A Consistent Model Selection Criterion for $L_2$-Boosting in High-Dimensional Sparse Linear Models

Tze Leung Lai, Stanford University
Ching-Kang Ing, Academia Sinica, Taipei
Zehao Chen, Lehman Brothers
High Dimensional Regression

- \( y_{in} = \sum_{j=1}^{p_n} \beta_{jn} x_{ij}(n) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \)

- \( p_n = O(\exp(n^\xi)) \) for some \( 0 < \xi < 1. \)

The \( n \)'s in \( \beta_{jn}, y_{in} \) and \( x_{ij}(n) \) may be suppressed.

- Some typical examples for \( n \ll p: \)
  - Gene Expression Data: \( n = \) sample size, \( p = \) \# genes.
  - Sparse Signal Reconstruction: \( n = \) \# measurements,
    \[ p = \# \text{ signals} \]
  - Image Recovery: \( n = \) \# grid points,
    \[ p = \# \text{ basis functions} \]
  - Portfolio Selection: \( p = \# \text{ assets}, \) \( n = \# \text{ time points} \)
\textbf{L}_2\text{-Boosting: Pure Greedy Algorithm (PGA)}

\textit{Step 1.} Define $R_0 = (y_1, \cdots, y_n)'$ and $x_j = (x_{1j}, \ldots, x_{nj})'$. Find a variable among $\{x_1, \ldots, x_{pn}\}$ that is most correlated to $R_0$. Call the variable $x_{\hat{s}_1}$ and generate residual vector $R_1 = R_0 - \hat{\beta}_{\hat{s}_1} x_{\hat{s}_1}$, where $\hat{\beta}_{\hat{s}_1}$ is the least squares estimate obtained by regressing $R_0$ on $x_{\hat{s}_1}$. 
Step 2. Find a variable among \( \{x_1, \ldots, x_{pn}\} \) that is **most** correlated to \( R_1 \). Call the variable \( x_{\hat{s}_2} \) and let
\[
R_2 = R_1 - \hat{\beta}_{\hat{s}_2} x_{\hat{s}_2}.
\]

If iterations stop at step \( m \), then the new outcome, \( y_{n+1} \), is predicted by
\[
\hat{y}_{\hat{s}_1, \ldots, \hat{s}_m} = \hat{\beta}_{\hat{s}_1} x_{n+1, \hat{s}_1} + \cdots + \hat{\beta}_{\hat{s}_m} x_{n+1, \hat{s}_m}.
\]
L₂-Boosting: Orthogonal Greedy Algorithm (OGA)

Step 1. Define $y_n = (y_1, \ldots, y_n)'$ and $x_j = (x_{1j}, \ldots, x_{nj})'$. Find a variable among $\{x_1, \ldots, x_{pn}\}$ that is most correlated to $y_n$. Call the variable $x_{s_1}$ and generate residual $R_1 = y_n - M_{s_1} y_n$, where $M_{s_1}$ is the projection matrix into $L(x_{s_1})$. 
**Step 2.** Find a variable among \( \{ x_1, \ldots, x_{p_n} \} \) that is **most correlated** to \( R_1 \). Call the variable \( x_{\tilde{s}_2} \) and let

\[
R_2 = y_n - M_{\tilde{s}_1, \tilde{s}_2} y_n,
\]

where \( M_{\tilde{s}_1, \tilde{s}_2} \) is the projection matrix into \( L(x_{\tilde{s}_1}, x_{\tilde{s}_2}) \).

\[
\vdots
\]

If iterations stop at step \( m \), then the new outcome, \( y_{n+1} \), is predicted by

\[
\hat{y}_{\tilde{s}_1, \ldots, \tilde{s}_m} = \hat{\beta}_{\tilde{s}_1}^{o} x_{n+1, \tilde{s}_1} + \cdots + \hat{\beta}_{\tilde{s}_m}^{o} x_{n+1, \tilde{s}_m},
\]

where \( \hat{\beta}_{\tilde{s}_1}^{o}, \ldots, \hat{\beta}_{\tilde{s}_m}^{o} \) are the least squares estimates.
Convergence Rates in $L_2$-Boosting

Assumptions:

(K.1) $Ee^{t\epsilon_1} + \sup_j E \exp(sx_{1j}) < \infty$ for $|t| \leq t_0$, $|s| \leq s_0$.

(K.2) $0 < \iota_1 < \lambda_{\min}(\Gamma_n) \leq \lambda_{\max}(\Gamma_n) < \iota_2 < \infty$ for all $n$, where $\Gamma_n = E(x_1x_1')$

(K.2*) $0 < \eta_1 \leq \min_{1 \leq j \leq p_n} E(x_{1j}^2) \leq \max_{1 \leq j \leq p_n} E(x_{1j}^2) < \eta_2 < \infty$.

(K.3) $p_n = O(\exp(Cn^\xi))$ for some $0 < \xi < 1$ and $C > 0$.

(K.4) $\sup_{n \geq 1} \sum_{j=1}^{p_n} |\beta_j| < \infty$. 
\textbf{Theorem 1.} Assume (K.1), (K.2), (K.3) and (K.4). Then, for any choice of $m = m_n$ satisfying $m_n = O(n^l)$ with $0 < l < (1 - \xi)/5$, 

$$E \left\{ (y_{n+1} - \hat{\gamma}_{\hat{s}_1, \ldots, \hat{s}_m})^2 \mid y_n, x_1, \ldots, x_{p_n} \right\} - \sigma^2 = O_p(m^{-1}).$$

\textbf{Theorem 2.} Assume (K.1), (K.2\star), (K.3) and (K.4). Then, for any choice of $m = m_n$ satisfying $m_n = o(\log n)$ and $0 < \theta < \frac{1}{3}$, 

$$E \left\{ (y_{n+1} - \hat{\gamma}_{\hat{s}_1, \ldots, \hat{s}_m})^2 \mid y_n, x_1, \ldots, x_{p_n} \right\} - \sigma^2 = O_p(m^{-\theta}).$$
Theorem 3. Assume (K.1), (K.2), (K.3), (K.4) and (K.5) For some $0 \leq \gamma < (1 - \xi)/5$,

$$\liminf_{n \to \infty} n^\gamma \min_{j \in O_n} \beta_j^2 > C^* > 0.$$  

Then, for any choice of $m = m_n$ satisfying $m_n = O(n^l)$ with $0 < l < (1 - \xi)/5$ and $\lim_{n \to \infty} m_n/n^\gamma = \infty$,

$$\lim_{n \to \infty} P(D_n^o) = 1,$$

where

$O_n = \{j : 1 \leq j \leq p_n, \beta_j \neq 0\}, \quad D_n^o = \left\{O_n \subseteq \{\hat{s}_1^o, \ldots, \hat{s}_{m_n}^o\}\right\}$.  

Theorem 3 indicates that with probability approaching 1, the variables chosen by the OGA after $m_n$ steps contains all relevant variables. To prevent overfitting, the algorithm should be stopped at the first time when all relevant variables are included, i.e., choose the smallest correct model along the boosting path.
High-Dimensional Hannan–Quinn Criterion

\[ HQ(k) = \log \hat{\sigma}^2(x_{s_1}, \ldots, x_{s_k}) + \frac{(\log p_n) C_n k}{n} \]

\[ \hat{k}_{HQ} = \arg \min_{1 \leq k \leq m_n} HQ(k), \]

where \( C_n \to \infty \) and

\[ \hat{\sigma}^2(x_{s_1}, \ldots, x_{s_k}) = \frac{1}{n} y_n'(I - M_{s_1, \ldots, s_k}) y_n. \]
The major difference between HQ and conventional Hannan–Quinn is the additional factor \( \log p_n \), which is related to the maximum of i.i.d. \( V_1, \ldots, V_{p_n} \sim \chi^2(1) \):

\[
\frac{\max_{1 \leq k \leq p_n} V_i}{2 \log p_n} \rightarrow 1 \text{ in probability.}
\]

Define \( A_j = \{ \hat{s}^o_1, \ldots, \hat{s}^o_j \} \), \( j \geq 1 \),

\[
\tilde{k}^o_n = \begin{cases} 
m_n, & \text{if } O_n \subset A_{m_n} \\
 \min\{j : 1 \leq j \leq m_n, O_n \subset A_j\}, & \text{if } O_n \subset A_{m_n}.
\end{cases}
\]
Theorem 4. Assume (K.1), (K.2), (K.3), (K.4) and (K.5). Then, for any choice of \( m = m_n \) satisfying \( m_n = O(n^l) \) with \( 0 < l < (1 - \xi)/5 \) and \( \lim_{n \to \infty} m_n/n^\gamma = \infty \), and any choice of \( C_n \) satisfying \( C_n \log p_n = o(n^{1-2\gamma}) \) and \( n^\gamma/C_n = o(1) \),

\[
\lim_{n \to \infty} P(A_{\hat{k}_{HQ}} = A_{\tilde{k}_n}) = 1.
\]

- For \( \gamma = 0 \), one can take \( C_n = \log n \), giving the high-dimensional version of BIC.
Sketch of Proof

1. Difference between OGA (or PGA) with its population version: Exponential bounds.

2. The convergence rates in Theorems 1 and 2 are used to prove Theorem 3.

3. Theorem 3, exponential bounds and extension of Hannan–Quinn-type arguments to prove Theorem 4.
Finite-Sample Performance

Example 1. \((\beta_1, \ldots, \beta_5) = (3, -3.5, 4, -2.8, 3.2)\),

\[\beta_j = 0 \text{ for } j \geq 6 \text{ in }\]

\[y_i = \sum_{j=1}^{5} \beta_j x_{ij} + \sum_{j=6}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n.\]

• \(\varepsilon_i\) i.i.d. \(N(0, 1)\), \(x_i = (x_{i1}, \ldots, x_{ip})^T\) independent of \(\varepsilon_i\).

• \(x_{ij} = z_{ij} + \eta w_i\), where \(\eta > 0\), \((z_i^T, w_i)^T \sim N(0, I)\).

• 1000 simulation runs for each result.
**Table 1:** Frequency of all 5 relevant and \( i \) additional variables selected \((m_n = 30)\) in 1000 simulations.

| \( \eta \) | \( n \) | \( p \) | method          | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 9  | 25–30 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0  | 100 | 2000 | OGA+HDBIC | 1000 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1000 |
|    |     |      | OGA+BIC    | 0    | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
| 200| 4000| OGA+HDBIC | 1000 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
|    |     | OGA+BIC    | 0    | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
| 2  | 100 | 2000 | OGA+HDBIC | 992  | 7  | 1  | 0  | 0  | 0  | 0  | 0  | 1000 |
|    |     | OGA+BIC    | 0    | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
| 200| 4000| OGA+HDBIC | 1000 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
|    |     | OGA+BIC    | 0    | 0  | 0  | 0  | 0  | 0  | 1000 | 1000 |
Example 2. \( q = 10, \ n = 400, \ p = 4000 \) in

\[
y_i = \sum_{j=1}^{q} \beta_j x_{ij} + \sum_{j=q+1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n,
\]

- \((\beta_1, \ldots, \beta_{10}) = (3.2, 3.2, 3.2, 3.2, 4.4, 4.4, 3.5, 3.5, 3.5, 3.5)\),

- \( \beta_j = 0 \) for \( j > 10 \).

- \( \varepsilon_i \) i.i.d. \( N(0, 2.25) \).

- \( \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \), same as in Example 1.

- 100 simulations for each result.
Table 2: Frequency of all 9 relevant and $i$ additional variables selected ($m_n = 40$) in 100 simulations.

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>method</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>18</th>
<th>9</th>
<th>10–19</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OGA+HDBIC</td>
<td>97</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>35</td>
<td>25</td>
<td>12</td>
<td>6</td>
<td>11</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>OGA+HDBIC</td>
<td>64</td>
<td>29</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>LASSO</td>
<td>10</td>
<td>13</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>100</td>
</tr>
</tbody>
</table>
Example 3. Whereas Example 2 satisfies the neighborhood stability condition of Meinshausen and Buhlmann (2006), which is an extension of the irrepresentable condition of Zhao and Yu (2006) to ensure model selection consistency of LASSO to random covariates, this condition is violated in taking $q = 10$ in Example 2 and

- $(\beta_1, \ldots, \beta_q) = 
  (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75).$
- $\beta_j = 0$ for $j \geq 11$, $(x_{i1}, \ldots, x_{iq})^T$ i.i.d. $N(0, I_q)$.
- $\varepsilon_i$ i.i.d. $N(0, 1)$ and independent of $(x_{i1}, \ldots, x_{ip})^T$.
- $x_{ij} = z_{ij} + (\sum_{l=1}^{q} b x_{il})$ for $q < j \leq p$, $z_i \sim N(0, I_{p-q}/4)$.
- $qb^2 + 1/4 = 1$ (i.e, $b = \sqrt{3/4q}$).
**Table 3:** Frequency of all 10 relevant and \( i \) additional variables selected \((m_n = 40)\) in 100 simulations.

<table>
<thead>
<tr>
<th>method</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>30 or more</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGA+HDBIC</td>
<td>0</td>
<td>9</td>
<td>87</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>LASSO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Conclusion

- By obtaining the convergence rates of orthogonal greedy $L_2$-boosting (OGA), classical model selection criteria such as BIC can be modified to a large number $p_n$ of regressors in sparse regression models. Consistency of model selection is formulated in terms of capturing all, and only, variables with nonzero coefficients under (K5).
• OGA with consistent model selection means that the estimate is asymptotically equivalent to OLS with the “correct” set of covariates.

• In the absence of (K5), consistency of model selection can be linked to the objective of the regression model, e.g., prediction, estimation of high-dimensional covariance matrix via modified Cholesky decomposition. For the latter, “consistent” model selection can be carried out via thresholding (Bickel & Levina, 2008, and extensions).