Random Weighting in LASSO Regression

JSM 2021 (Session ID: 220740)

Tun Lee Ng

Email: tng25@wisc.edu

Statistics PhD candidate, University of Wisconsin-Madison

Advisor: Prof. Michael A. Newton

8th August, 2021

Motivation

Consider the linear model

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

- Bayesian: Gaussian likelihood with Double Exponential Prior under the Bayesian LASSO (Park and Casella, 2008)

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp\left\{-\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - \lambda_n \sum_{j=1}^{p_n} |\beta_j|\right\}.$$

- Deploy MCMC to draw B posterior samples from $p(\boldsymbol{\beta}|\boldsymbol{y})$:

$$\left\{\boldsymbol{\beta}_{(1)}^{\text{MCMC}}, \cdots, \boldsymbol{\beta}_{(B)}^{\text{MCMC}}\right\}$$

- Other methods? 1 option: Weighted Bootstrap / **Random-weighting**

Random-weighting:

Newton et al. (2020):

- Assign random weights $W_i \overset{iid}{\sim} F_w$ on negative log posterior density $-\log p(\boldsymbol{\beta}|\boldsymbol{y})$ and optimize

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\}$$

  to get one solution $\widehat{\boldsymbol{\beta}}^w_{n(1)}$.

- Repeat: draw another set of random weights $W_i \overset{iid}{\sim} F_w$, assign it onto the objective function above, and optimize. Now we have 2 random-weighting (RW) samples $\left\{ \widehat{\boldsymbol{\beta}}^w_{n(1)}, \widehat{\boldsymbol{\beta}}^w_{n(2)} \right\}$.
- Repeat this B times, and obtain B random-weighting samples:

$$\left\{ \widehat{\boldsymbol{\beta}}^w_{n(1)}, \cdots, \widehat{\boldsymbol{\beta}}^w_{n(B)} \right\}.$$

* More on $W_{0,j}$ later.

Random-weighting:

- Assign random weights $W_i \overset{iid}{\sim} F_w$ on negative log posterior density $-\log p(\boldsymbol{\beta}|\boldsymbol{y})$ and optimize with a **2-step procedure** (more on this later)

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\}$$

  to get one solution $\widehat{\boldsymbol{\beta}}^w_{n(1)}$.

- Repeat: draw another set of random weights $W_i \overset{iid}{\sim} F_w$, assign it onto the objective function above, and optimize with the **2-step procedure**. Now we have 2 random-weighting (RW) samples $\left\{ \widehat{\boldsymbol{\beta}}^w_{n(1)}, \widehat{\boldsymbol{\beta}}^w_{n(2)} \right\}$.

- Repeat this B times, and obtain B random-weighting samples:

$$\left\{ \widehat{\boldsymbol{\beta}}^w_{n(1)}, \cdots, \widehat{\boldsymbol{\beta}}^w_{n(B)} \right\}.$$

* More on $W_{0,j}$ later.

The **two-step procedure**

- Step 1: Assign random weights, optimize

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\}$$

  to select variables. Let $\widehat{S}_n^w \subseteq \{1, \cdots, p_n\}$ be the set of selected variables.

- Step 2: Estimate the selected variables with weighted least squares (same random weights from Step 1).

$$\widehat{\boldsymbol{\beta}}_n^w := \begin{bmatrix} \widehat{\boldsymbol{\beta}}^w_{n,\widehat{S}_n^w} \\ \widehat{\boldsymbol{\beta}}^w_{n,(\widehat{S}_n^w)^c} \end{bmatrix} := \begin{bmatrix} \left(X'_{\widehat{S}_n^w} D_n X_{\widehat{S}_n^w} \right)^{-1} X'_{\widehat{S}_n^w} D_n Y \\ \mathbf{0} \end{bmatrix}, \qquad (1)$$

  where $D_n = diag\,(W_1, \cdots, W_n)$, and $X_{\widehat{S}_n^w}$ is the $n \times |\widehat{S}_n^w|$ submatrix of X.

Model Setup

- $\{\epsilon_i\}$ iid with mean 0, variance σ_ϵ^2, and finite 4^{th} moment. Unlike Bayesian LASSO, no normality assumption needed for our work.
- F_W: $P(W > 0) = 1$ and $\mathbb{E}(W^4) < \infty$ and $\mathbb{E}(W) = 1$.
- Other common linear-model-related regularity assumptions.

3 weighting schemes on the penalty terms are considered (Newton et al., 2020):

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p_n} W_{0,j} |\beta_j| \right\}$$

- `RW1`: $W_{0,j} = 1 \ \forall \ j,$
- `RW2`: $W_{0,j} = W_0 \ \forall \ j,$ where $(W_0, W_i) \overset{iid}{\sim} F_W \ \forall \ i,$
- `RW3`: $(W_{0,j}, W_i) \overset{iid}{\sim} F_W \ \forall \ i, j.$

Theorem 3.4 (Conditional Sparse Normality)

Assume the **strong irrepresentable condition** (Zhao and Yu, 2006)

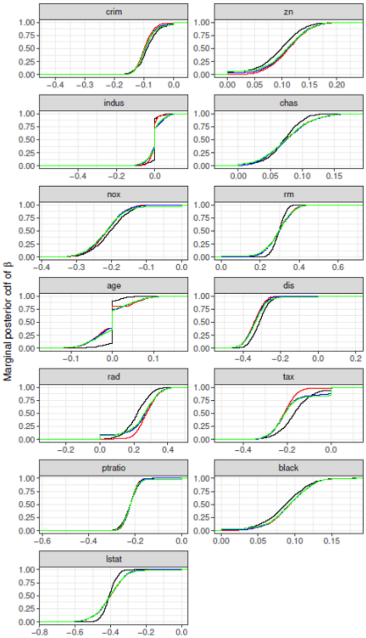
$$\left| C_{n(21)} \left(C_{n(11)} \right)^{-1} \operatorname{sgn} \left(\boldsymbol{\beta}_{0(1)} \right) \right| \leq \mathbf{1} - \boldsymbol{\eta},$$

where inequality holds element-wise for $0 < \eta_j \leq 1 \; \forall \; j = 1, \cdots, p_n - q$. Furthermore, assume proper growth rates of $\lambda_n = \mathcal{O}\left(n^{c_1}\right)$ and $p_n = \mathcal{O}\left(n^{c_2}\right)$, where $0 \leq c_2 < 1/2 < c_1 < 1$, and other regularity assumptions. Let $\widehat{\boldsymbol{\beta}}_n^w$ be the two-step random-weighting samples defined in (1), and let $\widehat{\boldsymbol{\beta}}_n^{LAS+LS}$ be the unweighted two-step LASSO+LS estimator (i.e. a LASSO variable selection step followed by least-squares estimation for the selected variables). Then, for all `RW1`, `RW2` and `RW3`, as $n \to \infty$,

$$P\left(\widehat{S}_n^w = S_0 \middle| \mathcal{F}_n \right) \to 1 \quad a.s.\ P_D,$$

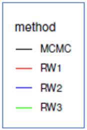
and

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{n(1)}^w - \widehat{\boldsymbol{\beta}}_{n(1)}^{LAS+LS} \right) \xrightarrow{\text{c.d.}} N_q\left(\mathbf{0}\ ,\ \sigma_W^2 \sigma_\epsilon^2 C_{11}^{-1} \right) \quad a.s.\ P_D.$$



Benchmark Dataset: Boston Housing

Abbreviation	Variable
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherw
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per \$10,000
ptratio	pupil-teacher ratio by town
black	proportion of blacks by town
lstat	lower status of the population (percent)



References

Newton, M., Polson, N. G., and Xu, J. (2020), "Weighted Bayesian Bootstrap for Scalable Posterior Distributions," *The Canadian Journal of Statistics*, 49, 421–437, https://doi.org/10.1002/cjs.11570.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Our manuscript for this presentation

Ng, T.L. and Newton, M.A. (2020), "Random-weighting in LASSO Regression," *arXiv:2002.02629*. In revision at *the Electronic Journal of Statistics*.