Extended Bayesian Information Criteria for Model Selection with Large Model Spaces

Jiahua Chen, University of British Columbia
Zehua Chen, National University of Singapore

(Biometrika, 2008)
Variable Selection

Classical Criteria

- Akaike Information Criterion (Akaike, 1973)
- Bayesian Information Criterion (Schwarz, 1978)
- Cross-Validation (Stone, 1974)

Large $P$, small $n$

- ex. genome-wide association studies
- The above criteria: select many spurious covariates
- A great challenge
Bayesian Information Criterion (BIC)

- \{(y_i, x_i) : i = 1, \ldots, n\}: independent observations
- \(f(y_i|x_i, \theta), \theta \in \mathbb{R}^P\)
- \(L_n(\theta) = \prod_{i=1}^{n} f(y_i|x_i, \theta)\)
- \(s \subset \{1, \ldots, P\}\)
- \(\theta(s)\): components outside \(s\) being set to 0

\[
\text{BIC}(s) = -2 \log L_n\{\hat{\theta}(s)\} + \nu(s) \log n,
\]

where \(\hat{\theta}(s)\): MLE of \(\theta(s)\), \(\nu(s)\): # of components in \(s\).
BIC: approximate Bayes approach

- $S$: model space under consideration.
- $p(s)$: prior probability of model $s$.
- Posterior probability of model $s$ is

\[
\Pr(s|Z) \propto \Pr(Z|s)p(s)
\]

where $\Pr(Z|s) = \int \Pr(Z|\theta(s), s)\Pr(\theta(s)|s)d\theta(s)$. 
BIC: approximate Bayes approach

- $s^*$ that maximizes $\log \Pr(Z|s)p(s)$
- Approximation to the integral followed by some simplifications

\[
\log \Pr(Z|s) = \log \ln \{\hat{\theta}(s)\} - \frac{\nu(s)}{2} \log n + O(1)
\]

- So,

\[
\log \Pr(Z|s)p(s) = \log \Pr(Z|s) + \log p(s)
= \log \ln \{\hat{\theta}(s)\} - \frac{\nu(s)}{2} \log n + O(1) + \log p(s)
= -\frac{1}{2} \text{BIC} + O(1) + \log p(s)
\]
Constant prior assumption behind BIC

- An implicit assumption underlying the use of BIC
  - $p(s)$ is constant for $s$ (uniform prior)
  - This may not be reasonable with large $P$
- $S_j$: class of models containing $j$ covariates
  - Ex. $S_1$: the collection of models with a single covariate.
- Under constant prior, $\Pr(S_j) \propto \tau(S_j)$, where $\tau(S_j)$: size of $S_j$
- Ex. $P = 1000$, $\tau(S_1) = 1000$ vs $\tau(S_2) = 1000 \times 999/2$
- The constant prior prefers larger models.
Extended BIC

- $S = \bigcup_{j=1}^{P} S_j$
- $\Pr(S_j) \propto \tau^\xi(S_j)$, where $0 \leq \xi \leq 1$
- $\Pr(s|S_j) = 1/\tau(S_j)$ for any $s \in S_j$
  (all models in $S_j$ are equally plausible)
- $p(s) \propto \tau^{-\gamma}(S_j)$ for $s \in S_j$, where $\gamma = 1 - \xi$
- Extended BIC

$$BIC_\gamma(s) = -2 \log L_n\{\hat{\theta}(s)\} + \nu(s) \log n + 2\gamma \log \tau(S_j),$$

where $0 \leq \gamma \leq 1$
Consistency of the Extended BIC

- $P = p_n = O(n^\kappa)$ as $n \to \infty$ for $\kappa > 0$
- Model

$$y_n = X_n \beta + \epsilon_n,$$

where $\epsilon_n \sim N(0, \sigma^2 I_n)$
- $s_0$: the smallest subset of $\{1, ..., p_n\}$ such that

$$\mu_n = E(y_n) = X_n(s_0) \beta(s_0),$$

where $X_n(s_0), \beta(s_0)$: design matrix and coefficients corresponding to $s_0$
- Call $s_0$ the true submodel
Asymptotic identifiability

- $K_0 = \nu(s_0)$
- $H_n(s)$: projection matrix of $X_n(s)$
- $\Delta_n(s) = \|\mu_n - H_n(s)\mu_n\|^2$
- Condition 1: Asymptotic identifiability. Model 1 with true submodel $s_0$ is asymptotically identifiable if

$$\lim_{n \to \infty} \min\{(\log n)^{-1} \Delta_n(s) : s \neq s_0, \nu(s) \leq K_0\} = \infty.$$ 

- Any other model of comparable size cannot predict the response well.
Consistency of the Extended BIC

Theorem

Assume that \( p_n = O(n^\kappa) \) for some constant \( \kappa \). If \( \gamma > 1 - 1/(2\kappa) \), then, under the asymptotic identifiability condition,

\[
\Pr \left[ \min \{ BIC_\gamma(s) : \nu(s) = j, s \neq s_0 \} > BIC_\gamma(s_0) \right] \to 1,
\]

for \( j = 1, ..., K \), as \( n \to \infty \) (\( K \) is an upper bound for \( K_0 \))

- If \( \gamma = 1 \), consistent for \( \kappa > 0 \)
- If \( \gamma = 0 \), consistent for \( \kappa < 0.5 \) (original BIC)
Simulation Studies

- $\gamma = 0, 0.5$ and 1
- cannot afford to compute $\text{BIC}_\gamma(s)$ for all possible $s$
- LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001)
- Increase $\lambda$ gradually
- $\text{BIC}_\gamma$ computed for sequence of nested models
- If $P \gg n$, a tournament (Chen and Chen, 2008)
  1. randomly partition covariates (each: $n/2$ covariates)
  2. apply LASSO or SCAD
  3. pool nonzero components
  4. repeat if necessary
Simulation Studies

- Linear model in all cases,

\[ y_i = x_i^T (s) \beta(s) + \epsilon_i, \]

for some \( s \), where \( \epsilon_i \sim N(0, 1) \)

- Case 1: \( P = 50, n = 200 \)

  (i) \( \text{cor}(x_j, x_k) = \rho \) for all \( (j, k) \)

  (ii) \( \text{cor}(x_j, x_k) = \rho \) for \( k = j \pm 1 \)

  (iii) \( \text{cor}(x_j, x_k) = \rho^{\left| k-j \right|} \)

- \( \beta(s) = (7, 9, 4, 3, 10, 2, 2, 1)^T / 10 \)
Results of Case 1

- $s^*$: selected by the extended BIC

- Positive Selection Rate (PSR): $\nu(s \cap s^*)/\nu(s)$

- False Discovery Rate (FDR): $\nu(s^* - s)/\nu(s^*)$

Table 2. Case 1. Pooled outcomes

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0.2$</th>
<th></th>
<th>$\rho = 0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{BIC}_0$</td>
<td>$\text{BIC}_{0.5}$</td>
<td>$\text{BIC}_1$</td>
</tr>
<tr>
<td>PSR</td>
<td>0.495</td>
<td>0.490</td>
<td>0.478</td>
</tr>
<tr>
<td>FDR</td>
<td>0.190</td>
<td>0.073</td>
<td>0.034</td>
</tr>
</tbody>
</table>

PSR, pooled positive selection rates; FDR, false discovery rates.
Simulation Studies: Case 2

- $P = 1000, \ n = 200$
- 20 groups of size 50
- First 10 groups: generated as Case 1
- The other 10 groups: generated from a discrete distribution

$\beta(s) = (10, 7, 5, 3, 2)^T/10$

- The tournament approach
  1. 1000 covariates: partitioned into subsets of size 100
  2. From each subset, 12 covariates selected
  3. 120 → 20 (final candidate covariates)
### Results of Case 2

#### Table 3. Simulation results for Case 2 with $P = 1000$, with standard deviations in brackets

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>BIC$_0$</th>
<th>BIC$_{0.5}$</th>
<th>BIC$_1$</th>
<th>BIC$_0$</th>
<th>BIC$_{0.5}$</th>
<th>BIC$_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>c</td>
<td>4.05 (0.69)</td>
<td>3.72 (0.66)</td>
<td>3.45 (0.56)</td>
<td>2.83 (0.68)</td>
<td>2.34 (0.72)</td>
<td>1.82 (0.67)</td>
</tr>
<tr>
<td>ic</td>
<td>7.11 (2.43)</td>
<td>0.72 (1.01)</td>
<td>0.09 (0.30)</td>
<td>8.98 (2.36)</td>
<td>0.85 (1.21)</td>
<td>0.05 (0.24)</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>c</td>
<td>4.03 (0.70)</td>
<td>3.71 (0.67)</td>
<td>3.43 (0.63)</td>
<td>2.87 (0.78)</td>
<td>2.26 (0.72)</td>
<td>1.74 (0.66)</td>
</tr>
<tr>
<td>ic</td>
<td>6.47 (2.57)</td>
<td>0.76 (1.06)</td>
<td>0.16 (0.43)</td>
<td>8.72 (2.66)</td>
<td>1.00 (1.30)</td>
<td>0.10 (0.31)</td>
</tr>
</tbody>
</table>

**Correct (c) and incorrect (ic) selections**

#### Positive selection rates (PSR) and false discovery rates (FDR)

<table>
<thead>
<tr>
<th></th>
<th>PSR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.810</td>
<td>0.637</td>
</tr>
<tr>
<td>2</td>
<td>0.806</td>
<td>0.616</td>
</tr>
<tr>
<td>1</td>
<td>0.744</td>
<td>0.162</td>
</tr>
<tr>
<td>2</td>
<td>0.690</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>0.566</td>
<td>0.760</td>
</tr>
<tr>
<td></td>
<td>0.468</td>
<td>0.266</td>
</tr>
<tr>
<td></td>
<td>0.364</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>0.348</td>
<td>0.054</td>
</tr>
</tbody>
</table>
Simulation Studies: Case 3

- A dataset from a genome-wide association study
- \( y \): mRNA level of a particular gene
- \( X \): 1414 single-nucleotide polymorphisms (SNPs)
- \( n = 233 \)
- Setting 1. \( y \): randomly permuted
- Setting 2. \( y \): randomly generated under assumption of no association
- Setting 3. \( y \): generated from linear model
  \[
  \beta(s) = (-1.56, -1.09, 1.22, -0.06, -0.08, -0.012, 0.067, -0.047, -0.07, 0.05)^T
  \]
- Setting 4.
  \[
  \beta(s) = (-0.31, 0.23, 0.42, -0.32, -0.33, -0.26, 0.41, 0.29, -0.35, -0.69)^T
  \]
# Results of Case 3

## Table 4. Simulation results for Case 3, with standard deviations in brackets

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\text{BIC}_0$</th>
<th>$\text{BIC}_{0.5}$</th>
<th>$\text{BIC}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6.19 (3.08)</td>
<td>0.23 (0.50)</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>2</td>
<td>6.61 (2.63)</td>
<td>0.34 (0.59)</td>
<td>0.01 (0.10)</td>
</tr>
<tr>
<td>3</td>
<td>3.51 (2.17)</td>
<td>0.19 (0.45)</td>
<td>0.02 (0.14)</td>
</tr>
<tr>
<td>4</td>
<td>3.97 (1.54)</td>
<td>0.94 (0.99)</td>
<td>0.47 (0.73)</td>
</tr>
</tbody>
</table>

**Incorrect selection number**

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\text{BIC}_0$</th>
<th>$\text{BIC}_{0.5}$</th>
<th>$\text{BIC}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2.98 (0.14)</td>
<td>2.98 (0.14)</td>
<td>2.98 (0.14)</td>
</tr>
<tr>
<td>4</td>
<td>6.38 (1.39)</td>
<td>5.80 (1.73)</td>
<td>5.15 (2.02)</td>
</tr>
</tbody>
</table>

**False discovery rate**

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\text{BIC}_0$</th>
<th>$\text{BIC}_{0.5}$</th>
<th>$\text{BIC}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.541</td>
<td>0.060</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>0.384</td>
<td>0.139</td>
<td>0.084</td>
</tr>
</tbody>
</table>

**Positive selection rate**

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\text{BIC}_0$</th>
<th>$\text{BIC}_{0.5}$</th>
<th>$\text{BIC}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>4</td>
<td>0.638</td>
<td>0.580</td>
<td>0.515</td>
</tr>
</tbody>
</table>
Summary: BIC$\gamma$

- $\gamma = 1$
  - effectively controls FDR
  - consistent for $\kappa > 0$ ($p_n = O(n^\kappa)$)
- $\gamma = 0$ (original BIC)
  - a slightly better PSR, much worse FDR
  - consistent for $\kappa < 0.5$ ($p_n = O(n^\kappa)$)
- BIC$\gamma$ incur a small loss in PSR, but tightly control FDR
- Another way of choosing $\gamma$
  1. $P = n^\kappa$
  2. $\gamma = 1 - 1/(2\kappa)$