High-dimensional Bayesian Birth-death MCMC Model Selection

Nanwei Wang[*]
Lunenfeld-Tanenbaum Research Institute
and
Laurent Briollais [†]
Lunenfeld-Tanenbaum Research Institute
and
Helene Massam [‡]
Department of Mathematics and Statistics, York Univeristy

November 21, 2018

Abstract

Regression and graphical model are two important statistical tools in data science and statistical genetics. While under high-dimensional setting, which is very common in this big data era, the model selection is a serious problem. In GWAS, It can be used to approximate the relationship between one phenotype variable and genotype covariate variables. They can also be used as predictive models to predict the response variable given specific values of covariates. A proper selected regression model can help us find correlated genotype variables from a huge number of unknown genes, as well as predict the phenotype variable. While, in most of the GWAS data cases, the dimensional of the data is very high, it is easily to overfit the data and to achieve false discoveries. In this paper, we will study a Bayesian birth-death MCMC method to explore the big model space. After we get the approximated posterior distribution of model given data $p(M_i|data)$, the bayesian model averaging is used to select the important covariates or interactions.

1 Introduction

Regression models, from the simple linear regression to the various generalized linear regressions, are widely used in data analysis. Nowadays, most of the data are in high-dimensional, as the number of sample points n is the in the same order of magnitude, or even smaller than the dimensional of the data p. Tibshirani (1996) first proposed Lasso method to perform variable selection in high-dimensional linear regression. Since then, a lot of researchers have

[*] The authors gratefully acknowledge *Funding*
[†] The authors gratefully acknowledge *Funding*
[‡] The authors gratefully acknowledge *Funding*

been working on various penalty vairable slection methods, such as smoothly clipped absolute deviation (SCAD) penalty Fan and Li (2001), Adaptive Lasso Zou (2006) and MCP Zhang et al. (2010). All theses methods can perform well under sparsity assumption, but model uncertainty remains a big challenging for these various penalized regression method, especially in today's big data world. One of the most promising strategies is called 'Bayesian model averaging'. The posterior probability of including a predictor x_v in the regression model is

$$p(x_v \in \mathcal{M}) = \sum_i \mathbf{1}(x_v \in \mathcal{M}_i) p(\mathcal{M}_i|D), \tag{1.1}$$

where $p(\mathcal{M}_i|D)$ is the posterior probability of model $\mathcal{M}_i$ given data $D = \{(y_i, x_i), i = 1 : n\}$, and can be computed as follows:

$$p(\mathcal{M}_i|D) = \frac{\int L(y|X,\beta)\pi(\beta|M_k)d\beta p(\mathcal{M}_i)}{p(D)} \tag{1.2}$$

There are two major problems in ”Bayesian model selection” or ”Bayesian model averaging”. First, the computation of the posterior probability of $p(\mathcal{M}_i|D)$. This probability requires integration over the parameter space.Only if the prior $\pi(\beta)$ is conjugate prior, we can get the exact result of this integration. Otherwise we have to use some approximation methods. Second, the search of the model space $\mathcal{M}$. The cardinality of the model space is usually exponential to the number of variables p, so an efficient MCMC sampling method is required to approximate the posterior distribution. Ye et al. (2018) recently proposed a Sparsity Oriented Importance Learning(SOIL) method, which is similar to 'Bayesian model averaging'. SOIL is a two-step method: first, some sparse candidate models are selected by using several popular penalized likelihood methods; second, use a weighting method to compute the importance of the variables. Ye et al. (2018) didn't point out Bayesian methodology in their work, but the weight of the models is an approximation of the posterior probabilities. The only difference is that SOIL method using selected candidate models, instead of MCMC sampling from model space.

To solve the first problem, we use BIC value of regression models to approximate the posterior probabilities, as given in Wasserman (2000). For the second problem, we will apply Birth-death continuous time MCMC method to approximate the posterior distribution of regression models. Stephens (2000) proposed using Birth-death MCMC procedure to study the mixture models with unknown number of components. Later Cappé et al. (2002) compared Reversible jump MCMC to Birth-death MCMC, and proved some important theoretical results. Recently, Mohammadi et al. (2015), Dobra et al. (2018) applied the Birth-death MCMC method on graphical model learning for continuous data and discrete data,respectively. For the high-dimensional regression problems, the adding or removing a covariate can be treate as a poisson process with some birth or death rate. As we show see in section 3, the birth-death MCMC process with converge to the target posterior distribution with some specific birth, death rate.

2 Introduction to Regression Models

Most of the studies are tying to find relations between a primary interest variable Y, which is also called *response variable or outcomes*, and a series of *explanatory variables* $X = \{X_1, X_2, \cdots, X_p\}$. The explanatory variables can be discrete or continuous. The simplest regression case is linear regression, which we assume Y is a continuous variable following normal distribution with equal variance:

$$Y \sim N(X\beta, \sigma^2),$$

or in the regression equation form, we use y to denote the $n \times 1$ observations of outcomes of Y, X to denote the $n \times p$ design matrix, β to denote the $p \times 1$ coefficient parameters and $\epsilon \sim N(0, \sigma^2)$ to denote the random error term:

$$y = X\beta + \epsilon. \tag{2.1}$$

If Y_i is binary variable, we can fit logistic regression, which we assume the conditional distribution of Y_i given X_i is Bernoulli distribution:

$$Y_i|X_i \sim Bernoulli(p_i),$$

where p is the conditional probability of Y_i taking value 1. Another assumption is that the logarithm of odds $\frac{p_i}{1-p_i}$ is linear in the design matrix x_i:

$$\log \frac{p_i}{1-p_i} = x_i\beta, \tag{2.2}$$

Linear regression and logistic regression are the most popular regression models in data analysis. In the more general cases, we can use the generalized linear regression(GLM). In GLM, the response variable y_i is assumed to follow a distribution with mean μ_i. The second assumption is that there exists a link function g, such that

$$g(\mu_i) = x_i\beta.$$

In theory, the conditional distribution of Y given X can be any distribution, but we use exponential family distributions a lot. The detailed theory won't be covered in this paper.

3 Birth Death MCMC method

Model selection is a classical problem both in frequentist statistics and in Bayesian statistics. Under the regression setting, the model selection problem simply becomes variable selection problem, i.e. which variable or interaction to include in the regression. Readers refer to a review paper Wasserman (2000) for more details on Bayesian model selection and model averaging.

Under Bayesian methodology, a finite model space $\mathcal{M} = \{M_1, M_2, \cdots, M_K\}$ follows a prior distribution $p(M_k), k = 1, 2, \cdots, K$. Given any model M_k, let $\pi_k(\theta)$ be the prior parameters of model M_k. Then from Bayes' theorem, the posterior distribution of model M_k given $D = (y, X)$ is

$$p(M_k|D) = \frac{p(M_k)p(data|M)}{p(data)},$$

where $p(D|M_k) = \int L(y|X, \theta)\pi(\theta|M_k)d\theta$ is the marginal likelihood of data given the model M_k. Bayesian model selection is to find the model which maximizes $p(M|D)$. Since the cardinality of the model space is too large in high dimensional regression problems, as well as the complexity of the marginal likelihood, the computation of posterior distribution $p(M_k|D)$ is intractable most of the time. A straightforward solution is to use a MCMC method to sample from the posterior distribution $p(M|D), M \in \mathcal{M}$. Bayesian MCMC method can not only help select one "best" model, but also average results over different models. The use of

MCMC methods to sample the posterior distribution of some parameters in statistical models is very popular, but the MCMC methods which can do model selection, or travel through models with different dimensions in model space is not well explored. To our best knowledge, the reversible jump MCMC and birth-death continuous time MCMC (Cappé et al., 2002), which we are going to use in this paper, are two such methods.

Birth-death Markov process is a continuous Markov process the model space $\mathcal{M} = \cup_k M_k$, where M_k are disjoint. This process explores the model space by adding and removing covariates corresponding to birth and death jumps. Given the current model M with parameter $\theta_M \in \Theta_M$, the birth and death events are defined as independent Poisson processes:

- Birth event: each variable $X_i \notin M$ is born independently as Poisson process with rate $B_i(M, \theta_M)$.
- Death event: each variable $X_j \in M$ dies independently of other variables as a Poisson process with rate $D_j(M, \theta_M)$.

Now given the occurrence of the birth of $X_i \notin M$, the kernal

$$K_{B_i(M,\theta_M)}(\theta_M, F) = \frac{B_i(M,\theta_M)}{\sum_{i,X_i\notin M} B_i(M,\theta_M)} \int_{\theta_i,\theta_M\cup\theta_i\in F} b_i(\theta_i|\theta_M)d\theta_i$$

denotes the probability that the birth jump leads to a parameter in set $F \in \theta_{M+i}$. The pdf $b_i(\theta_i|\theta_M)$ is where we sample the new parameter in the new model with X_i.

Similarly, given the occurrence of the death of $X_j \in M$, the kernal

$$K_{D_j(M,\theta_M)}(\theta_M, F) = \frac{D_j(M,\theta_M)}{\sum_{j,X_j\in M} D_j(M,\theta_M)} 1(\theta_{M-i} \in F)$$

denotes the probability that the death jump leads to a parameter in set $F \in \theta_{M-i}$.

Definition 3.1. *The distribution $p(M, \theta_M|x)$ satisfies detailed balance conditions if*

$$\int_F \sum_{i,X_i\notin M} B_i(M,\theta_M)p(M,\theta_M|x)d\theta_M = \sum_i \int_{\Theta_{M+i}} \sum_i D_i(M+i,\theta_{M+i})K_{D_i}(\theta_{M+i},F)p(M+i,\theta_{M+i})d\theta_{M+i} \tag{3.1}$$

and

$$\int_F \sum_{j,X_j\in M} D_j(M,\theta_M)p(M,\theta_M|x)d\theta_M = \sum_j \int_{\Theta_{M-j}} \sum_j B_j(M-j,\theta_{M-j})K_{B_j}(\theta_{M-j},F)p(M-j,\theta_{M-j})d\theta_{M-j} \tag{3.2}$$

Lemma 3.2. *The birth-death process has the stationary distribution $p(M|D)$, if the following detailed balance condition is satisfied:*

$$B_i(M)p(M|D) = D_i(M^{+i})p(M^{+i}|D)$$

Proof. Now we try to proof the first detailed balanced equation in definition3.1:

The left side of the equation is

$$\begin{aligned}
LHS &= \textstyle\int_F \sum_{i,X_i\notin M} B_i(M,\theta_M)p(M,\theta_M|D)d\theta_M \\
&= \textstyle\sum_i \int I(\theta_M \in F)B_i(M,\theta_M)p(M,\theta_M|D)d\theta_M \\
&= \textstyle\sum_i \int_{\Theta_M} I(\theta_M \in F)B_i(M,\theta_M)p(M,\theta_M|D)[\int_{\Theta_i} b_i(\theta_i|\theta_M)]d\theta_M \\
&= \textstyle\sum_i \int_{\Theta_M} \int_{\Theta_i} I(\theta_M \in F)B_i(M,\theta_M)p(M,\theta_M|D)b_i(\theta_i|\theta_M)d\theta_M d\theta_i
\end{aligned}$$

The right side of the equation is

$$RHS = \sum_i \int_{\Theta_{M+i}} I(\theta_M \in F) D_i(M+i, \theta_{M+i}) p(M+i, \theta_{M+i}) d\theta_{M+i}$$

Therefore, In order to get LHS=RHS, the following equation needs to be satisfied:

$$B_i(M, \theta_M) p(M, \theta_M|data) b_i(\theta_i|\theta_M) = D_i(M+i, \theta_{M+i}) p(M+i, \theta_{M+i}).$$

Intehrating over θ_{M+i}, we have

$$B_i(M) p(M|data) = D_i(M+i) p(M+i|data).$$

We can also proof the second euqation in the same way. □

Based on the Lemma 3.2, the birth and death rate are defined as follows:

$$\begin{aligned} b_i(M) &= \tfrac{p(M^{+i}|data)}{p(M|X)}, \ X_i \notin M \\ d_i(M) &= \tfrac{p(M^{-i}|data)}{p(M|data)}, \ X_i \in M \end{aligned}$$

At current model M_k, the waiting time to next birth or death event follows an exponential distribution with mean equals to $\frac{1}{\sum_{X_i \in M_k} d_i(M_k) + \sum_{X_j \notin M_k} b_j(M_k)}$. Based on Rao-Blackwellized estimator (Cappé et al., 2002), the posterior probability of each sampled model is proportional to the expectation of the length of its waiting time. The birth-death MCMC algorithm can be summarized as follows:

1. Given the input data (y, X) and set the starting model M_0 as $y = \beta_0 + \epsilon$;
2. at the k^{th} iteration in the MCMC process, compute the birth and death rate of each variable $X_i, i = 1 : p$;
3. Calculating the waiting time for M_k by $W(M_k) = \frac{1}{\sum_{X_i \in M_k} d_i(M_k) + \sum_{X_j \notin M_k} b_j(M_k)}$;
4. Sample a birth or death event based on the birth or death rates. Move to the next model M_{k+1}. If the birth event of x_i is sampled, then $M_{k+1} = M_k \cup x_i$; while if the death event of x_j is sampled, then $M_{k+1} = M_k \setminus x_i$
5. Repeat step 2 to step 4 until the distribution stable.

After the BDMCMC process, we get a sample from $p(M|data)$. There are two ways to select a model from the Bayesian model selection framework. We can either select the model with the highest sampled posterior probability, or take the average of some models with relative high posterior probabilities. The remaining problems are how to compute the marginal likelihood of data given a model $l(data|M)$ for different regressions, which we will study in the following sections.

4 Computation of birth-death rate

To simplify the notation, let $r_i(M_0, \theta_0) = \frac{p(M_1|D)}{p(M_0|D)}$ denote the change rate from old model M_0 jump to new model M_1. Immediately, we have

$$r_i(M_0, \theta_0) == \begin{cases} b_i(M_0, \theta_0) & x_i \notin M_0 \\ d_i(M_0, \theta_0) & x_i \in M_0 \end{cases}$$

In this section, we will offer a fast and accurate estimation of $r_i(M_0, \theta_0)$:

$$r_i(M_0, \theta_0) = \frac{p(M_1|D)}{p(M_0|D)} = \frac{p(D|M_1)}{p(D|M_0)} \times \frac{p(M_1)}{p(M_0)},$$

so the change rate r_i is the product of bayes factor $BF(M_1, M_0)$, and the ratio of model prior.

4.1 BIC and Extended-BIC

Let's look at the computation of Bayes factor first. Most of the circumstance, computing the exact Bayes factor value is difficult. While in regression models, we can use Bayesian information criterion(BIC) value of the model to approximate $p(D|M)$. (Wasserman, 2000) has showed that

$$\log p(D|M) = l(\hat{\beta}) - \frac{d}{2}\log n + \mathcal{O}(1),$$

Where $l(\hat{\beta})$ is the log-likelihood function value with the MLE of β, d is the dimension of the regression, i.e. the length of β and n is the sample size. Therefore, we can use the BIC value $BIC(M) = -2l(\hat{\beta}) + d\log(n)$ to approximate the marginal likelihood:

$$p(D|M) \approx \exp(-BIC(M)/2)$$

This approximation requires no integration and does not depend on the prior of parameters in the model. The error term $\mathcal{O}(1)$ doesn't converge to 0 as $n \to \infty$, but it is a relative small value compared to $\log p(D|M)$ as $n \to \infty$. However, in high-dimensional problems, which is the main study objections in this paper, the original BIC doesn't work very will. Chen and Chen (2008) proposed Extended Bayesian Information Criteria(E-BIC), whichs take into account both the number of unknown parameters and the complexity of the model space, for small-n-big-p problems. The E-BIC formula is as follows:

$$E - BIC(M, \theta) = -2\log L(\theta)) + d\log(n) + 2\gamma\log\tau(M),\ 0 \le \gamma \le 1 \tag{4.1}$$

where $\tau(M)$ is the number of models with same dimension d as model M. In regression models, assume the number of variables in model M is k, then $\tau(M) = \binom{p}{k}$.

5 Prior on model space $\mathcal{M}$

For the prior of the model space, It makes sense to put more weight on models with fewer variables in small-n-big-p, i.e. we would like to choose an prior to give us a sparse model selection result. We offer three options in this paper:

1. Let k denote the number of variables in model M, the prior is

   $$p(M) \propto \alpha^k, a \in (0, 1]$$

   The prior is similar to the one given in Dobra et al. (2018). The smaller α is, the bigger the ratio between a dense model and a sparse model is.

2. The second prior is modified from the one given in Nan and Yang (2014):

   $$p(M) \propto \exp(-\gamma C_M),$$

   where $0 \le \gamma \le 1$, $C_M = \log\binom{p}{k} + 2\log(k)$. This prior is similar to the E-BIC's idea. It take the model complexity into consideration.

3. The third prior is given in Scott and Berger (2010):

   $$p(M) = \alpha^k(1-\alpha)^{p-k}, \quad 0 \le \alpha \le 1.$$

   This prior is to treat variable inclusions as exchangeable Bernoulli trials with common success probability α.

Combine the Bayes factor $BF(M_1, M_0)$ and model space prior $p(M)$, we can get the results of all the change rates.

References

Allen, G. I., Liu, Z., et al. (2013). A local poisson graphical model for inferring networks from sequencing data. *IEEE Trans NanoBiosci*, 12(3):189–98.

Cappé, O., Robert, C. P., Rydén, T., and Enz, T. R. (2002). Reversible jump mcmc converging to birth-and-death mcmc and more general continuous time samplers.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Cheng, J., Li, T., Levina, E., and Zhu, J. (2017). High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378.

Dobra, A., Mohammadi, R., et al. (2018). Loglinear model selection and human mobility. *The Annals of Applied Statistics*, 12(2):815–845.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Massam, H. and Wang, N. (2018). Local conditional and marginal approach to parameter estimation in discrete graphical models. *Journal of Multivariate Analysis*, 164:1–21.

Mohammadi, A., Wit, E. C., et al. (2015). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

Nan, Y. and Yang, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics*, 23(3):636–656.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of statistics*, pages 40–74.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107.

Yang, E., Baker, Y., Ravikumar, P., Allen, G., and Liu, Z. (2014). Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics*, pages 1042–1050.

Ye, C., Yang, Y., and Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association*, pages 1–16.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.