High-dimensional Baybesian Model Selection

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

December 11, 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. Lasso and some other varieties are classical model selection methods in Frequentist inference, but these methods usually suffer two problems: the estimation of likelihood function and the choice of penalty parameters. On the other hands, in Bayesian model selection, the computation of Bayes factor, and the model search in high dimensional problems are also difficult to handle. In this paper, we will modify the Birth-death MCMC(BDMCMC) method proposed by Stephens (2000) and apply it to regression and graphical model problems.With the BDMCM method, we can quickly get samples from the approximated posterior distribution of model given data $p(M_i|data)$, then 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-

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

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 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

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

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

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

There are two major problems in "Bayesian model selection" or "Bayesian model averaging". First, the computation of the posterior probability of $p(\mathcal{M}_i|data)$. 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 Model Selection in Regressions and Graphical Models

Linear regression is used to build a linear relationship between the response variable y and given predict variables $(x_1, x_2, \cdots, x_k)$:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_j x_j + \epsilon$$

If the response variable y is binary, we then fit a logistic regression:

$$\log \frac{p(y=1|x)}{p(y=0|x)} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j.$$

Given data $D = (X, y)$, where y is $n \times 1$ sample vector from response variable, X is $n \times k$ covariate matrix, one wants to find the best regression model as given in above formulas, or equivalently, select the best subset of variables, to fit the data.

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.

The second model selection problem we are dealing with in this paper is graphical model. Graphical model is a powerful statistical tool to model the conditional independence relationship among given variables $x = (x_1, x_2, \cdots, x_p)$. There are two popular types of graphical models: Gaussian graphical models and Markov random field.

Suppose X follow the Multivariate Normal distribution $\mathcal{N}(0, \Sigma)$. Normal distribution belongs to exponential family, so we can write the probability density function as follows:

$$f(x|K) = \exp\{\langle -\frac{1}{2} x x^t, K \rangle - (\frac{p}{2} \log(2\pi) - \frac{1}{2} det(K))\},$$

Where $K = \Sigma^{-1}$ is the canonical parameter, $\langle , \rangle$ denotes the inner product between the canonical parameter and sufficient statistic $-\frac{1}{2}xx^t$ and $(\frac{p}{2}\log(2\pi) - \frac{1}{2}det(K))$ is the log-partition function. Let$G = (V, E)$ denote an undirected graph with vertices $V = \{1, 2, \cdots, p\}$ and edges $E \subset V \times V$. If a continuous p-dimensional random vector X follows Multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with

$$(\Sigma^{-1})_{ij} = 0 \text{ for all } (i, j) \notin E,$$

then X is said to satisfy the Gaussian graphical model Markov with respect to G. Given data from the true distribution of X, the model selection in Gaussian graphical model is to estimate the non-zeros entries of the inverse covariance matrix K, which also corresponding to the graph G.

On the other hand, Markov Random Field, mostly used in machine learning, is a graphical model for discrete data, specially binary data. Assume p-dimensional random vector X takes value from $\{0, 1\}^p$ and G is the graph as above. If

$$X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}, \text{ for all } (i, j) \notin E,$$

then X is said to satisfy the Markov Random Field with G. The probability Mass function of X is given as follows:

$$p(x) = \frac{1}{z(\theta)} \exp(\sum_{v=1}^{p} \theta_v x_v + \sum_{e \in E} \theta_{ij} x_i x_j)$$

then the model selection is to find the graph structure G, or estimate the parameter $\theta_{ij}, (i, j) \in E$.

Recently, a new type of graphical model, mixed graphical model, has been studied a lot due to the complexity of the variable type in big data problems. Lauritzen (1996) firstly studied the *Conditional Gaussian* density, the continuous variables follow Gaussian distribution conditioned on other discrete variables, and its Markov properties. Yang et al. (2014) proposed mixed graphical models via exponential family distributions, Cheng et al. (2017) studied the Simplified Mixed Graphical Model for Gaussian and binary variables. Although the definition of mixed graphical models are different in above papers, they use the same model learning method: Lasso. In this section, we will first define our new local mixed graphical models and later in the paper, we will use the BDMCMC method to learn the best local mixed graphical models.

The joint distribution of mixed graphical models are given in both papers Cheng et al. (2017) and Yang et al. (2014). While in Cheng et al. (2017), only Gaussian and binary variables are taken into consideration. Yang et al. (2014)'s mixed graphical models are also limited. As pointed in the paper, the parameters in Poisson graphical models can only be negative, otherwise, the normalization constant is infinity. Therefore in this paper, we proposed our local mixed graphical models(LMGM).

Let $X = (X_1, \cdots, X_{p1}), Y = (Y_1, \cdots, Y_{p2}), Z = (Z_1, \cdots, Z_{p3})$ denote continuous, binary, and counts variables, respectively. Since the joint distribution of all the mixed variables imposes restrictions on parameters. we define the mixed graphical models locally without specifying the joint distribution.

Definition 2.1. *The mixed variables $W = (X, Y, Z)$ follow **Local Markov Property** (see Lauritzen (1996)) with respect to an undirected graphic $G = (V, E)$. The local mixed graphical model is a series of conditional distribution of each variable W_v given its neighbours W_{N_v}, and we assume the conditional distributions follow exponential family distributions:*

$$p(W_v|W_{N_v}) = \exp(\theta_0 W_v + \sum_{j \in N_v} \theta_j W_v W_j + A(W_v) - K(\theta))$$

Remark 2.1. *The local mixed graphical model is an extension from the local Poisson graphical model in Allen et al. (2013). Under the high-dimensional setting, the fact that the global model follows a probability distribution is difficult to satisfy, and sometimes requires additional constrains on parameter space. The other problems for global model in high-dimensional setting is the model learning and parameter estimation problem. The two problems are untractable through global MLE when the dimension of the model is very high. Therefore, even through the joint distribution of mixed graphical models are given in both papers Cheng et al. (2017) and Yang et al. (2014), their model learning method is still penalized pseudolikelihood. i.e they are still using the local model estimates to approximate the global model parameters. For the performance of local model estimation, Massam and Wang (2018) has studied different local estimation methods, and showed how close they are to the global parameters.*

In this paper, we only consider three types of common variables: Gaussian, binary, count.

- For Gaussian variable x_i, the condition distribution follow a Normal distribution with mean:
$$E(x_i|x_{-i}, y, z) = \theta_0 + \theta^x_{-i} x_{-i} + \theta^y y + \theta^z z;$$
- For binary variable y_i, the conditional distribution is a Bernoulli distribution with mean:
$$E(y_i|x, y_{-i}, z) = \frac{\exp(\theta_0 + \theta^x x + \theta^y_{-i} y + \theta^z z)}{1 + \exp(\theta_0 + \theta^x x + \theta^y_{-i} y + \theta^z z)}$$
or in GLM, we write the following regression
$$\log \frac{p(y_i = 1|x, y_{-i}, z)}{p(y_i = 0|x, y_{-i}, z)} = \theta_0 + \theta^x x + \theta^y_{-i} y + \theta^z z$$
- For counts variable z_i, we assume the conditional distribution follows a Poisson distribution. The GLM is
$$\log E(z_i|x, y, z_{-i}) = \theta_0 + \theta^x x + \theta^y y + \theta^z_{-i} z$$

3 Birth Death MCMC method

In frequentist statistics, Lasso is mostly used to do model selection in regression and graphical models, but we know there are some limits in Lasso. In this section we will talk about the model selection in Bayesian framework, and explore the new Birth-death MCMC bayesian method. Readers refer to a review paper Wasserman (2000) for more details on Bayesian model selection and model averaging.

Assume 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 data D is

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{j=1}^{K} p(D|M_j)p(M_j)}, \tag{3.1}$$

where $p(D|M_k) = \int p(D|\theta)\pi(\theta|M_k)d\theta$ is the marginal likelihood of data given the model M_k, it is also called the evidence of model M_k. Based on the posterior distribution, one can perform pairwise comparison between any two models M_1, M_2:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)} \times \frac{p(M_1)}{p(M_2)}.$$

While when the model space $\mathcal{M}$ is very large, pairwise model comparison to find the best model is not feasible, neither the computation of equation 3.1. Markov Chain Monte Carlo (MCMC) can be used to simulate an approximate sample from the posterior distribution. This sample is then can be used to find the model with largest posterior density, or get the average of models with high posterior densities. 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*

$$\int_F \sum_{i,X_i\notin M} B_i(M,\theta_M)p(M,\theta_M|x)d\theta_M = \sum_i \int_{\Theta_{M+i}} \sum_i D_i(M+i,\theta_{M+i})K_{D_i}(\theta_{M+i},F)p(M+i,\theta_{M+i})d\theta_{M+i} \tag{3.2}$$

and

$$\int_F \sum_{j,X_j\in M} D_j(M,\theta_M)p(M,\theta_M|x)d\theta_M = \sum_j \int_{\Theta_{M-j}} \sum_j B_j(M-j,\theta_{M-j})K_{B_j}(\theta_{M-j},F)p(M-j,\theta_{M-j})d\theta_{M-j} \tag{3.3}$$

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

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

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

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

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:

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

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 $p(M|D)$. 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 evidence of model M_k: $p(D|M_k)$, 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:

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

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 $\tau(M) = \binom{p}{k}$.

4.2 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 $0 \leq \gamma \leq 1$, $C_M = \log \binom{p}{k} + 2\log(k)$. 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):

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

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.

5 Numerical Simulations

In this section, We will generate data sets from different regression models and analyze the performance of the BDMCMC method. We compare the results with with Lasso and SOIL method gave in Ye et al. (2018). We use the F_1-score Baldi et al. (2000) to evaluate the performance of the three methods:

$$F_1 - score = \frac{2TP}{2TP + FP + FN},$$

where TP, FP,FN are the true positive, false positive, false negative respectively.

5.1 Linear Regression Model Selection

First, we use a simulation example to see how the BDMCMC method works. Data $(y_i, x_i)_{i=1}^n$ are generated from the linear model $y_i = x_i^t \beta + \epsilon$, $\epsilon \sim N(0,1)$. The covariates x_i is generate from $N(0, \Sigma)$, where $\Sigma_{i,j} = |0.5|^{|i-j|}$. Consider the following setting:

$$p = 1000, \ \beta = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, rnorm(5, 0, 2), rnorm(10, 0, 0.1), 0, \cdots, 0).$$

There are 20 variables out of 1000 predictors in the true model, the first 10 coefficients are large(five given values, five generated from $\mathcal{N}(0, 2^2)$), and the next 10 coefficients are relative small(generated from $\mathcal{N}(0, 0.1^2)$). In the Figure 1, we can see that BDMCMC reached good convergence results after around 20 steps. Therefore, we only use around 100 iterations for the linear regression problems.

To evaluate the performance of the three methods, we first generate sample from the above true linear regression model setting with different sample size, then we compute the F_1-scores of three methods. The process is repeated 100 times to get an average value of the F_1-scores

and the variance. The result is shown in Table 1, the BDMCMC method out-performances Lasso method and is better than SOIL method by a small margin.

Table 1: The F_1-scores of the three methods, the mean and variance are computed from 100 experiments

Sample size	LASSO		SOIL		BDMCMC	
	mean	variance	mean	variance	mean	variance
100	0.3372	0.0041	0.6071	0.0041	0.6079	0.0030
300	0.4368	0.0068	0.6411	0.0019	0.6446	0.00188
500	0.5163	0.0088	0.6753	0.0021	0.6787	0.0018
700	0.5518	0.0094	0.6954	0.0032	0.6931	0.0019
1000	0.6001	0.0094	0.7111	0.0030	0.7114	0.0027

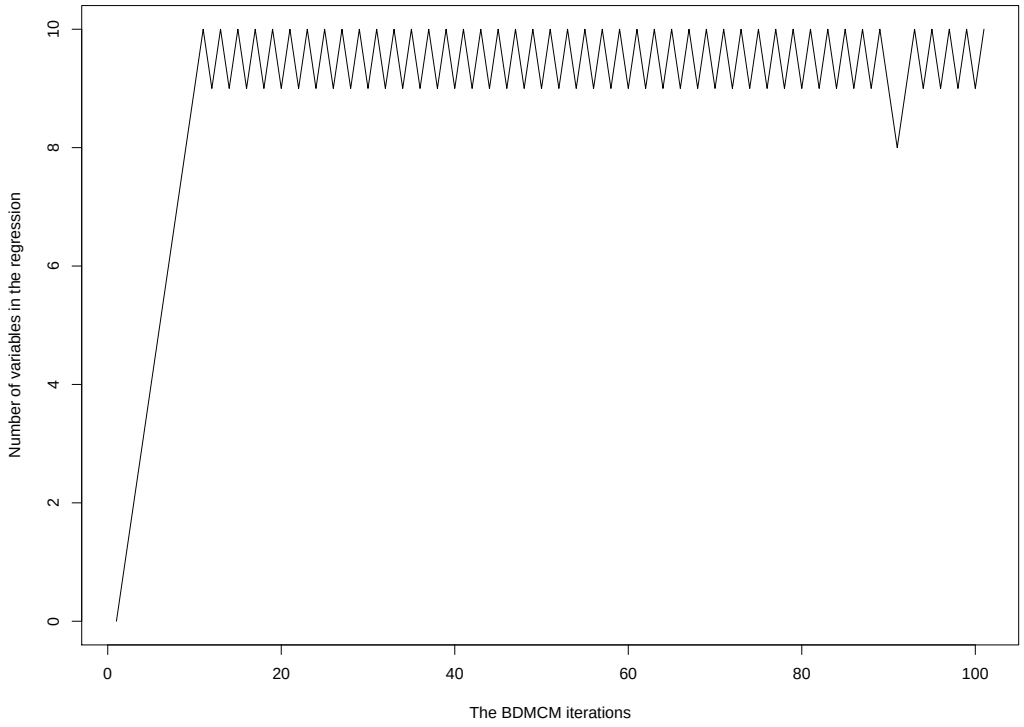

Figure 1: The convergence speed of the BDMCMC method

5.2 Logistic Regression Model Selection

Similar to linear regression, we can have a look at the logistic regression simulation results. Assume there are $p = 1000$ predictor variables, we generate the predictor vectors $x_i, i = 1, 2, \cdots, n$ from p-dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{j,k} = |0.5|^{|j-k|}$. The binary response variable y_i is generated from the Bernoulli distribution with the probability of $y_i = 1$ as

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

and $\beta = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, rnorm(5, 0, 2), rnorm(10, 0, 0.1), 0, \cdots, 0)$. The coefficients are the same as the linear regression simulation and we perform the same experiments as the linear regression. The result is shown in Table 2. The BDMCMC method is better than Lasso, but slightly worse than SOIL method in the logistic regression models.

Table 2: The F_1-scores of the three methods, the mean and variance are computed from 100 experiments

Sample size	LASSO		SOIL		BDMCMC	
	mean	variance	mean	variance	mean	variance
100	0.285	0.016	0.158	0.023	0.190	0.021
300	0.553	0.025	0.589	0.029	0.562	0.023
500	0.646	0.019	0.712	0.014	0.688	0.015
700	0.705	0.017	0.760	0.013	0.738	0.013
1000	0.743	0.0.013	0.786	0.014	0.767	0.014

5.3 BDMCMC algorithm for LMGM and numerical experiments

For each variable X_v, we can apply the BDMCM method to select the variables in the neighbourhood N_v. The process is the same as the linear regression problem, and we still use the BIC value to approximate the marginal likelihood $p(X_v|G)$. After the neighbourhood structure of every variable is selected, then we take the intersection of the local models to get the global graph G. One can also take the union of the local models.

Next, we use an simulated example to show the performance of the BDMCMC algorithm and compare it to the regular lasso. First, we generate a random graph in which edges were randomly generation for 100 continuous variables, 50 binary variables and 50 counting variables. Second, sample the parameters from standard normal distribution $\mathcal{N}(0, 1)$. Third, generate sample points using Gibbs sampling method.

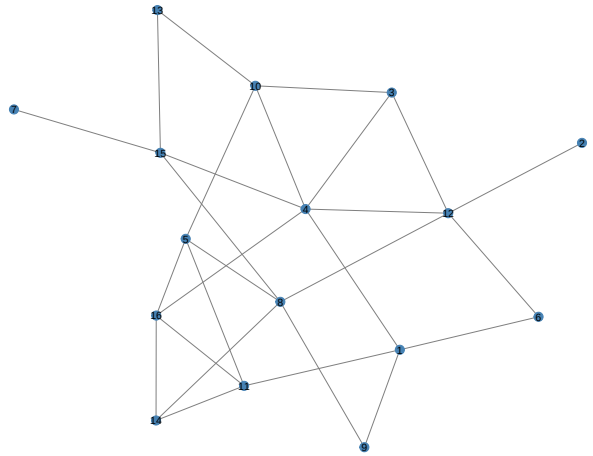

Figure 2: The graphical model with 16 gaussian variables

	BDMCMC	LASSO	BDMCMC(MOHAMMADI)
TP	0.74	0.72	0.37
FP	0	0.05	0

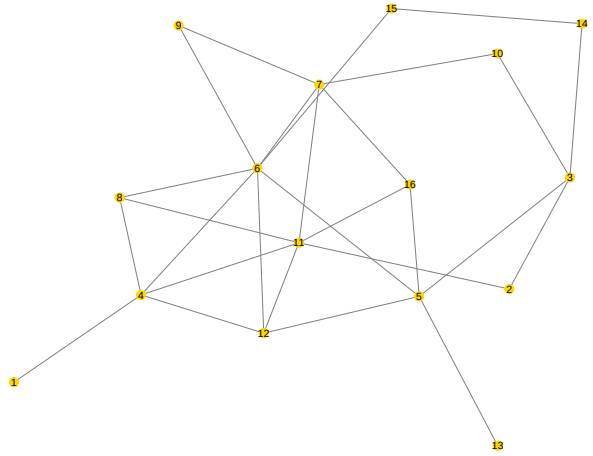

Figure 3: The graphical model with 16 binary variables

	BDMCMC	LASSO	BDMCMC(MOHAMMADI)
TP	0.55	0.41	0.34
FP	0.03	0.27	0

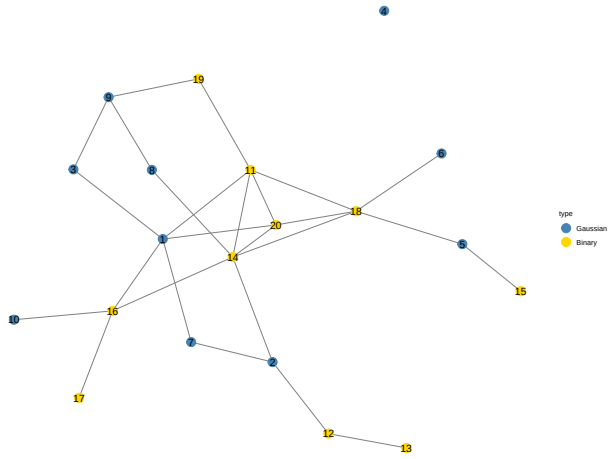

Figure 4: The mixed graphical model with 10 gaussian variables and 10 binary variables

	BDMCMC	LASSO
TP	0.65	0.61
FP	0.24	0.11

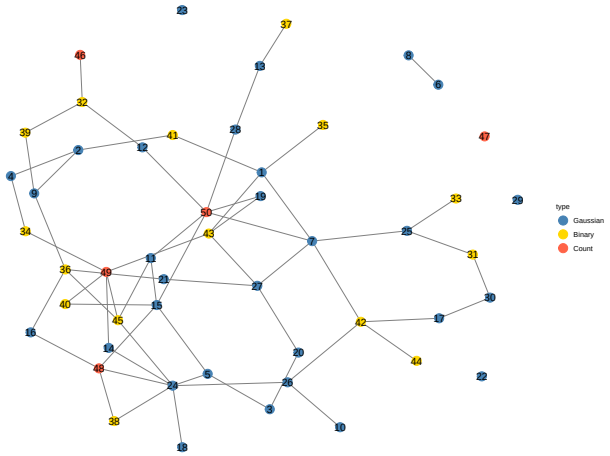

Figure 5: The mixed graphical model with 30 gaussian variables and 15 binary variables and 5 counts

	BDMCMC	LASSO
TP	0.65	0.61
FP	0.24	0.11

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.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

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.