Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters

H. Wang, B. Li, and C. Leng

Presenter: Xu He

April 9, 2010

Unpenalized estimators

- AIC
  - Loss efficient
  - Selection inconsistent
- BIC
  - Consistent
  - Computationally expensive for an exhautive search

Penalized estimators

- LASSO (least absolute shrinkage and selection operator)
- SCAD (smoothly clipped absolute deviation)
- Consistent if tuning parameters is appropriate, fixed or diverging predictor dimension

Tuning parameters

- GCV
  - Loss efficient
  - Selection inconsistent, at least for SCAD
- BIC
  - Consistent for SCAD under fixed predictor dimension
  - Consistent for adaptive LASSO under fixed predictor dimension
- Slightly modified BIC
  - Serving as a unpenalized estimator itself, consistent
  - Consistent for LASSO and SCAD, for fixed and diverging predictor dimension

Notations

- Y: response by n iid observations
- X: d-dimentional predictor; standardized
- $S = \{j_1, \ldots, j_c\}$: a candidate model
- $|S|$: size of the model S
- S_F: Full model
- S_T: True model
- $d_0 = |S_T|$
- $\hat{\sigma}_S^2 = SSE_S/n$

Modified BIC criterion

- $$BIC_S = \log(\hat{\sigma}_S^2) + |S| \times \frac{\log(n)}{n} \times C_n$$
- $C_n > 0$, $C_n \to \infty$
- If $C_n = 1$, this is the traditional BIC
- Traditional BIC is consistent for fixed predictor dimension
- It is hard to prove that traditional BIC is consistent for diverging predictor dimension
- In this paper proved that Modified BIC is consistent for diverging predictor dimension

BIC consistently not overfitting

- Suppose S is an arbitrary overfitted model, i.e., $S \supset S_t$, $|S| > |S_t|$.
- $$BIC_S - BIC_{S_T} = \log(\frac{\hat{\sigma}_S^2}{\hat{\sigma}_{S_T}^2}) + (|S| - |S_t|) \times \frac{\log(n)}{n} \times C_n$$
- $$\log(\frac{\hat{\sigma}_S^2}{\hat{\sigma}_{S_T}^2}) = O_p(2\log\frac{n - |S|}{n - d_0}) = O_p(n^{-1})$$
- $$(|S| - |S_t|) \times \frac{\log(n)}{n} \times C_n > C_n\frac{\log(n)}{n}$$
- $$P(BIC_S > BIC_{S_T}) \to 1$$
- $$P(\min_{S \supset S_T} BIC_S > BIC_{S_T}) \to 1$$

Technical Conditions

- $$(C1) \max_{1 \leq j \leq d} EX_{ij}^4 < \infty$$
- (C2) There exists a $\kappa > 0$ such that $\tau_{\min}(\Sigma) \geq \kappa$ for every $d > 0$
  - Σ is the covariance matrix of X_i
  - $\tau_{\min}(A)$ is the minimal eigenvalues of an arbitrary positive definite matrix A
- $$(C3) \limsup d/n^q < 1$$

  for some $q < 1$
- $$(C4) C_n d \log n/n \to 0$$

  and

  $$(C_n d \log n/n) \times \lim \inf_{n \to \infty} \{\min_{j \in S_t} |\beta_{0,j}|\}^{-2} \to 0$$

BIC with unpenalized estimators

Theorem (1)

Assume conditions (C1)-(C4), $C_n \to \infty$, ϵ normally distributed, then

$$P(\min_{S \not\supseteq S_t} BIC_S > BIC_{S_F}) \to 1$$

$C_n \to \infty$ but the rate can be arbitrarily slow. For example, $C_n = \log \log d$

Theorem (2)

Assume conditions (C1)-(C4), $C_n \to \infty$, ϵ normally distributed, then

$$P(\min_{S \supset S_t} BIC_S > BIC_{S_T}) \to 1$$

Modified BIC criterion is concsistent

Proof of Theorem 1

Define $\tilde{\beta}$ be the unpenalized full model estimator. By condition C1, C2 and C3, we know that

$$E||\tilde{\beta} - \beta_0||^2 = trace(cov(\tilde{\beta})) = \sigma^2 trace((X^T X)^{-1})$$

$$\leq dn^{-1}\sigma^2\tau_{\min}^{-1}(n^{-1}X^T X) = O_p(d/n)$$

This implies that $||\tilde{\beta} - \beta_0||^2 = O_p(d/n)$.

Next, for an arbitrary model S, define $\hat{\beta}^{(S)} = \arg\min_{\{\beta:\beta_j=0,\forall j\notin S\}} ||Y - X\beta||^2$. We then have

$$\min_{S\not\supset S_T} ||\hat{\beta}^{(S)} - \tilde{\beta}||^2 \geq \min_{S\not\supset S_T} ||\hat{\beta}^{(S)} - \beta_0||^2 - ||\tilde{\beta} - \beta_0||^2 \geq \min_{j\in S_T} \beta_{0,j}^2 - O_p(d/n)$$

Proof of Theorem 1

By C4, we know $\min_{j \in S_T} \beta_{0,j}^2 - O_p(d/n)$ is positive with probability tending to one. Next,

$$\min_{S \not\supseteq S_T} (BIC_S - BIC_{S_F}) \geq \min_{S \not\supseteq S_T} \log(\hat{\sigma}_S^2 / \hat{\sigma}_{S_F}^2) - C_n d \log n/n$$

Note that the right hand side of the above equation can be written as

$$\min_{S \not\supseteq S_T} \log \left(1 + \frac{(\hat{\beta}^{(S)} - \tilde{\beta})^T (n^{-1} X^T X)(\hat{\beta}^{(S)} - \tilde{\beta})}{\hat{\sigma}_{S_F}^2} \right) - C_n d \log n/n$$

$$\geq \min_{S \not\supseteq S_T} \log \left(1 + \frac{\hat{\tau}_{\min} ||\hat{\beta}^{(S)} - \tilde{\beta}||^2}{\hat{\sigma}_{S_F}^2} \right) - C_n d \log n/n$$

where $\hat{\tau}_{\min} \doteq \tau_{\min}(n^{-1} X^T X)$.

Proof of Theorem 1

One can varify that $\log(1+x) \geq \min\{0.5x, \log 2\}$ for any $x > 0$. Consequently, it is further bounded by

$$\geq \min_{S \not\supseteq S_T} \min \left(\log 2, \frac{\hat{\tau}_{\min} ||\hat{\beta}^{(S)} - \tilde{\beta}||^2}{\hat{\sigma}^2_{S_F}} \right) - C_n d \log n/n$$

By C4, we have $\log 2 - C_n d \log n/n \geq 0$ with probability tending to one. Therefore, we only need to show that

$$\min_{S \not\supseteq S_T} \left(\frac{\hat{\tau}_{\min} ||\hat{\beta}^{(S)} - \tilde{\beta}||^2}{\hat{\sigma}^2_{S_F}} \right) - C_n d \log n/n$$

is positive.

Proof of Theorem 1

As ϵ is Normally distributed, $\hat{\sigma}^2_{S_F} \to_p \sigma^2$. Also, $\hat{\tau}_{\min} \to \tau_{\min} = \tau_{\min}(\Sigma)$ with probability tending to one.

Therefore, it is further bounded by

$$\geq \frac{\tau_{\min}}{\sigma^2}(\min_{j \in S_T} \beta^2_{0,j} - O_p(d/n))(1 + o_p(1)) - C_n d \log n/n$$

$$= C_n d \log n/n \times \frac{\tau_{\min}}{\sigma^2}(C_n d \log n/n \times \min_{j \in S_T})(1 + o_p(1)) - C_n d \log n/n$$

which is guaranteed to be positive asymptotically under C4.

Therefore, with probability tending to one,

$$\min_{S \not\supseteq S_T} \log \left(1 + \frac{\hat{\tau}_{\min}||\hat{\beta}^{(S)} - \tilde{\beta}||^2}{\hat{\sigma}^2_{S_F}}\right) - C_n d \log n/n$$

is positive. Therefore asymptotically

$$\min_{S \not\supseteq S_T} (BIC_S - BIC_{S_F}) > 0.$$

BIC with penalized estimators

- Shrinkage estimators:

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

- $\dot{p}_{\lambda,j}()$ is first order derivatiove of $p_{\lambda,j}()$
- resulting estimator by $\hat{\beta}_\lambda$

BIC with penalized estimators

- $$BIC_\lambda = \log(\hat{\sigma}^2_\lambda) + |S_\lambda| \frac{\log n}{n} C_n$$
- $\hat{\sigma}^2_\lambda = SSE_\lambda / n$
- S_λ is the model identified by $\hat{\beta}_\lambda$
- SSE_{S_λ} is the residual sum squares with the unpenalized estimator based on S_λ
- Use the optimal tuning parameter $\hat{\lambda} = \arg\min_\lambda BIC_\lambda$, which gives the model $S_{\hat{\lambda}}$

BIC with penalized estimators

- Assume $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda,a}, \hat{\beta}_{\lambda,b})$ where $\hat{\beta}_{\lambda,a}$ for nonzero coefficients and $\hat{\beta}_{\lambda,b}$ for zero coefficients
- There exist a tuning parameter $\lambda_n \to 0$ such that with probability tending to one $\hat{\beta}_{\lambda,b} = 0$ and $\hat{\beta}_{\lambda,a}$ efficient
- Asymptotically we must have $\hat{\beta}_{\lambda_n,a}$ being the minimizer of

$$Q_\lambda^*(\beta_{S_T}) = n^{-1}||Y - X_{S_T}\beta_{S_T}||^2 + \sum_{j=1}^{d_0} p_{\lambda_n,j}(|\beta_j|)$$

BIC with penalized estimators

- With probability tending to one, we must have

$$\begin{aligned}\hat{\beta}_{\lambda_n,a} &= \{n^{-1}X_{S_T}^T X_{S_T}\}^{-1}\{n^{-1}X_{S_T}^T Y + 1/2\mathrm{sgn}(\hat{\beta}_{\lambda_n,a})\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,a}|)\} \\ &= \hat{\beta}_{S_T} + 1/2\{n^{-1}X_{S_T}^T X_{S_T}\}^{-1}\mathrm{sgn}(\hat{\beta}_{\lambda_n,a})\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,a}|)\end{aligned}$$

- $\hat{\beta}_{S_T} = \{n^{-1}X_{S_T}^T X_{S_T}\}^{-1}\{n^{-1}X_{S_T}^T Y\}$
- $\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,a}|) = \{\dot{p}_\lambda(|\hat{\beta}_{\lambda_n,j}|) \mid j = 1, \ldots, d_0\}$
- $\mathrm{sgn}(\hat{\beta}_{\lambda_n,a})$ is a diagonal matrix with the jth diagonal component given by $\mathrm{sgn}(\hat{\beta}_{\lambda_n,j})$.

BIC with penalized estimators

- We need to show that BIC_{λ_n} and $BIC_{S_{\lambda_n}}$ are sufficiently similar
- It suffices to show that

$$SSE_{\lambda_n} = SSE_{S_{\lambda_n}} + o_p(\log_n)$$

- It suffices to show that

$$||\dot{p}_\lambda(\hat{\beta}_{\lambda_n,a})||^2 = o_p(\log n/n)$$

which is reasonable

Theorem (3)

Assume conditions (C1)-(C4), $C_n \to \infty$, ϵ normally distributed, $||\dot{p}_\lambda(\hat{\beta}_{\lambda_n,a})||^2 = o_p(\log n/n)$ then

$$P(S_{\hat{\lambda}} = S_T) \to 1$$

Proof of Theorem 3

Define $\Omega_- = \{\lambda > 0 : S_\lambda \not\supseteq S_T\}$, $\Omega_0 = \{\lambda > 0 : S_\lambda = S_T\}$, and $\Omega_+ = \{\lambda > 0 : S_\lambda \supset S_T\}$.

Case 1, with underfitted model, i.e., $\lambda \in \Omega_-$.

Firstly, we have $BIC_{\lambda_n} = BIC_{S_{\lambda_n}} + o_p(\log n/n)$. Then with proability tending to 1, we have

$$\inf_{\lambda \in \Omega_-} BIC_\lambda - BIC_{\lambda_n} \geq \inf_{\lambda \in \Omega_-} BIC_{S_\lambda} - BIC_{S_{\lambda_n}} + o_p(\log n/n)$$

$$\geq \min_{S \not\supseteq S_T} BIC_{S_\lambda} - BIC_{S_{\lambda_n}} + o_p(\log n/n)$$

By Theorem 1 and Theorem 2,

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

Proof of Theorem 3

Case 2, with voerfitted model, i.e., $\lambda \in \Omega_+$.

Similarly,

$$\inf_{\lambda \in \Omega_+} BIC_\lambda - BIC_{\lambda_n} \geq \min_{S \supset S_T} BIC_{S_\lambda} - BIC_{S_{\lambda_n}} + o_p(\log n/n)$$

We can find a positive number η such that $\min_{S \supset S_T} BIC_{S_\lambda} - BIC_{S_{\lambda_n}} > \eta \log n/n$ with probability tending to 1.

Similarly,

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

Numerical studies

- Example 1: $d = [4n^{1/4}] - 5$, $d_0 = 5$
- Example 2: $d = [7n^{1/4}]$, $d_0 = [d/3]$
- Median of the relative model error (MRME)
- Average model size (MS)
- Percentage of the correctly identified true models (CM)