Bayesian Linear Regression

Lesi Chen

School of Data Science, Fudan University

2022.5.10

1. Choices of priors
2. Variable selection
3. Bayesian regression workflow

Problem set-up

We focus on the following Bayesian regression problem:

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n),$$

where the dataset is given by $y^{n\times 1}$ and $X^{n\times p}$.

Semiconjugate prior for Bayesian regression

In the course we have discussed the semiconjugate prior:

$$\beta \sim \mathcal{N}(\mu_0, \Lambda_0^{-1})$$
$$\sigma^2 \sim \mathrm{IG}(a_0/2, b_0/2),$$

say, the prior of σ^2 satisfies inverse Gamma distribution, while β satisfies multivariable normal distribution whose parameters is independent on σ^2.

Can we obtain a conjugate prior when β is dependent on σ^2?

Conjugate prior for Bayesian regression

If we let

$$\sigma^2 \sim \mathrm{IG}(a_0/2, b_0/2)$$
$$\beta \mid \sigma^2 \sim \mathcal{N}(\mu_n, \sigma^2 \Lambda_0^{-1}).$$

Then we can show that

$$\sigma^2 \mid X, y \sim \mathrm{IG}(a_n/2, b_n/2)$$
$$\beta \mid X, y, \sigma^2 \sim \mathcal{N}(\mu_n, \sigma^2 \Lambda_n^{-1}).$$

(continue on next slide)

Conjugate and informative prior for Bayesian regression

The parameters of posterior distribution are as follows [4]:

$$
\begin{aligned}
\mu_n &= (X^\top X + \Lambda_0)^{-1}(\Lambda_0 \mu_0 + X^\top X \hat{\beta}) \\
\Lambda_n &= X^\top X + \Lambda_0 \\
a_n &= a_0 + n \\
b_n &= b_0 + (y^\top y + \mu_0^\top \Lambda_0 \mu_0 - \mu_n^\top \Lambda_n \mu_n).
\end{aligned}
$$

where $\hat{\beta}$ is the least square estimator.

We may use $\pi(\beta, \sigma^2) \propto 1/\sigma^2$ as a non-informative prior, which also preserve conjugation.

1. Choices of priors
2. Variable selection
3. Bayesian regression workflow

Bayesian variable selection

In this section, I will introduce two representative Bayesian variable selection methods:

1. Stochastic Search Variable Selection (SSVS) [2]
2. Bayesian Lasso [1, 3]

The role of latent variables in SSVS

Consider the following mixture model:

$$\beta_j \mid \gamma_j \sim (1-\gamma_j)\mathcal{N}(0, \tau_j^2) + \gamma_j \mathcal{N}(0, c_j^2 \tau_j^2),$$

where $\gamma_j = \{0, 1\}$ are the latent variables.

If we let τ_j^2 small but $c_j^2 \tau_j^2$ large, then γ_j selects the significant coefficients for us.

Gibbs sampler for SSVS

Conditional on γ_j, the full conditional distributions of σ^2, β follows from standard conclusions. For γ_j, we let

$$\gamma_j \sim \text{Bernoulli}(p_j).$$

Then by the Bayesian rule,

$$\begin{aligned}
\mathbb{P}(\gamma_j = 1 \mid \beta_j) \propto \mathbb{P}(\beta_j \mid \gamma_j = 1)\mathbb{P}(\gamma_j = 1) \\
\mathbb{P}(\gamma_j = 0 \mid \beta_j) \propto \mathbb{P}(\beta_j \mid \gamma_j = 0)\mathbb{P}(\gamma_j = 0).
\end{aligned}$$

Hence, we can easily maintain the full conditional distribution of γ_j.

A hierarchical model for SSVS

The distribution of p_j does matter. We can treat p_j as a parameter to estimate. Let

$$\begin{aligned} \gamma_j &\sim \text{Bernoulli}(p_j) \\ p_j &\sim \text{Beta}(a_j, b_j). \end{aligned}$$

The we can obtain

$$p_j \mid \gamma_j \sim \text{Beta}(a_j + \gamma_j, b_j + 1 - \gamma_j),$$

which may be of help when selecting variables.

Main advantage of SSVS

SSVS uses latent variables to identify the most promising subsets, avoiding the overwhelming problem of calculating the posterior probabilities of all 2^p subsets.

The Bayesian Lasso

Lasso problem can be viewed regression with Laplace prior:

$$\pi(\beta \mid \sigma^2, \lambda) = \prod_{j=1}^{k} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right),$$

where λ serves as the shrinkage parameter.

How to obtain conjugation for Bayesian Lasso?

Gibbs sampler for Bayesian Lasso

A key observation is that Laplace density is a scale mixture of normal distributions:

$$\begin{aligned}
\pi\left(\beta \mid \sigma^2, \lambda\right) &= \prod_{j=1}^{k} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{-\frac{\lambda\left|\beta_j\right|}{\sqrt{\sigma^2}}\right\} \\
&= \prod_{j=1}^{k} \int_0^{\infty} \frac{1}{\sqrt{2\pi\sigma^2 s_j}} \exp\left(-\frac{\beta_j^2}{2\sigma^2 s_j}\right) \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2} s_j\right) \mathrm{d}s_j \\
&= \prod_{j=1}^{k} \int_0^{\infty} \pi(\beta_j \mid s_j) \times \pi(s_j) \mathrm{d}s_j .
\end{aligned}$$

Gibbs sampler for Bayesian Lasso

After introducing auxiliary variables $\{s_j\}_{j=1}^k$, if we let

$$\begin{aligned}\sigma^2 &\sim \mathrm{IG}(a_0/2, b_0/2)\\ \beta \mid \sigma^2, \{s_j\}_{j=0}^k &\sim \mathcal{N}(0, \sigma^2\Lambda_0^{-1}),\end{aligned}$$

where $\Lambda_0 = \mathrm{diag}\{1/s_1, \cdots, 1/s_k\}$. Then,

$$\begin{aligned}\sigma^2 \mid X, y, \{s_j\}_{j=1}^k &\sim \mathrm{IG}(a_n/2, b_n/2)\\ \beta \mid X, y, \sigma^2, \{s_j\}_{j=1}^k &\sim \mathcal{N}(\mu_n, \sigma^2\Lambda_n^{-1}).\end{aligned}$$

(continue on next slide)

Gibbs sampler for Bayesian Lasso

The parameters are given by [1, 3]:

$$
\begin{aligned}
\mu_n &= (X^\top X + \Lambda_0)^{-1} X^\top y \\
\Lambda_n &= X^\top X + \Lambda_0 \\
a_n &= a_0 + p + n \\
b_n &= b_0 + (y - X\beta)^\top (y - X\beta) + \beta^\top \Lambda_0 \beta.
\end{aligned}
$$

(See references for proof details)

Gibbs sampler for Bayesian Lasso

Note that the prior of each s_j is given by:

$$s_j \sim \text{Exponential}(\lambda^2/2).$$

We can derive the conditional distribution, which is

$$1/s_j \mid \sigma^2, \beta_j \sim \text{InverseGaussian}(\mu', \lambda'),$$

where $\mu' = \sqrt{\lambda\sigma^2/\beta_j^2}$ and $\lambda' = \lambda^2$.

Advantages of Bayesian Lasso

Compared to the frequent Lasso, the Bayesian Lasso

1. is easy to implement
2. automatically provides interval estimates for parameters
3. enable us to integrate prior beliefs

A brief summary

For now, we have the following options:

Prior	Solver	Additional Usage
Semiconjugate	Gibbs sampler	
Conjugate	Close form	
Non-informative	Close form	
Mixture	Gibbs sampler	Variable selection
Laplace	Gibbs sampler	Variable selection

Step 1: Description of the data

We use the abalone dataset from LIBSVM [1], the features are scaled to $[-1, 1]$. In this dataset, $n = 4177$, $p = 8$.

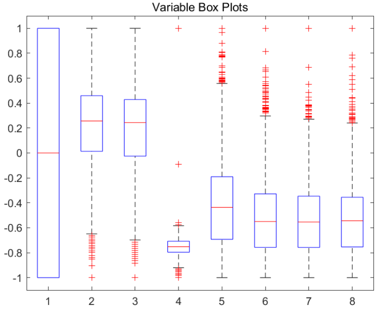


Figure 1: Features of abalone dataset

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools

Step 2: Choose a model

See Slide 19 for different models and priors for details. They include non-hierarchical and hierarchical models and two methods for variable selection (Bayesian Lasso and SSVS). Both informative and weakly informative priors can be applied.

Step 3: Choose a prior

We display the mixture prior for SSVS.

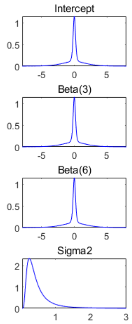
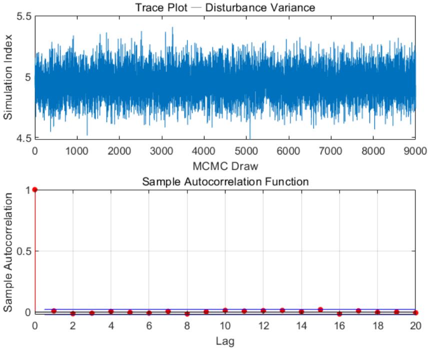


Figure 2: Prior of SSVS

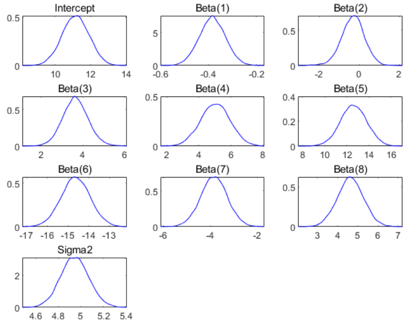
Step 4: Convergence diagnostics

We use Gibbs for both SSVS and Bayesian Lasso. We iterate 10000 times and discard the first 1000 samples. For the 9000 samples estimating σ^2 in SSVS, the effective sample size is 8840.



Step 5: Posterior predictive checking

We first display the mixture posterior for SSVS.



(continue on next slide)

Step 5: Posterior predictive checking

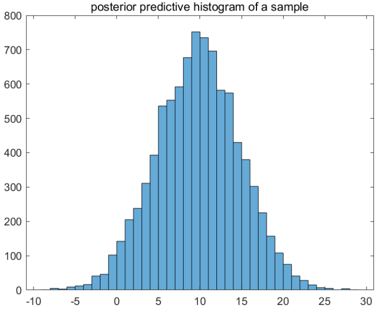
Since we have the samples of β_n, σ_n^2. We can directly use them to form an approximation of posterior predictive distribution for $\tilde{y}$ by Monte-Carlo method:

$$\begin{aligned} p(\tilde{y}) &= \int p(\tilde{y}, \beta_n, \sigma_n^2) \mathrm{d}\beta_n \mathrm{d}\sigma_n^2 \\ &\propto \int p(\tilde{y} \mid \beta_n, \sigma_n^2) p(\beta_n, \sigma_n^2) \mathrm{d}\beta_n \mathrm{d}\sigma_n^2. \end{aligned}$$

(continue on next slide)

Step 5: Posterior predictive checking

For instance, we give prediction to the first sample in the testing data whose ground truth is 9.



Step 6: Model comparison

Since we use non-informative or weak informative priors, the regression results are data-driven. We measured above methods by

$$\text{FMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{real}}^{(i)} - y_{\text{pred}}^{(i)})^2}.$$

(continue on next slide)

Step 6: Model comparison

We can see that data dominates in this example.

Prior	FMSE
Semiconjugate	2.222
Conjugate	2.222
Non-infomative	2.222
Mixture	2.222
Laplace	2.222

Table 1: FMSE on training data for different methods

Step 6: Model comparison (variable selection with SSVS)

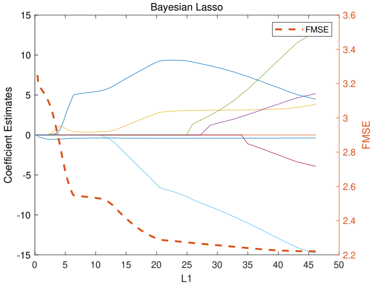
Recall that p_j indicated the probability that β_j is significant.

Coefficient	p_j
Intercept	1
β_1	0.1082
β_2	0.1521
β_3	0.9976
β_4	0.9997
β_5	1
β_6	1
β_7	0.9987
β_8	0.9999

We can conclude that β_1, β_2 may be less important.

Step 6: Model comparison (variable selection with Lasso)

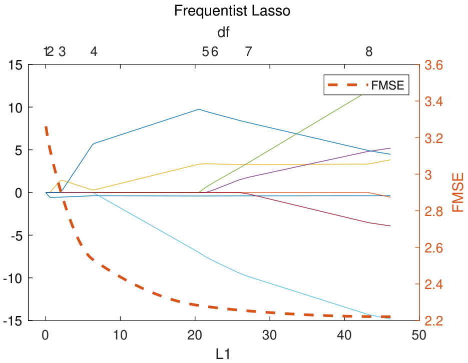
We plot the solution path of Bayesian Lasso.



The insignificant variable selected are the same as SSVS.

Step 6: Model comparison (variable selection with Lasso)

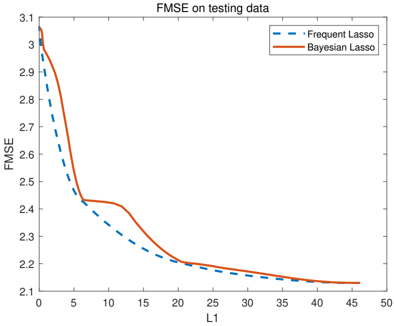
We compared Bayesian Lasso with frequentist Lasso.



Their behaviours are close to each other.

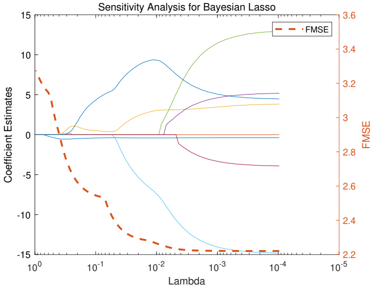
Step 7: Predictive performance assessment

We also evaluate their performances on testing data.



Step 8: Sensitivity analysis with respect to prior choices

Take a look at the path of different λ for Bayesian Lasso.



Further discussion

Interesting problems include:

1. How does these methods work for rare events?
2. Can we apply Bayesian framework to penalties like SCAD?
3. Can we use EM algorithm for SSVS in the frequentist view?

Reference I

[1] Rahim Alhamzawi and Haithem Taha Mohammad Ali. A new gibbs sampler for bayesian lasso. *Communications in Statistics-Simulation and Computation*, 2020.

[2] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 1993.

[3] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 2008.

[4] Wikipedia. Bayesian linear regression. `https://en.wikipedia.org/wiki/Bayesian_linear_regression`. Accessed May 10, 2022.

Thanks!