Some Bayesian Lasso Perspectives

Lei Yang

Bayesian interpretation of lasso

The form of lasso

$$\hat{\beta}_{lasso}(\lambda) = \underset{\beta}{\text{argmin}} \, ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda \, ||\beta||_1$$

May be interpreted as a Bayesian posterior mode estimate when β_i have independent and identical double exponential (Laplace) priors.

$$\hat{\beta}_{lasso}(\lambda) = \arg\max_{\beta} p(\beta | y, \sigma^2, \tau)$$

where $p(\beta|\tau) = (\tau/2)^p \exp(-\tau \, ||\beta||_1)$ and $\lambda = 2\tau\sigma^2$ [1].

Mode is not natural in Bayesian perspective

1. mode is not preserved under marginalisation. Mean is in the sense that $E(\beta|\sigma^2, \tau) = E_y(E(\beta|y, \sigma^2, \tau))$ [3].
2. mode is not a common Bayes rule (mean, median are), though it's the limit of of a sequence of Bayes rules [1].
3. the mode of posterior predictive distribution $\arg\max_{\tilde{y}} p(\tilde{y}|y, \sigma^2, \tau)$ is not the lasso prediction $\hat{y} = \tilde{X}\hat{\beta}_{lasso}$. (The posterior mean facilitates both point estimation and prediction. $E(\tilde{y}|y, \sigma^2, \tau) = \tilde{X}E(\beta|y, \sigma^2, \tau))$ [1].

Model Selection in Bayesian lasso

The usual lasso variable selection property is *ad hoc* from a Bayesian perspective. The posterior probability of $\beta_i = 0$ is 0.

In the Bayesian setting. Let γ be a p-vector where $\gamma_j = 1$ if predictor variable x_j is included in the regression model and $\gamma_j = 0$ otherwise. If the prior probability of a particular model is $\pi(\gamma)$, the posterior probability $\pi(\gamma|y) = \frac{m_\gamma(y)\pi(\gamma)}{\sum_{\gamma' \in \mathcal{T}} m_{\gamma'}(y)\pi(\gamma')}$ is often used to identify variables.

Where $\mathcal{T}$ is the collection of all possible models and $m_\gamma(y) = \int N(y|X_\gamma\beta_\gamma, \sigma^2 I_n)\pi(\beta_\gamma, \sigma^2|\gamma) d\beta_\gamma d\sigma^2$, the marginal likelihood of the observed data under model γ. [2].

Bayesian implementation 1

Exploits the representation of double exponential distribution as a scale mixture of normals [3]:

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds$$

hierarchical representation:

$$\begin{aligned} y|\mu, X, \beta, \sigma^2 &\sim N_n(\mu 1_n + X\beta, \sigma^2 I_n) \\ \beta|\tau_1^2, ..., \tau_p^2, \sigma^2 &\sim N_p(0_p, \sigma D_r), D_r = diag(\tau_1^2, ..., \tau_p^2) \\ \tau_1^2, ..., \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \tau_1^2, ..., \tau_p^2 > 0 \\ \sigma^2 &\sim \pi(\sigma^2) d\sigma^2 \end{aligned}$$

Bayesian implementation 2: Univariable case

Or use direct characterization [1], the kernel of univariate posterior:

$$(y - x\beta)^2 + \lambda|\beta| = x^2\beta^2 - 2x\beta + \lambda|\beta| + y^2$$

When $\beta > 0$, it's:

$$x^2\beta^2 - (2x - \lambda)\beta + y^2$$

When $\beta < 0$, it's:

$$x^2\beta^2 - (2x + \lambda)\beta + y^2$$

Then positive part and negative part of β are both normal distributed.

Bayesian implementation 2

Multivariable case:

$$N^{[z]}(\beta|m,S) \equiv \frac{N(\beta|m,S)}{P(z,m,S)1(\beta \in O_z)}, P(z,m,S) = \int_{O_z} N(t|m,S)dt$$

$$p(\beta|y,\sigma^2,\tau) = \sum_{z \in Z} w_z N^{[z]}(\beta|\mu_z,\Sigma)$$

where $\mu_z = \hat{\beta}_{OLS} - \tau\sigma^{-1}\Sigma z$ and

$$w_z = \left\{ \frac{P(z,\mu_z,\Sigma)}{N(0|\mu_z,\Sigma)} \right\} \Bigg/ \left\{ \sum_{z \in Z} \frac{P(z,\mu_z,\Sigma)}{N(0|\mu_z,\Sigma)} \right\}$$

Results

1. Provide Bayesian credible intervals that can guide variable selection [3].
2. Model-based Bayesian predictions for the two examples performed on average, as well as or better than the lasso predictions [1].
3. Variable inclusion probability instead of in-or-out rule can be provided[2].

References

[1] Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

[2] Chris Hans. Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20(2):221–229, 2010.

[3] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.