Random-weighting in LASSO Regression

Institute for Foundations of Data Science (IFDS) Seminar

Tun Lee Ng   Michael A. Newton

13th May, 2019

Outline

Motivation

- data : $\boldsymbol{y} = (y_1, \ldots, y_n)$
- parameters : $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$
- Bayes : $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

MCMC:

- works well for moderate-sized models
- computationally intensive & mixing hard to verify for large models

Optimization:

- Efficient algorithms available
- Optimization feasible in many models, eg. MAP estimation

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$$

Q: Any alternative method to obtain posterior samples?

A: How about random-weighting (Newton et al., 2019)

Motivation

Example 1: Diabetes Study (Park and Casella, 2008)

Table 1 shows a small part of the data for our main example.

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. A prediction model was desired for the response variable, a measure of disease progression one year after baseline.

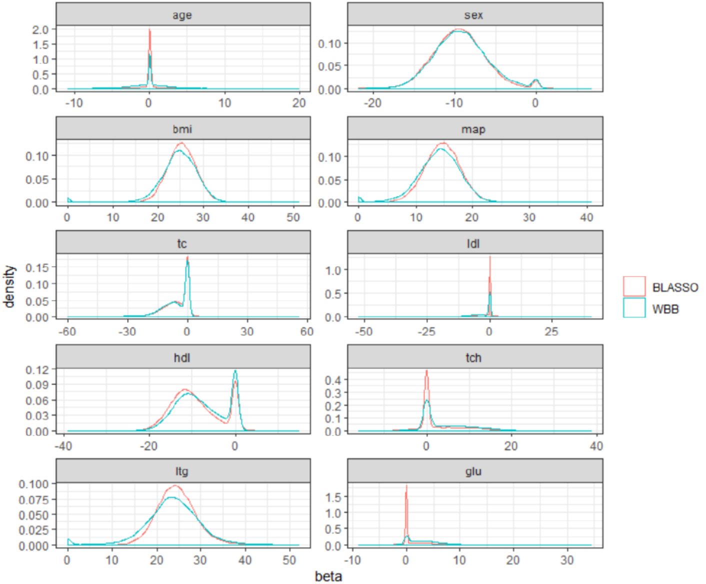
Random-weighting with LASSO regression (Newton et al., 2019):

For $j = 1, \ldots, B$,

1. Draw random weights $W_{j1}, \ldots, W_{jn} \overset{iid}{\sim} Exp(1)$.
2. Solve $\widehat{\boldsymbol{\beta}}^w_{n(j)} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n W_{ji}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^p |\beta_j| \right\}$.

Motivation

Example 1: Diabetes Study (Park and Casella, 2008); Random-weighting (Newton et al., 2019)



Outline

Setup

$$Y = X\beta + \epsilon$$

- $\{\epsilon_i\}$ iid with mean 0, variance σ_ϵ^2, and finite 4^{th} moment.
- All predictors are bounded.
- $\beta = (\beta_1, \ldots, \beta_p)'$ sparse.
- WLOG, partition $\boldsymbol{\beta}_0' = \left[\boldsymbol{\beta}_{0(1)}' \ \boldsymbol{\beta}_{0(2)}'\right]$, where
  - $\boldsymbol{\beta}_{0(1)}$ is $q \times 1$ vector of true non-zero regression parameters,
  - $\boldsymbol{\beta}_{0(2)}$ is $(p-q) \times 1$ vector of zeroes.
- q is fixed.
- Correspondingly, partition $X = [X_{(1)} \ X_{(2)}]$, and denote

$$C_{n(11)} = \frac{1}{n} X_{(1)}' X_{(1)} \qquad \text{and} \qquad C_{n(21)} = \frac{1}{n} X_{(2)}' X_{(1)}$$

Main Results

$$\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^p |\beta_j| \right\}, \quad \text{where} \quad W_i \overset{iid}{\sim} Exp(1)$$

Conditional on data, $\widehat{\boldsymbol{\beta}}_n^w$ has the following properties:

- conditional consistency (for fixed p)
- conditional asymptotic normality (for fixed p)
- conditional model selection consistency (for both fixed p and growing p_n).

Main Results

Theorem 1

p is fixed. Assume $\frac{1}{n}X'X \to C$ for some non-singular C.

(a) **(Conditional Consistency)** If $\dfrac{\lambda_n}{n} \to 0$, then

$$\widehat{\boldsymbol{\beta}}_n^w \xrightarrow{\text{c.p.}} \boldsymbol{\beta}_0 \quad a.s.\ P_D.$$

(b) If $\dfrac{\lambda_n}{n} \to \lambda_0 \in (0, \infty)$, then

$$\left(\widehat{\boldsymbol{\beta}}_n^w - \boldsymbol{\beta}_0\right) \xrightarrow{\text{c.p.}} \arg\min_{\boldsymbol{u}} g(\boldsymbol{u}) \quad a.s.\ P_D,$$

where

$$g(\boldsymbol{u}) = \boldsymbol{u}'C\boldsymbol{u} + \lambda_0\|\boldsymbol{\beta}_0 + \boldsymbol{u}\|_1.$$

Main Results

Theorem 2 (Asymptotic Conditional Distribution)

p is fixed. Assume $\frac{1}{n}X'X \to C$ for some non-singular C. If $\frac{\lambda_n}{\sqrt{n}} \to 0$, then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^w - \widehat{\boldsymbol{\beta}}_n^{\text{OLS}}\right) \xrightarrow{\text{c.d.}} N\left(\mathbf{0}, \sigma_\epsilon^2 C^{-1}\right) \quad a.s.\ P_D,$$

where $\widehat{\boldsymbol{\beta}}_n^{\text{OLS}}$ is the ordinary least squares estimator of $\boldsymbol{\beta}$ in the linear model.

Main Results

Theorem 3 (Posterior Model Selection Consistency) – fixed p

Assume $\frac{1}{n}X'X \to C$ for some non-singular C, and 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, and $0 < \eta_j \leq 1 \;\forall\; j = 1, \cdots, p-q$. Then, for any $\frac{1}{2} < c_1, c_2 < 1$ such that $c_1 + c_2 < 1.5$, and for all λ_n that satisfies

$$\frac{\lambda_n}{n^{c_2}} \to \infty \qquad \text{but} \qquad \frac{\lambda_n}{n^{1.5-c_1}} \to 0,$$

we have

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \stackrel{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n \right) \to 1 \quad a.s.\ P_D.$$

Main Results

Theorem 4 (Posterior Model Selection Consistency) – growing p_n

Assume the **strong irrepresentable condition** (Zhao and Yu, 2006), and that for some $M_2 > 0$,

$$\boldsymbol{\alpha}' C_{n(11)} \boldsymbol{\alpha} \geq M_2 \quad \forall \quad \|\boldsymbol{\alpha}\|_2 = 1.$$

For any $0 < c_3 < \frac{1}{2} < c_1, c_2 < 1$ such that $c_3 < 2\min(c_1, c_2) - 1$ and $c_1 + c_2 < 1.5$, for which $p_n = \mathcal{O}\left(n^{c_3}\right)$ and $\lambda_n = \mathcal{O}\left(n^{c_2}\right)$, we have

$$P\left(\widehat{\boldsymbol{\beta}}_n^w(\lambda_n) \stackrel{s}{=} \boldsymbol{\beta}_0 \middle| \mathcal{F}_n\right) \to 1 \quad a.s.\ P_D.$$

Connection to Bayesian Approach

- If $\lambda_n = o(\sqrt{n})$, first-order approximation to Bayesian samples.
- If $\lambda_n = \mathcal{O}(n^c)$ for $\frac{1}{2} < c < 1$, posterior model selection consistency.

Outline

Non-Standard-Exponential Weights

Q: What if W_i is any positive r.v. with $\mu_W = \sigma_W^2 = 1$ and $\mathbb{E}(W_i^4) < \infty$?
A: Same asymptotic results in Theorems 1 - 3.

Q: What if μ_W and/or σ_W^2 not equal to 1?
A: Nice properties like Theorems 1 - 4, with asymptotics scaled accordingly with μ_W and σ_W^2.

Weighting the Penalty Term

Q: What if

1. $\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n W_0 \sum_{j=1}^p |\beta_j| \right\}$
2. $\widehat{\boldsymbol{\beta}}_n^w = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n W_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^p W_j |\beta_j| \right\}$

A: Nice properties like Theorems 1 - 4, with penalty terms in asymptotics weighted accordingly.

Outline

1. Motivation
2. Main Results
3. Other Issues
   - Non-Standard-Exponential Weights
   - Weighting the Penalty Term
4. Future Work

Future Work

- Growing p_n?
- Other likelihood and/or penalty structure?

References

Newton, M., Polson, N. G., and Xu, J. (2019), "Weighted Bayesian Bootstrap for Scalable Bayes," *revision in press at the Canadian Journal of Statistics*.

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.