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 β_{jn}, y_{in} and $x_{ij}(n)$ may be suppressed.

- Some typical examples for $n << p$:

  ◇ Gene Expression Data: n = sample size, $p = \#$ genes.

  ◇ Sparse Signal Reconstruction: $n = \#$ measurements,

  $p = \#$ signals

  ◇ Image Recovery: $n = \#$ grid points,

  $p = \#$ basis functions

  ◇ Portfolio Selection: $p = \#$ assets, $n = \#$ time points

L_2-Boosting: Pure Greedy Algorithm (PGA)

Step 1. Define $\boldsymbol{R}_0 = (y_1, \cdots, y_n)'$ and $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$. Find a variable among $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\}$ that is most correlated to $\boldsymbol{R}_0$. Call the variable $\boldsymbol{x}_{\hat{s}_1}$ and generate residual vector $\boldsymbol{R}_1 = \boldsymbol{R}_0 - \hat{\beta}_{\hat{s}_1}\boldsymbol{x}_{\hat{s}_1}$, where $\hat{\beta}_{\hat{s}_1}$ is the least squares estimate obtained by regressing $\boldsymbol{R}_0$ on $\boldsymbol{x}_{\hat{s}_1}$.

Step 2. Find a variable among $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\}$ that is most correlated to $\boldsymbol{R}_1$. Call the variable $\boldsymbol{x}_{\hat{s}_2}$ and let $\boldsymbol{R}_2 = \boldsymbol{R}_1 - \hat{\beta}_{\hat{s}_2}\boldsymbol{x}_{\hat{s}_2}$.

$$\vdots$$

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

$$\hat{y}_{\hat{s}_1,\ldots,\hat{s}_m} = \hat{\beta}_{\hat{s}_1} x_{n+1,\hat{s}_1} + \cdots + \hat{\beta}_{\hat{s}_m} x_{n+1,\hat{s}_m}.$$

L_2-Boosting: Orthogonal Greedy Algorithm (OGA)

Step 1. Define $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$ and $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$. Find a variable among $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\}$ that is most correlated to $\boldsymbol{y}_n$. Call the variable $\boldsymbol{x}_{\hat{s}_1^o}$ and generate residual $\boldsymbol{R}_1 = \boldsymbol{y}_n - \boldsymbol{M}_{\hat{s}_1^o}\boldsymbol{y}_n$, where $\boldsymbol{M}_{\hat{s}_1^o}$ is the projection matrix into $\mathcal{L}(\boldsymbol{x}_{\hat{s}_1^o})$.

Step 2. Find a variable among $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\}$ that is most correlated to $\boldsymbol{R}_1$. Call the variable $\boldsymbol{x}_{\hat{s}^o_2}$ and let $\boldsymbol{R}_2 = \boldsymbol{y}_n - \boldsymbol{M}_{\hat{s}^o_1, \hat{s}^o_2} \boldsymbol{y}_n$, where $\boldsymbol{M}_{\hat{s}^o_1, \hat{s}^o_2}$ is the projection matrix into $\mathcal{L}(\boldsymbol{x}_{\hat{s}^o_1}, \boldsymbol{x}_{\hat{s}^o_2})$.

$$\vdots$$

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

$$\hat{y}_{\hat{s}^o_1, \ldots, \hat{s}^o_m} = \hat{\beta}^o_{\hat{s}^o_1} x_{n+1, \hat{s}^o_1} + \cdots + \hat{\beta}^o_{\hat{s}^o_m} x_{n+1, \hat{s}^o_m},$$

where $\hat{\beta}^o_{\hat{s}^o_1}, \ldots, \hat{\beta}^o_{\hat{s}^o_m}$ are the least squares estimates.

Convergence Rates in L_2-Boosting

Assumptions:

(K.1) $Ee^{t\varepsilon_1} + \sup_j E\exp(sx_{1j}) < \infty$ for $|t| \le t_0$, $|s| \le s_0$.

(K.2) $0 < \iota_1 < \lambda_{\min}(\Gamma_n) \le \lambda_{\max}(\Gamma_n) < \iota_2 < \infty$ for all n,
where $\Gamma_n = E(\boldsymbol{x}_1\boldsymbol{x}_1')$

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

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

(K.4) $\sup_{n\ge1} \sum_{j=1}^{p_n} |\beta_j| < \infty$.

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* $0 < l < (1 - \xi)/5$,

$$E\left\{(y_{n+1} - \hat{y}_{\hat{s}_1^o,\ldots,\hat{s}_m^o})^2 | \boldsymbol{y}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\right\} - \sigma^2 = O_p(m^{-1}).$$

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

$$E\left\{(y_{n+1} - \hat{y}^P_{\hat{s}_1,\ldots,\hat{s}_m})^2 | \boldsymbol{y}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}\right\} - \sigma^2 = O_p(m^{-\theta}).$$

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

$$\mathrm{HQ}(k) = \log \hat{\sigma}^2(x_{\hat{s}_1^o}, \ldots, x_{\hat{s}_k^o}) + \frac{(\log p_n) C_n k}{n}$$

$$\hat{k}_{\mathrm{HQ}} = \arg \min_{1 \le k \le m_n} \mathrm{HQ}(k),$$

where $C_n \to \infty$ and

$$\hat{\sigma}^2(x_{\hat{s}_1^o}, \ldots, x_{\hat{s}_k^o}) = \frac{1}{n} \boldsymbol{y}_n' (\boldsymbol{I} - \boldsymbol{M}_{\hat{s}_1^o, \ldots, \hat{s}_k^o}) \boldsymbol{y}_n.$$

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. $V_1, \ldots, V_{p_n} \sim \chi^2(1)$:

$$\frac{\max_{1 \le k \le p_n} V_i}{2 \log p_n} \to 1 \text{ in probability.}$$

Define $A_j = \{\hat{s}_1^o, \ldots, \hat{s}_j^o\}$, $j \ge 1$,

$$\tilde{k}_n^o = \begin{cases} m_n, & \text{if } O_n \subsetneq A_{m_n} \\ \min\{j : 1 \le j \le m_n,\ O_n \subseteq A_j\}, & \text{if } O_n \subseteq A_{m_n}. \end{cases}$$

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* $n^\gamma / C_n = o(1)$,

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

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

$\beta_j = 0$ for $j \geq 6$ in

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

- ε_i i.i.d. $\mathcal{N}(0,1)$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ independent of ε_i.
- $x_{ij} = z_{ij} + \eta w_i$, where $\eta > 0$, $(\boldsymbol{z}_i^T, w_i)^T \sim N(\boldsymbol{0}, \boldsymbol{I})$.
- 1000 simulation runs for each result.

Table 1: Frequency of all 5 relevant and i additional variables selected ($m_n = 30$) in 1000 simulations.

				i									
η	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	0	1000
			OGA+BIC	0	0	0	0	0	0	0	0	1000	1000
	200	4000	OGA+HDBIC	1000	0	0	0	0	0	0	0	0	1000
			OGA+BIC	0	0	0	0	0	0	0	0	1000	1000
2	100	2000	OGA+HDBIC	992	7	1	0	0	0	0	0	0	1000
			OGA+BIC	0	0	0	0	0	0	0	0	1000	1000
	200	4000	OGA+HDBIC	1000	0	0	0	0	0	0	0	0	1000
			OGA+BIC	0	0	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 $j > 10$.
- ε_i i.i.d. $N(0, 2.25)$.
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, 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.

		i											
η	method	0	1	2	3	4	5	6	7	18	9	10–19	Total
1	OGA+HDBIC	97	3	0	0	0	0	0	0	0	0	0	100
	LASSO	35	25	12	6	11	4	2	2	0	1	2	100
3	OGA+HDBIC	64	29	6	1	0	0	0	0	0	0	0	100
	LASSO	10	13	14	10	11	14	7	7	0	3	11	100

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. $N(\mathbf{0}, \boldsymbol{I}_q)$.
- ε_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 $q < j \leq p$, $\mathbf{z}_i \sim N(\mathbf{0}, \boldsymbol{I}_{p-q}/4)$.
- $qb^2 + 1/4 = 1$ (i.e, $b = \sqrt{3/4q}$).

Table 3: Frequency of all 10 relevant and i additional variables selected ($m_n = 40$) in 100 simulations.

	i									
method	0	1	2	3	4	5	6	7	30 or more	Total
OGA+HDBIC	0	9	87	4	0	0	0	0	0	100
LASSO	0	0	0	0	0	0	0	0	100	100

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).