Variable Selection via Gibbs sampling

Min-Yi Chen

December 4, 2022

1 Introdiction

Linear regression is a popular model in statistics. Generally, we solve regression model by using linear algebra. In recent years, George & McCulloch [1] propose a regression variable selection method under Bayesian framework. In this note, I write down the process in variable selection via Gibbs sampling.

2 Regression Setup

First, we identify the regression.

$$Y|\beta,\sigma^2 \sim N_n(X\beta,\sigma^2 I) \tag{1}$$

where Y is $n\times 1$,and $X=[X_1,...,X_p]$ is $n\times p$. Both β and σ^2 are unknown. Our goal is to simulate and select the subset of $\beta=[\beta_1,...,\beta_p]'$ to fit our regression model. The process is called variable selection. In Bayesian Variable Selection, we do not set $\beta_i=0$ directly since it's hard to define, the method we used is 'spike and slab' mixtures.That is,

$$\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0+\gamma_i N(0,\sigma^2\tau_i^2) \tag{2}$$

where the latent variable $\gamma_i=0$ or 1 ,and δ_0 is the Dirac delta function where

$$\delta_0=\delta(x)\simeq\{\begin{array}{l}1,x=0\\0,x\neq 0\end{array} \tag{3}$$

$$\int_{-\infty}^{\infty}\delta(x)dx=1 \tag{4}$$

the Dirac delta function δ_0 is the spike and τ^2 is the variance of the slab.

Note :'spike' and 'slab' can change to any distribution you want to meet the condition you want.

The reason why 'spike and slab' works in Bayesian Variable Selection is $(1-\gamma_i)\delta_0+\gamma_i N(0,\sigma^2\tau_i^2)$ can be shown as the PDF of $\beta_i|\gamma_i$. I provide a simple discussion in example 2.1.

Assume Equation 2 make sense, we use Bernoulli model to choose whether $\gamma_1 = 0$ or 1

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i \tag{5}$$

$$\gamma \sim Bern(p_i) \tag{6}$$

The prior distribution of $p(\gamma)$ can be shown as:

$$p(\gamma) = \prod p_i^{\gamma_i}(1 - p_i)^{1-\gamma_i} \tag{7}$$

We can see that there's 2^p possible values of γ. Then, we denote the prior of σ^2, we use inverse gamma conjugate prior

$$\tau^2|\gamma \sim IG(\frac{v_\gamma}{2}, \frac{v_\gamma \lambda_\gamma}{2}) \tag{8}$$

where v_γ and λ_γ are the hyperparameters depend on γ, and we can just set them as constant.

Example 2.1. *Suppose that U, Z_0 ans Z_1 are independent random variables on the probability space $(\Omega, \mathcal{F}, P)$, and v is a σ-finite measure on $(\mathbb{R}, B(\mathbb{R}))$. Suppose that $P(U = 1) = 1 - P(U = 0) \in (0, 1)$ and for $i \in \{0, 1\}$, Z_i has a* PDF f_i w.r.t. v. Define

$$X = \begin{cases} Z_0, if\ U = 0 \\ Z_1, if\ U = 1 \end{cases} \tag{9}$$

Let $\pi_0 = P(U = 0), \pi_1 = P(U = 1)$ and define

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$$

for $x \in \mathbb{R}$. Show that f is a PDF of X w.r.t. v.

Proof. Let event $A \in \mathcal{F}$, we want to show

$$P(X \in A) = \int_A f(x)dv$$

We start with the left side

$$\begin{aligned} P(X \in A) &= P((X \in A) \cap (U = 0)) + P((X \in A) \cap (U = 1)) \\ &= P((X = Z_0) \cap (U = 0)) + P((X = Z_1) \cap (U = 1)) \end{aligned} \tag{10}$$

Since U, Z_0, Z_1 are independent

$$P((X = Z_0) \cap (U = 0)) = P(X = Z_0) \times P(U = 0)$$

$$P((X = Z_1) \cap (U = 1)) = P(X = Z_1) \times P(U = 1)$$

This implies

$$\begin{aligned} P(X \in A) &= P(X = Z_0) \times P(U = 0) + P(X = Z_1) \times P(U = 1) \\ &= P(X = Z_0) \times \pi_0 + P(X = Z_1) \times \pi_1 \end{aligned} \tag{11}$$

Since Z_i has PDF f_i w.r.t. v for $i \in \{0, 1\}$

$$\frac{dP(X=Z_0)}{dv} = f_0$$

$$\frac{dP(X=Z_1)}{dv} = f_1$$

We have

$$P(X = Z_0) = \int_A f_0(x)dv(x)$$
$$P(X = Z_1) = \int_A f_1(x)dv(x)$$

We can conclude

$$\begin{aligned} P(X \in A) &= \pi_0 \times \int_A f_0(x)dv(x) + \pi_1 \times \int_A f_1(x)dv(x) \\ &= \int_A \pi_0 f_0(x)dv(x) + \int_A \pi_1 f_1(x)dv(x) \\ &= \int_A [\pi_0 f_0(x) + \pi_1 f_1(x)]dv(x) \\ &= \int_A f(x)dv(x) \end{aligned} \tag{12}$$

We complete the proof. □

3 Identifying the best model

By 'spike and slab', we known that the way we identify the best model is to modify γ. Therefore, we obtain

$$p(\gamma|Y) \propto f(Y|\gamma)p(\gamma) \tag{13}$$

3.1 prior distribution $p(\gamma)$

$p(\gamma|Y)$ can help us to select the more promising model since the updated information of γ. The prior distribution is shown in Equation 7.

3.2 the choice of τ

The choice of τ in Equation 2 should be such that if $\beta_i \sim N(0, \sigma^2\tau_i^2)$, then a non-0 estimate of β_i should be include in the final model we identify. The marginal density can be shown as:

$$\begin{aligned} &(\hat{\beta}_i|\sigma_{\beta_i}, \gamma_i = 0) \sim \delta_0 \\ &(\hat{\beta}_i|\sigma_{\beta_i}, \gamma_i = 1) \sim N(0, \sigma_{\beta_i}^2 + \tau_i^2) \end{aligned} \tag{14}$$

3.3 The relation between all random variable

There's still some prior distributions haven't been decided. Those distributions will be set by us with

$$\begin{aligned} &\theta \sim Beta(a, b) \\ &\sigma^2 \sim IG(\alpha_1, \alpha_2) \end{aligned} \tag{15}$$

where $\theta = p_i$ in Equation 5.
Finally, we can visualize the relations between all random variables in a DAG.

We can write down the joint probability distribution factors:

$$p(y, \beta, \tau^2, \gamma, \theta, \sigma^2) = p(y|\beta, \sigma^2)p(\sigma^2)p(\beta|\gamma, \tau^2)p(\gamma|\theta)p(\theta)p(\tau^2) \tag{16}$$

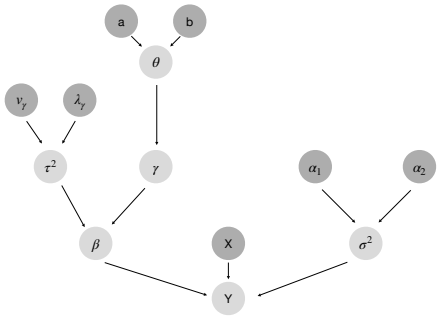

Figure 1: Directed acyclic graphs

4 Gibbs sampling the best subsets

However, the joint distribution is hard to construct since the complex structure. Hence, George & McCulloch [1] reconstruct the Bayesian framework to multiple conditional posterior distribution, and used computational statistics method, Gibbs sampling, to simulate the joint distribution. With the Gibbs sampler, the goal is to generate a sequence

$$\gamma_1, ..., \gamma_m \tag{17}$$

which converges in distribution to $\gamma \sim p(\gamma|Y)$.

To achieve this goal, George & McCulloch [1] generate a Gibbs sequence of the unknown parameters which is relevant to variable selection.

$$\gamma_0, \beta_0, \sigma_0^2, \tau_0^2, ..., \gamma_j, \beta_j, \sigma_j^2, \tau_j^2, ... \tag{18}$$

based on the conditional posterior distribution of each parameter given the data and all other parameters.

$$\begin{aligned}
&p(\theta|y, \beta, \gamma, \tau^2, \sigma^2) = p(\theta|\gamma) \\
&p(\tau^2|y, \beta, \gamma, \theta, \sigma^2) = p(\tau^2|\beta, \gamma) \\
&p(\sigma^2|y, \beta, \gamma, \theta, \tau^2) = p(\sigma^2|y, \beta) \\
&p(\beta|y, \gamma, \theta, \tau^2, \sigma^2) = p(\beta|y, \gamma, \tau^2, \sigma^2) \\
&p(\gamma|y, \beta, \theta, \sigma^2, \tau^2) = p(\gamma|\beta, \theta, \tau^2)
\end{aligned} \tag{19}$$

4.1 Conditional postetrior $p(\theta|\gamma)$

We have

$$\begin{aligned} p(\theta|\gamma) &\propto p(\gamma|\theta)p(\theta) \\ &= \theta^{a-1}(1-\theta)^{b-1}\theta^{\gamma}(1-\theta)^{1-\gamma} \\ &= \theta^{(a+\gamma)-1}(1-\theta)^{(b+1-\gamma)-1} \end{aligned} \tag{20}$$

we get the kernel of the Beta distribution, which implies the posterior distribution as a new Beta distribution:

$$\theta|\gamma \sim Beta(a+\gamma, b+1-\gamma) \tag{21}$$

4.2 Conditional posterior $p(\tau^2|\beta,\gamma)$

We know that $p(\tau^2|\beta,\gamma)$ can expand as:

$$p(\tau^2|\beta,\gamma) = \frac{p(\beta^2|\tau^2,\gamma)p(\gamma)p(\tau^2)}{\int p(\beta^2|\tau^2,\gamma)p(\gamma)p(\tau^2)d\tau^2} = \frac{p(\beta^2|\tau^2,\gamma)p(\tau^2)}{\int p(\beta^2|\tau^2,\gamma)p(\tau^2)d\tau^2} \tag{22}$$

Furthermore, we know γ is the latent variable which can be either 0 or 1. Thus, we can consider the case separately.
Case 1: $\gamma = 1$

$$\begin{aligned} p(\tau^2|\beta,\gamma=1) &\propto p(\beta^2|\tau^2,\gamma)p(\tau^2) \\ &= (2\pi\sigma^2\tau^2)^{\frac{-1}{2}} exp(-\frac{\beta^2}{2\sigma^2\tau^2})\frac{(\frac{v_\gamma\lambda_\gamma}{2})^{\frac{v_\gamma}{2}}}{\Gamma(\frac{v_\gamma}{2})}(\tau^2)^{-\frac{v_\gamma}{2}-1}exp(-\frac{\frac{v_\gamma\lambda_\gamma}{2}}{\tau^2}) \\ &= (\tau^2)^{-\frac{v_\gamma}{2}-1-\frac{1}{2}}exp(-\frac{\beta^2}{2\sigma^2\tau^2}-\frac{\frac{v_\gamma\lambda_\gamma}{2}}{\tau^2}) \\ &= (\tau^2)^{-\frac{(v_\gamma+1)}{2}-1}exp(-\frac{(\frac{\beta^2}{2\sigma^2}+\frac{v_\gamma\lambda_\gamma}{2})}{\tau^2}) \end{aligned} \tag{23}$$

which is a new Inverse Gamma distribution:

$$(\tau^2|\beta,\gamma=1) \sim IG(\frac{v_\gamma+1}{2}, \frac{\beta^2}{2\sigma^2}+\frac{v_\gamma\lambda_\gamma}{2}) \tag{24}$$

Case 2: $\gamma = 0$, then $\beta = 0$, we sample τ^2 from the prior:

$$(\tau^2|\beta,\gamma=0) \sim IG(\frac{v_\gamma}{2}, \frac{v_\gamma\lambda_\gamma}{2}) \tag{25}$$

4.3 Conditional posterior $p(\sigma^2|y,\beta)$

We know that $p(\sigma^2|y,\beta)$ can expand as:

$$\begin{aligned}
p(\sigma^2|y,\beta) &\propto p(y|\beta,\sigma^2)p(\beta)p(\sigma^2) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} exp(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta x_i)^2) \\
&\times \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}(\sigma^2)^{-\alpha_1-1}exp(-\frac{\alpha_2}{\sigma^2}) \\
&\propto (\sigma^2)^{-\frac{n}{2}} exp(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta x_i)^2) \\
&\times (\sigma^2)^{-\alpha_1-1}exp(-\frac{\alpha_2}{\sigma^2}) \\
&= (\sigma^2)^{-(\alpha_1+\frac{n}{2})-1}exp(-\frac{1}{\sigma^2}\left[\alpha_2+\frac{\sum_{i=1}^{n}(y_i-\beta x_i)^2}{2}\right])
\end{aligned} \tag{26}$$

which is a new Inverse Gamma distribution:

$$\sigma^2|y,\beta \sim IG(\alpha_1+\frac{n}{2}, \alpha_2+\frac{\sum_{i=1}^{n}(y_i-\beta x_i)^2}{2}) \tag{27}$$

4.4 Conditional posterior $p(\beta|y,\gamma,\tau^2,\sigma^2)$

In equation 14, we know that there's two cases.
Case 1: $\gamma = 0$

$$\beta|y,\gamma,\tau^2,\sigma^2 \sim \delta_0 \tag{28}$$

Case 2: $\gamma = 1$

$$\begin{aligned}
p(\beta|y,\gamma=1,\tau^2,\sigma^2) &= \frac{p(y|\beta,\gamma=1,\tau^2,\sigma^2)p(\beta|\gamma=1,\tau^2)p(\gamma=1)p(\tau^2)}{\int p(y|\beta,\gamma=1,\tau^2,\sigma^2)p(\beta|\gamma=1,\tau^2)p(\gamma=1)p(\tau^2)d\beta} \\
&= \frac{p(y|\beta,\gamma=1,\tau^2,\sigma^2)p(\beta|\gamma=1,\tau^2)}{\int p(y|\beta,\gamma=1,\tau^2,\sigma^2)p(\beta|\gamma=1,\tau^2)d\beta} \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta x_i)^2) \\
&\times (2\pi\sigma^2\tau^2)^{-\frac{1}{2}} exp(-\frac{\beta^2}{2\sigma^2\tau^2}) \\
&\propto exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta x_i)^2)exp(-\frac{\beta^2}{2\sigma^2\tau^2}) \\
&= exp(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\beta x_i)^2+\frac{\beta^2}{\tau^2}\right]) \\
&= exp(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}y_i^2-2\beta\sum_{i=1}^{n}y_ix_i+\beta^2\sum_{i=1}^{n}x_i^2+\frac{\beta^2}{\tau^2}\right])
\end{aligned} \tag{29}$$

By calculation, we can conclude:

$$\begin{aligned} p(\beta|y,\gamma=1,\tau^2,\sigma^2) &\propto exp(-\frac{1}{2\sigma^2}\left[\beta^2\sum_{i=1}^{n}x_i^2+\frac{\beta^2}{\tau^2}-2\beta\sum_{i=1}^{n}y_ix_i\right]) \\ &= exp(-\frac{(\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2})}{2\sigma^2}\left[\beta^2-\frac{2\beta\sum_{i=1}^{n}y_ix_i}{\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2}}\right]) \\ &\propto exp(-\frac{(\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2})}{2\sigma^2}\left[(\beta-\frac{\sum_{i=1}^{n}y_ix_i}{\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2}})^2\right]) \end{aligned} \tag{30}$$

Hence, we have

$$\beta|y,\gamma,\tau^2,\sigma^2 \sim N(\frac{\sum_{i=1}^{n}y_ix_i}{\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2}},\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2+\frac{1}{\tau^2}}) \tag{31}$$

4.5 Conditional posterior $p(\gamma|\beta,\theta,\tau^2)$

In equation 19, we know that $\gamma\perp y|\beta$ which means that the Bernoulli distribution can be shown as:

$$p(\gamma|\beta,\theta,\tau^2)=\frac{p(\beta|\gamma,\tau^2)p(\gamma|\theta)p(\theta)}{p(\beta|\gamma=1,\theta,\tau^2)p(\gamma=1|\theta)+p(\beta|\gamma=0,\theta,\tau^2)p(\gamma=0|\theta)} \tag{32}$$

We discuss the equation in two cases which is similar above.
Case 1: $\gamma=1$

$$p(\gamma=1|\beta,\tau^2)\propto(2\pi\sigma^2\tau^2)^{-\frac{1}{2}}exp(-\frac{\beta^2}{2\sigma^2\tau^2})\theta \tag{33}$$

Case 2: $\gamma=0$

$$p(\gamma=0|\beta,\tau^2)\propto\delta_0(1-\theta) \tag{34}$$

Then we can conclude the posterior distribution:

$$(\gamma|\beta,\tau^2)\sim Bern(\frac{(2\pi\sigma^2\tau^2)^{-\frac{1}{2}}exp(-\frac{\beta^2}{2\sigma^2\tau^2})\theta}{(2\pi\sigma^2\tau^2)^{-\frac{1}{2}}exp(-\frac{\beta^2}{2\sigma^2\tau^2})\theta+\delta_0(1-\theta)}) \tag{35}$$

Finally, we have all posterior distribution. Now, we can do variable selction via Gibbs Sampling.

4.6 Gibbs sampling

Before using Gibbs sampling in Variable selection, we first introduce the basic concept of the simulated method. Gibbs sampling is one of most famous computational simulation based on Naive Bayes in MCMC.

Recall: In Naive Bayes, we have the equation:

$$P(\theta|\mathcal{X})=\frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}$$

The goal in Gibbs sampling is to approach $P(\theta|\mathcal{X}) \approx \pi(\theta)$ by iterative process. Let $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$, $d \in \mathbb{N}$. Gibbs sampling update θ one dimension per time during iteration. That is
before update: $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})$, after update: $\theta' = (\theta^{(1)}, \ldots, \theta'^{(i)}, \ldots, \theta^{(d)})$
We define the process as $\mathcal{T}_i(\theta \to \theta')$ and the probability to θ' is

$$\mathcal{T}_i(\theta \to \theta') = \frac{\pi(\theta^{(1)},\ldots,\theta^{(i-1)},\theta'^{(i)},\ldots,\theta^{(d)})}{\sum_{\mathcal{Z}} \pi(\theta^{(1)},\ldots,\theta^{(i-1)},\mathcal{Z},\ldots,\theta^{(d)})}$$

A practical example is exhibited below.

Example 4.1. *Let* $\theta = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$, $\theta^{(i)} \in \{1, 2, 3, 4, 5\}$,$\theta = (2, 3, 5)$

$$\mathcal{T}_2(\theta \to \theta') = \frac{\pi(2,\theta'^{(2)},5)}{\pi(2,1,5)+\ldots+\pi(2,5,5)}$$

4.7 Gibbs sampling from a bivariate Gaussian

We implement a concrete example to show how Gibbs sampling work.

Let X_1 and $X_2 \sim$ bivariate normally distribution with mean zero , unit variance and correlation ρ. We have the conditional Gaussian distribution of X_1 given $X_2 = x_2$

$$X_1|X_2 = x_2 \sim N(\rho x_2, (1-\rho)^2) \tag{36}$$

and the conditional Gaussian distribution of X_2 given $X_1 = x_1$

$$X_2|X_1 = x_1 \sim N(\rho x_1, (1-\rho)^2) \tag{37}$$

The Gibbs sampler generate a Gibbs sequence repeatedly from these two conditional distributions:

$$(x_1^1, x_2^1), ..., (x_1^n, x_2^2), ... \tag{38}$$

Generally, we will remove the earlier samples and retain the later samples. We gather those remain samples which is our unique stationary distribution. The step is called burn-in. The result is presented shown below:

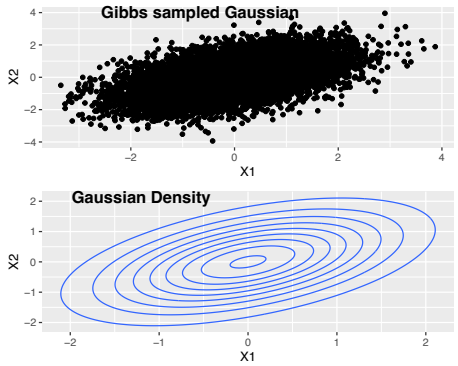

Figure 2: Gibbs sampling and the real density

We can see the data generated by Gibbs sampler approach to the real density.

5 Code implement

Implement Variable Selection via Gibbs Sampling by R.(To be done in the future)

References

[1] George, E. I., McCulloch, R. E. (1993). Variable selection via Gibbs Sampling. Journal of the American Statistical Association, 88(423), 881-889.

[2] Dablander, F. (2019, March 31). Variable selection using Gibbs Sampling. Fabian Dablander. Retrieved November 20, 2022, from https://fabiandablander.com/r/Spike-and-Slab.html