Shrinkage tuning parameter selection with a diversing number of parameters

Donggyu Kim

May 2, 2012
The consistency of selection of shrinkage methods (LASSO and SCAD) relies on an appropriate choice of the tuning parameters. Then, how could we select the tuning parameters with a diversing number of parameters?

In the past literature, the method of generalized cross-validation (GCV) has been widely used. However, the asymptotic behavior of GCV is similar to that of AIC, which is a well-known loss efficient but selection inconsistent variable selection criterion. So, we can reasonably conjecture that the shrinkage parameter that is selected by GCV might not be able to identify the true model consistently.
Under fixed predictor dimension, the shrinkage parameter that is selected by a BIC-type criterion can identify the true model consistently (Wang et al. (2007b)). So, the authors thought that such a BIC or its slightly modified version can still find the true model consistently with a diversing number of parameters.

Modified BIC:

\[ BIC_S = \log(\hat{\sigma}_S^2) + |S| \frac{\log(n)}{n} C_n \]
Setting a model and notations

\[ Y = X^T \beta + \epsilon_i \]
where \( X \in \mathbb{R}^{n \times d} \) with \( E(X_i) = 0 \) and \( \text{var}(X_i) = 1 \) and \( \epsilon_i : \text{iid} \)
with \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = \sigma^2 \)
d can go to \( \infty \)

\( \beta_0 = (\beta_{01}, ..., \beta_{0d})^T \): true regression coefficient
\( \beta_{0j} \neq 0, \forall 1 \leq j \leq d_0, \beta_{0j} = 0, \forall j > d_0 \)

\( S = \{j_1, ..., j_{d_*}\} \): generic notation
\( X_S = \{X_{j_1}, ..., X_{j_{d_*}}\}^T \), \( \beta_S = \{\beta_{j_1}, ..., \beta_{j_{d_*}}\}^T \)
\( S_F = \{1, ..., d\} \): Full model, \( S_T = \{1, ..., d_0\} \): true model

\( \hat{\sigma}^2_S = \frac{SSE_S}{n} = \inf_{\beta_S} (\|Y - X_S\beta_S\|^2)/n \)

\( \Sigma \): covariance matrix of \( X_i \)

\( \tau_{\text{min}}(A) \): minimal eigenvalue of \( A \)
Technical conditions

- **Condition 1.** $X_i$ has componentwise finite fourth-order moment. i.e. $\max_{1 \leq j \leq d} E(X_i^4) < \infty$.

- **Condition 2.** There is a positive number $k$ such that $\tau_{\min}(\Sigma) \geq k$ for every $d > 0$.

- **Condition 3.** The predictor dimension satisfies that $\limsup (d/n^{k^*}) < 1$ for some $k^* < 1$.

- **Condition 4.** $\sqrt{\left[ n/(C_n d \cdot \log(n)) \right]} \min_{j \in S_T} |\beta_{0j}| \to \infty$, and $C_n d \cdot \log(n)/n \to 0$
Theorem 1: Assume technical conditions 1-4, $C_n \to \infty$, and that $\epsilon$ is normally distributed: we then have

$$P \left\{ \min_{S \not\supset S_T} (BIC_S) > (BIC_{S_F}) \right\} \to 1$$

Theorem 2: Assume technical conditions 1-4, $C_n \to \infty$, and that $\epsilon$ is normally distributed: we then have

$$P \left\{ \min_{S \neq S_T, S \supset S_T} (BIC_S) > (BIC_{S_T}) \right\} \to 1$$

Combining theorem 1 and 2 shows that the modified BIC can identify the true model consistently.
Estimators are obtained by optimizing the penalized least squares objective function

\[
Q_\lambda(\beta) = n^{-1} \| Y - X\beta \|^2 + \sum_{j=1}^{d} p_{\lambda,j}(|\beta_j|)
\]

where \( p_{\lambda,j}(\cdot) \) : penalty function.

\[
\hat{\beta}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,d})^T : \text{estimator}
\]

If \( p_{\lambda,j}(\cdot) \) is a function with its first-order derivative given

\[
p'_{\lambda,j}(t) = \lambda [I(t \leq \lambda) + I(t > \lambda)(a\lambda - t)_+] / [(a - 1)\lambda]
\]

with \( a=3.7 \), \( \hat{\beta}_\lambda \) becomes the SCAD estimator.

If \( p_{\lambda,j}(t) = \lambda w_j t \), \( \hat{\beta}_\lambda \) becomes the adaptive LASSO estimator.
Define the modified BIC for the shrinkage method

\[ BIC_S = \log(\hat{\sigma}^2_\lambda) + |S_\lambda| \frac{\log(n)}{n} C_n \]

where \( \hat{\sigma}^2_\lambda = \frac{SSE_\lambda}{n} \), \( SSE_\lambda = \| Y - X \hat{\beta}_\lambda \|^2 \), \( S_\lambda = \{ j : \hat{\beta}_\lambda \neq 0 \} \)

\( BIC_\lambda, BIC_{S_\lambda} \) are different notations.

\[ BIC_{S_\lambda} = \log(\hat{\sigma}^2_{S_\lambda}) + |S_\lambda| \frac{\log(n)}{n} C_n \]

where \( \hat{\sigma}^2_{S_\lambda} = \frac{SSE_{S_\lambda}}{n} = \inf_{\beta_{S_\lambda}} (\| Y - X_S \beta_{S_\lambda} \|^2) / n \)

Thus,

\[ BIC_\lambda \geq BIC_{S_\lambda} \]
They suggest the tuning parameter such that

$$\hat{\lambda} = \arg\min_\lambda (BIC_\lambda)$$

How can we be sure that $\hat{\lambda}$ is the optimal tuning parameter? Many researchers have demonstrated that there is a tuning parameter sequence $\lambda_n \to 0$ s.t $P(S_{\lambda_n} = S_T) \to 1$

We know that

$$BIC_{S_{\lambda_n}} = BIC_{S_T} + o_p(1)$$

$$BIC_{\hat{\lambda}} \leq BIC_{\lambda_n}$$

Now, intuitively, if $BIC_{S_{\lambda_n}} \approx BIC_{\lambda_n}$, we’ll obtain the optimal tuning parameter.
Condition 5. \[ \left\| p'_\lambda(\hat{\beta}_{\lambda_n,a}) \right\|^2 = o_p(\log(n)/n) \]
where \[ \hat{\beta}_{\lambda_n} = (\hat{\beta}_{\lambda_n,a}^T, \hat{\beta}_{\lambda_n,b}^T)^T, \] with \[ \hat{\beta}_{\lambda_n,a} = (\hat{\beta}_{\lambda_n,1}, \ldots, \hat{\beta}_{\lambda_n,d_0})^T \]
and \[ \hat{\beta}_{\lambda_n,b} = (\hat{\beta}_{\lambda_n,d_0+1}, \ldots, \hat{\beta}_{\lambda_n,d})^T \]

**Theorem 3:** Assume technical conditions 1-5, \( C_n \to \infty \), and that \( \epsilon \) is normally distributed: we then have

\[ P \{ S_{\hat{\lambda}} = S_T \} \to 1 \]
Assume condition 5. Then, $\text{BIC}_{\lambda_n} = \text{BIC}_{S\lambda_n} + o_p(\log(n)/n)$ (proof outline)

$\text{BIC}_{\lambda_n} - \text{BIC}_{S\lambda_n} = \log\left(\frac{\text{SSE}_{\lambda_n}}{\text{SSE}_{S\lambda_n}}\right)$

$= \log(1 + [\text{SSE}_{\lambda_n} - \text{SSE}_{S\lambda_n}]/\text{SSE}_{S\lambda_n})\cdots(\ast)$

By Bai and Silverstein (2006), the condition 5 implies that

$\text{SSE}_{\lambda_n} = \text{SSE}_{S\lambda_n} + o_p(\log(n))$

By Taylor expansion, (\ast) will be

$\text{BIC}_{\lambda_n} - \text{BIC}_{S\lambda_n} = o_p(\log(n)/n)$

If $\left\|p'_\lambda(\hat{\beta}_{\lambda_n,a})\right\|^2 = o_p(1)$, $\text{BIC}_{\lambda_n} = \text{BIC}_{S\lambda_n} + o_p(1)$.

(Proof is similar with the previous one.)

Why do we need $'o_p(\log(n)/n)'$ convergent rate?
Proof of Thm 3)
Define \( \Omega_+ = \{ \lambda > 0 : S_\lambda \not\supseteq S_T \} \), \( \Omega_0 = \{ \lambda > 0 : S_\lambda = S_T \} \) and \( \Omega_- = \{ \lambda > 0 : S_\lambda \supset S_T, S_\lambda \neq S_T \} \)

**Case 1**: underfitted model (i.e. \( \lambda \in \Omega_- \))

By the condition 5, we have \( BIC_{\lambda_n} = BIC_{S_{\lambda_n}} + o_p(\log(n)/n) \)

Then, with probability tending to 1, we have
\[
\inf_{\lambda \in \Omega_-} (BIC_{\lambda}) - BIC_{\lambda_n} \geq \inf_{\lambda \in \Omega_-} (BIC_{S_{\lambda}}) - BIC_{S_T} + o_p(\log(n)/n)
\]

by \( BIC_{\lambda} \geq BIC_{S_{\lambda}}, BIC_{S_{\lambda}} = BIC_{S_T} + o_p(1) \)

\[
\geq \min_{S \not\supseteq S_T} (BIC_S) - BIC_{S_T} + o_p(\log(n)/n)
\]

By Thm1 and Thm2,
\[
P(\min_{S \not\supseteq S_T} (BIC_S) - BIC_{S_T} + o_p(\log(n)/n) > 0) \to 1
\]

Thus,
\[
P(\inf_{\lambda \in \Omega_-} (BIC_{\lambda}) - BIC_{\lambda_n} > 0) \to 1
\]
Case 2: Overfitted model (i.e. $\lambda \in \Omega_+$)

By the similar argument, we can obtain the inequality

$$\inf_{\lambda \in \Omega_+} (BIC_\lambda) - BIC_{\lambda_n} \geq \min_{S \supsetneq S_T, S \neq S_T} (BIC_S) - BIC_{S_T} + o_p\left(\frac{\log(n)}{n}\right)$$

By thm2,

$$P\left(\min_{S \supsetneq S_T, S \neq S_T} (BIC_S) - BIC_{S_T} + o_p\left(\frac{\log(n)}{n}\right) > 0\right) \to 1$$

Thus,

$$P\left(\inf_{\lambda \in \Omega_+} (BIC_\lambda) - BIC_{\lambda_n} > 0\right) \to 1$$

By these results, we can know that

$$P(\hat{\lambda} \in \Omega_0) \to 1$$

It means that

$$P(\hat{S}_{\hat{\lambda}} = S_T) \to 1$$
How many $n$ do we need when $X$ variables are highly correlated?

Simulation setting:
   ex) Uniform, Exponential, Poisson, Log normal, etc. $\varepsilon_i$ are Double exponential $(0, 1/\sqrt{2})$

2. Correlation Structure:
   $$\rho_{i,j} = \begin{cases} 
   1 & \text{if } i = j \\
   0.99 & \text{if } j = i + d_0 \text{ and } i \leq d_0 \\
   0 & \text{if o.w.}
   \end{cases}$$

3. Dimension: $d = [4n^{1/4}] - 5$, $d_0 = 5$
Figure 1: MRME (Median of the relative model error), MS (average model size), and CM (percentage of the correctly identified true models)
Questions

\[ \Sigma = \begin{bmatrix}
\Sigma_1 & 0 & \ldots & 0 \\
0 & \Sigma_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \Sigma_k
\end{bmatrix} \quad \text{Blockmized diagonal matrix} \]

Let \( p_i \) be the dimension of group \( i \).

Q1, If \( p_i \) is satisfied technical conditions 3 and 4, can we apply this argument to the case, \( p > n \)? \((p = \sum_{i=1}^{k} p_i)\)

Idea)

\( Y = \beta X + \gamma Z + \epsilon, \ X \ and \ Y \ linearly \ indep \)

If we estimate \( \beta \ and \ \gamma \) for each models s.t.

\( Y = \beta X + \epsilon, \ Y = \gamma Z + \epsilon \)

this result is the same with the full model’s one.

Q2, If so, when \( k \rightarrow \infty \), what kind of condition do we need?
Figure 2: MRME (Median of the relative model error), MS (average model size), and CM (percentage of the correctly identified true models)