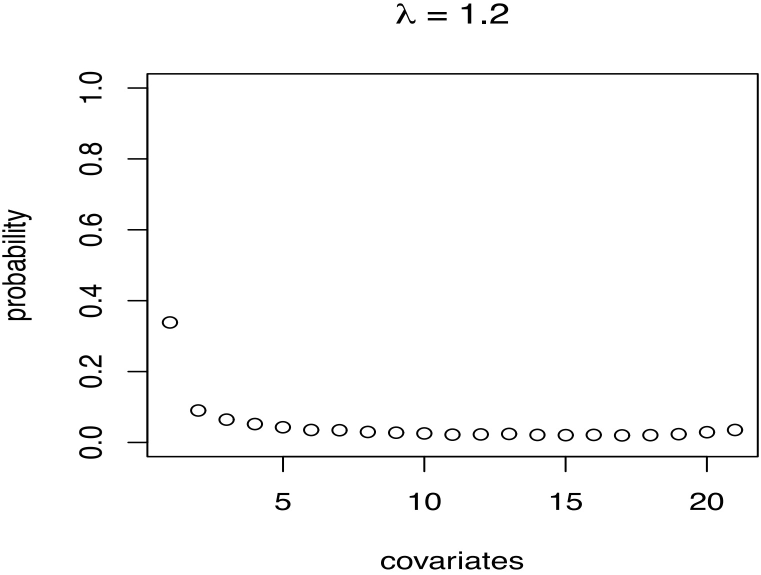
Simulation of $\{S_k(1.4)\}$



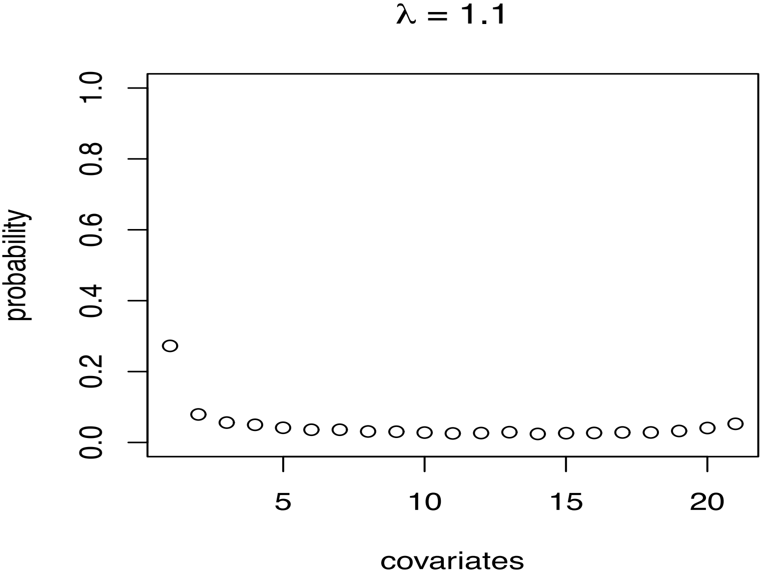
Simulation of $\{S_k(1.3)\}$



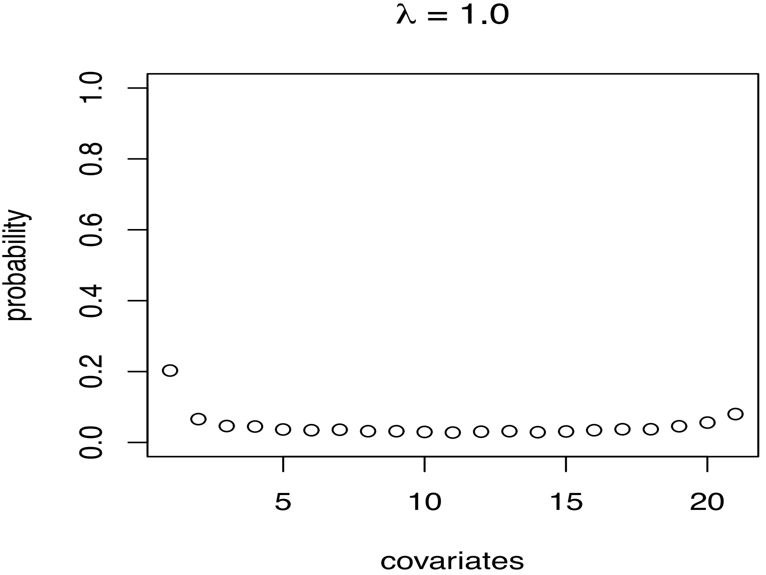
Simulation of $\{S_k(1.2)\}$



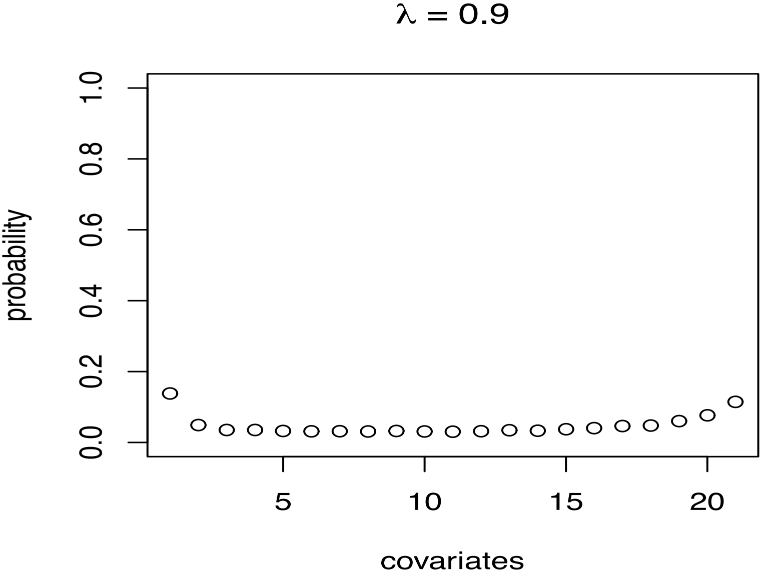
Simulation of $\{S_k(1.1)\}$



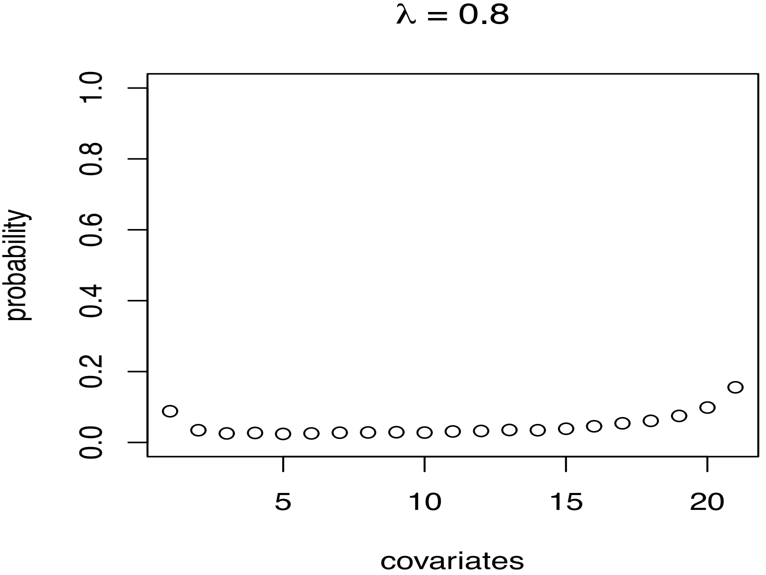
Simulation of $\{S_k(1.0)\}$



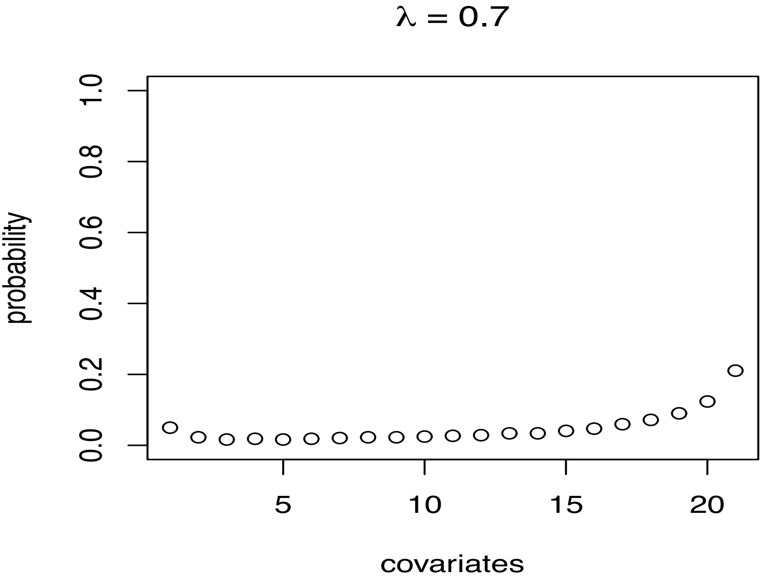
Simulation of $\{S_k(0.9)\}$



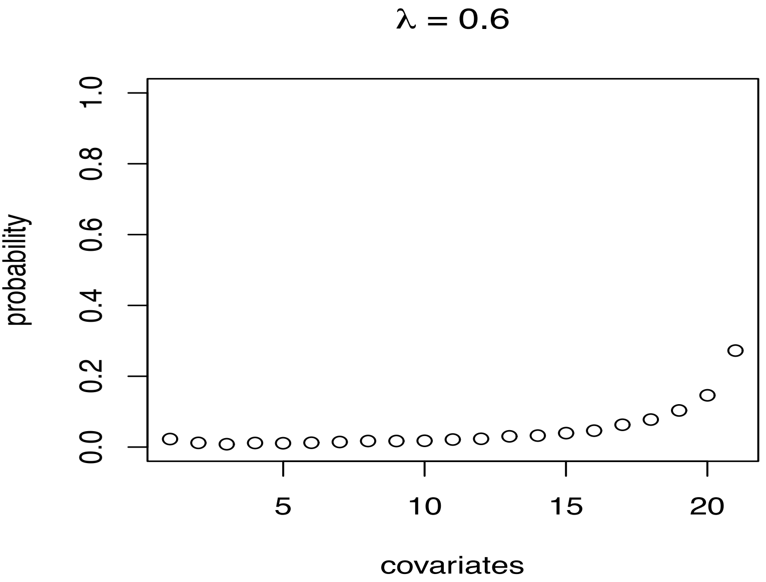
Simulation of $\{S_k(0.8)\}$



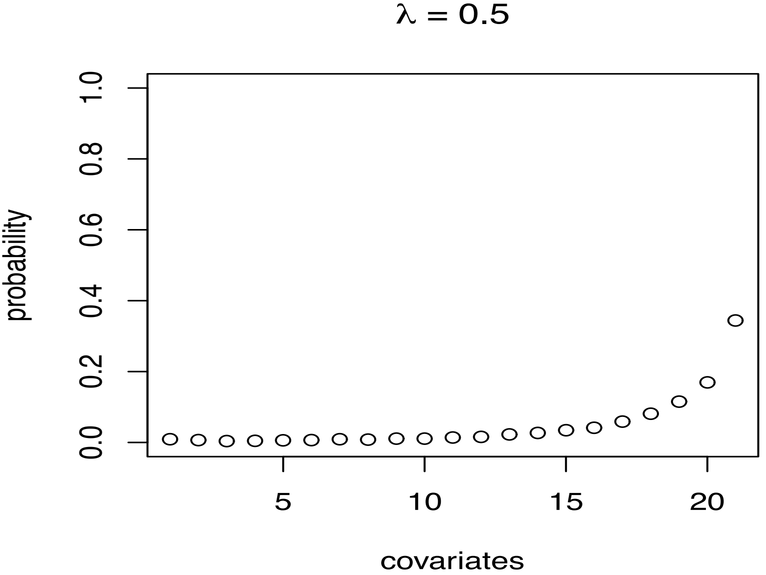
Simulation of $\{S_k(0.7)\}$



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



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



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.