BAYESIAN VARIABLE SELECTION: EXTENSIONS OF NONLOCAL PRIORS

By

Guiling Shi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2017

ABSTRACT

BAYESIAN VARIABLE SELECTION: EXTENSIONS OF NONLOCAL PRIORS

By

Guiling Shi

In presence of high dimensional cavariates, variable selection is an important technique for any further data analysis. Bayesian analysis can reach the aim of model selection based on shrinkage priors. First I would explain Bayesian variable selection technique through three methods, which have been demonstrated giving plausible performance when working on high dimensional model selection problems. I also compared these methods based on both simulation results and real data application. Further I extend the method based on Dirichlet-Laplace prior from normal means problem to linear regression model, and show the minimax contraction rate still holds under mild conditions.

While most developments in Bayesian model selection literature are based on local prior on regression parameters, Johnson and Rossell(2012, 2013) proposed a nonlocal prior distribution for model selection. Enlightened by this idea, I applied nonlocal prior while performing spike and slab variable selection method. I used a point mass density for spike prior, while applied nonlocal prior as slab density, this setting could make overlap between spike and slab prior very little, which could achieve variable selection result efficiently. Following I proved the consistency for variable selection of proposed method.

At last, I extended nonlocal prior model selection method from Johnson and Rossell's method to logistic regression and to generalized linear models. Laplace approximations are used in implementation process due to complicated likelihood. Also, convergence rate is derived under some regularity conditions. The selection based on a nonlocal prior eliminates unnecessary variables and recommends a simple model. This method is validated by

simulation study and illustrated by real data example.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Taps Maiti for the continuous support of my Ph.D study and related research, for his encouragement, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

At the same time, I would like to express my special appreciation to my co-advisor Dr. Chae Young Lim, not only for her tremendous academic support, but also for her patience, genuine caring and concern. Without all her contributions of time and ideas, it is not possible to make this thesis completed and comprehensive.

Besides my advisors, I would like to thank my thesis committee Dr. Jongeun Choi and Dr. Hyokyoung Hong, for their encouragement and serving as member of committee.

Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this thesis and my life in general.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

With the emerge of high dimensional data in many industry, especially in clinical and genetic research, variable selection is one of the most commonly used technique now. The aim of variable selection is to select the best subset of predictors. Remove the redundant predictors is a good way to explain the data in simple way, since unnecessary predictors will add noise to the estimation of interested variables. Many research has been done related with variable selection under both frequentist and Bayesian framework.

Under frequentist statistics, the simple method for variable selection procedure is backward elimination, forward selection and stepwise regression. While some issues are related with stepwise regression, for example, collinearity can be a major issue, some variables may be removed from the model when they are deemed important. Criterion based model selection procedures is popular under frequentist study. Some popular methods include Akaike information criterion, Bayesian information criterion, Mallows C_p. Recommendation could be provided based on comparison of these criterion for each subset model. However, with p potential predictors, there are 2^p possible models. If p is large, it is impossible to fit all these models and choose the best one according to some criterion. This limitation encourages the emergence of many penalization based ideas. Among which, some popular and proved to be effective methods incorporate least absolute shrinkage and selection operation(LASSO), smoothly clipped absolute deviation(SCAD), elastic net and ridge regression. Variable selection methods for generalized linear model are also based on criterion based procedure, or

regularization path.

In Bayesian analysis, variable selection is achieved by shrinking priors. Under appropriate priors, the solution for Bayesian methods could be equivalent to frequentist penalized approach. For example, Laplacian prior for each coefficient could result to LASSO solution, Gaussian prior is corresponding with L_2 penalty result. However, instead of searching through model space and selection criteria for choosing between competing models, Bayesian methods focus on the marginal posterior probability of covariates that should be in the model. Many shrinkage priors have been proposed in Bayesian literature, most of them are local priors, which means prior function value is positive when parameter value is zero. Examples of local priors include Gaussian, Cauchy and Laplacian prior. The positive density around zero could shrink coefficient value to zero and achieve the aim of variable selection result.Among all local priors, one special example is horseshoe prior. horseshoe prior is unbounded with singularity at zero, it is formulated to obtain marginals having a high concentration around zero with heavy tails. The singularity at zero point coupled with tail robustness properties leads to excellent empirical performance of the horseshoe. Compared with common shrinkage priors, horseshoe concentrates more along sparse regions of the parameter space, reference as in [Carvalho, Polson, and Scott(2009)] and [Carvalho, Polson, and Scott(2010)]. While another prior used to contrary with horseshoe is nonlocal prior density. Nonlocal prior densities are exactly zero whenever a model parameter equals its null value. It is first defined by [Johnson and Rossell(2010)] in the context of hypothesis testing, then it is extended in model selection problems in [Johnson and Rossell(2012)], where product moment and product inverse moment prior densities were introduced as priors on a vector of regression coefficients. Model selection consistency property was demonstrated for model selection procedures based on these nonlocal priors when $p \leq n$. More recently, [Rossell et al. (2013)] proposed product

exponential moment prior density with similar behavior to product inverse moment prior. However, model selection property of nonlocal priors in $p \gg n$ settings remain understudied.

In high-dimensional data, one of the most useful technique for Bayesian variable selection is spike and slab prior, which introduces a latent variable for each covariate to indicate whether the covariate is active in the model or not. First proposed by [Mitchell and Beauchamp(1988)], then generalized by [Ishwaran and Rao (2005)], various selection procedures with spike and slab structure have been proposed, they essentially differ in the form of spike priors and slab priors. For example, Gaussian distributions for both spike and slab prior in [George and McCulloch (1993)], uniform distribution for the slab prior in [Mitchell and Beauchamp(1988)]. In addition, variable selection consistency have been established for spike and slab prior in [Ishwaran and Rao (2011)] and [Narisetty and He(2014)].

In this thesis, I did thoroughly study on Bayesian variable selection method for both linear model and generalized linear model. First of all, I explained Bayesian variable selection procedure by reviewing three demonstrated performance methods, and extended one of the method(Dirichlet-Laplace prior) from normal means problem to linear model. In this review study, nonlocal prior method outperforms the other method from application study, also it has some unique properties. This is the motivation of why I believe extensions of nonlocal prior is an interesting topic. This is covered by Chapter 2. In chapter 3, I explored spike and slab variable selection method when slab prior is nonlocal density. The performance of proposed method can be validated by simulation and data application results. In chapter 4, I extended variable selection method based on nonlocal prior to generalized linear model. Convergence rate under proposed method is derived under some regularity conditions. For clarification, each chapter uses independent notations.

Chapter 2

Comparison study on high-dimensional Bayesian variable selection methods

2.1 Introduction

Variable selection is a problem where we identify a subsets of the covariates $\{x_1, x_2, ..., x_n\}$ thats forms a causal features of the given data y. It becomes a great challenge in a small n large p situation. The penalised likelihood approach is to find a linear model where a subset of $\{x_1, x_2, ..., x_n\}$ minimize a penalized likelihood. For example L_1 penalty norm leads to the LASSO method.

[Tibshirani(1996)] pointed out that the LASSO estimator can be interpreted as the maximum a posteriori (MAP) estimator when the regression parameters have independent and identical Laplace priors. For large n small P regression, [Liang, Truong and Wong(2001)] established an explicit relationship between the Bayesian approach and the penalised likelihood approach for linear regression. They showed empirically that Bayesian subset regression (BSR) is choosing priors such that the resulting negative log-posterior probability of the subset model can be approximately reduced to frequentists subset model selection statistic upto a multiplicative constant.

Extending from the idea of spike and slab prior, [Narisetty and He(2014)] introduced shrinking and diffusing priors to establish the strong selection consistency of the approach for $p = e^{o(n)}$. Under the global-local prior, [Bhattacharya et al.(2014)] proposed a Dirichlet-Laplace prior, and showed the minimax optimal rate of posterior contraction. [Johnson(2013)] proposed a coupled Metropolis-Hastings algorithm, which allows moving between any two models. These three paper are focusing on high-dimensional Bayesian variable selection, while through different prior and algorithm. In this chapter, we review the idea for these three paper, and compare the result with application results.

2.2 Methodology

In this section, we present three Bayesian methods solving problem in high-dimensional case. They take Dirichlet-Laplace prior(DL), shrinking and diffusing prior(SD) and product moment prior respectively. See following for specific method.

2.2.1 Dirichlet-Laplace prior(DL)

This part is based on paper Dirichlet-Laplace(DL) priors for optimal shrinkage([Bhattacharya et al.(2014)]), which proposed DL prior to shrink some coefficients into zero.

In high-dimensional settings, most penalization approaches have a Bayesian interpretation as corresponding to the mode of a posterior distribution under a shrinkage prior. Different penalty method can be explained by different priors. [Polson and Scott(2010)] showed that essentially all such shrinkage priors can be represented as global-local(GL) mixtures of Gaussians. In this part, we consider normal means problem, expressed by $y_i = \theta_i + \epsilon_i, \epsilon_i \sim N(0,1), 1 \leq i \leq n$.

We can describe the setting of GL prior specifically.

$$(\theta_i|\tau, \psi_i) \sim N(0, \psi\tau)$$

$$\psi_i \sim f(\psi_i)$$

$$\tau \sim g(\tau)$$

Here, ψ is a vector with components ψ_i, and each ψ_i is called a local variance component, it allows deviations in the degree of shrinkage, while τ is the global variance component, it controls global shrinkage towards the origin. In the penalized-likelihood formulation, τ plays the role of regularization parameter.

If we choose f and g appropriately, a lot of frequentist regularization procedures such as ridge, lasso, bridge and elastic net can be explained under GL prior.

One example is double exponential prior. If θ_i follows double exponential prior, that is $\theta_i \sim \frac{1}{2b} exp\left(-\frac{|\theta_i|}{b}\right)$, then to maximize a posteriori probability estimator corresponding to L_1 or LASSO penalty. If $f(\psi_i)$ is an exponential distribution, after integrating out the local scales ψ_i, θ_i follows a double exponential prior. And LASSO solution is obtained in this setting. In a specific case, if $f(\psi_i)$ is an exponential distribution with scale parameter $\frac{1}{2}$, that is $\psi_i \sim Exp(1/2)$, then θ_i follows a double exponential distribution with parameter τ, that is $\theta_i \sim \frac{1}{2\tau} exp\left(-\frac{|\theta_i|}{\tau}\right)$. We assume this is the case throughout this paper for DL prior.

Another example is horseshoe prior. If f is half Cauchy prior, $\psi_i^{1/2} \sim Ca_+(0, 1)$, this resulting θ_i is horseshoe prior. Horseshoe prior is unbounded with a singularity at zero. Along with tail robustness property leads to excellent empirical performance of the horseshoe([Carvalho, Polson, and Scott(2010)]).

In proposed Dirichlet-Laplace prior setting, instead the single global scale τ, we use a vector of scales $(\phi_1\tau, \ldots, \phi_n\tau)$, where $\phi = (\phi_1, \ldots, \phi_n)$ satisfy $\{\phi_j \geq 0, \sum_{j=1}^{n} \phi_j = 1\}$, and is assigned a Dirichlet density prior, $\phi \sim Dir(a, \ldots, a)$. Additionally, we assume τ follows a Gamma density prior, $\tau \sim Gamma(\lambda, 1/2)$, with $\lambda = pa$, p is the number of covariates. Then the full DL_a prior can be represented as

$$\theta_i \sim N(0, \psi_i\phi_i^2\tau^2), \psi_i \sim Exp(1/2), \phi \sim Dir(a, \ldots, a), \tau \sim Gamma(na, 1/2)$$

The posterior sampler cycles can be obtained based on (a) $\theta|\psi, \phi, \tau, y$ and (b) $\psi, \phi, \tau|\theta$, holds with the fact that $\psi, \phi, \tau|\theta$ is independent of y. Also, we have $(\psi, \phi, \tau|\theta)$
$= (\psi|\phi, \tau, \theta)(\tau|\phi, \theta)(\phi|\theta)$, so the complete cycle to get posterior sampler is: (1) draw $\theta|\psi, \phi, \tau, y$, by sample θ_j independently follows $N(\mu_j, \sigma_j^2)$, here $\mu_j = \frac{y_j}{1+1/(\psi_j\phi_j^2\tau^2)}$ and $\sigma_j^2 = \frac{1}{1+1/(\psi_j\phi_j^2\tau^2)}$ (2) draw $\psi|\phi, \tau, \theta$, by sample ψ_i independently from $\psi_i \sim giG(\frac{1}{2}, 1, \frac{\theta_i^2}{\phi_i^2\tau^2})$ (3) sample $\tau|\phi, \theta$ from $\tau \sim giG(pa-n, 1, 2\sum_{i=1}^{p} \frac{|\theta_i|}{\phi_i})$ (4) sample T_i independently from $T_i \sim giG(a-1, 1, 2|\theta_i|)$, and $T = \sum_i T_i$, then $\phi_i|\theta$ will has the same distribution with T_i/T. Note, $Y \sim giG(\lambda, \rho, \chi)$ if $f(y) \propto y^{\lambda-1}e^{-0.5(\rho y+\chi/y)}$ for $y > 0$.

[Bhattacharya et al.(2014)] explained specifically the process to derive these posterior densities, also proved the minimax rate of convergence on posterior contraction with appropriate choice of the Dirichlet prior parameter a. With Gibbs samples, we can get posterior estimation and Bayesian credible region.

2.2.2 Shrinking and diffusing prior(SD)

This part is according to paper Bayesian variable selection with shrinking and diffusing priors([Narisetty and He(2014)]). This paper worked on model with spike and slab prior,

and calculate the posterior probability of each covariate included in the model.

A natural assumption in high dimensional settings is that the regression function is sparse, only a small number of covariates have nonzero coefficients. If one covariate has nonzero coefficient, it is active in the model. The purpose of this paper is to develop a Bayesian methodology for selecting the active covariates that is asymptotically consistent and computationally convenient. If the selected model equals the true model with probability converging to one, this is called selection consistency. In Bayesian methods, if the posterior probability of the true model converges to one, then it is referred as strong selection consistency by [Bondell and Reich(2012)].

In the hierarchical model, we can place prior distributions on the regression coefficients, also we can put prior on model space. The linear regression model considered is $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1}+\epsilon_{n\times 1}$, the subscripts is used to specify the dimension. We introduce latent binary variables for each of the covariates to be denoted by $Z = (Z_1, \ldots, Z_p)$. Each Z_i indicates whether the i-th covariate X_i is active in the model or not. The prior distribution on the regression coefficient β_i under $Z_i = 0$ is a point mass at zero, but a diffused prior under $Z_i = 1$. The concentrated prior of β_i under $Z_i = 0$ is referred as the spike prior, and the diffused prior under $Z_i = 1$ is called the slab prior. The prior on model space is applied by assuming a prior distribution on the binary random vector Z. A Bayesian variable selection method then selects the model with highest posterior probability. Different form of spike and slab priors yield different selection procedures.

In [Narisetty and He(2014)], shrinking and diffusing priors are introduced as spike and slab priors, and established strong selection consistency of the approach for $p = e^{o(n)}$. This approach is computationally advantageous because a standard Gibbs sampler can be used to sample from the posterior.

From here, we use p_n to denote the number of covariates to indicate that it grows with n. Specific model can be described as following:

$$
\begin{aligned}
&Y|(X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I), \\
&\beta_i|(\sigma^2, Z_i = 0) \sim N(0, \sigma^2 \tau_{0,n}^2), \\
&\beta_i|(\sigma^2, Z_i = 1) \sim N(0, \sigma^2 \tau_{1,n}^2), \\
&P(Z_i = 1) = 1 - P(z_i = 0) = q_n, \\
&\sigma^2 \sim IG(\alpha_1, \alpha_2)
\end{aligned}
$$

where i runs from 1 to p_n. Also, q_n, $\tau_{0,n}$ and $\tau_{1,n}$ are constants that depend on n. We use the posterior probabilities of the latent variable Z to identify the active covariates. Some threshold value can be set here. In general, if the posterior probability of $Z_i = 1$ is greater than 0.5, then covariate X_i will be included in the model.

In the simple case, we consider the case where the number of covariates $p_n < n$, and assume that the design matrix X is orthogonal, that is $X'X = nI$. We also assume σ^2 to be known. After some calculation, the posterior probability of Z_i can be obtained from

$$
P(Z_i = 0|\sigma^2, Y) = \frac{(1 - q_n) E_{\hat{\beta}_i}(\pi_0(B))}{(1 - q_n) E_{\hat{\beta}_i}(\pi_0(B)) + q_n E_{\hat{\beta}_i}(\pi_1(B))}
$$

here $\hat{\beta}_i$ is OLS estimator of β_i, and for $k = 0$ and 1,

$$
E_{\hat{\beta}_i}(\pi_k(B)) = \frac{1}{\sqrt{2\pi} a_{k,n}} exp\left\{ -\frac{\hat{\beta}_i{}^2}{2a_{k,n}^2} \right\}
$$

with $a_{k,n} = \sqrt{\sigma^2/n + \tau_{k,n}^2}$.

First, assume all the parameters are fixed. Fix $\tau_{0n}^2 = \tau_0^2 < \tau_{1n}^2 = \tau_1^2$ and $q_n = q = 0.5$. Then the limiting value of $P(Z_i = 1|\sigma^2, Y)$ will be less than 0.5 as $n \to \infty$. This implies that even as $n \to \infty$, we would not be able to identify the active coefficient in this case. Second, consider the case when shrinking $\tau_{0,n}^2$, fixed $\tau_{1,n}^2$ and q_n. $\tau_{0,n}^2$ goes to 0 with n. After some argument, we can get $P(Z_i = 0|\sigma^2, Y) \to^P I(\beta_i = 0)$. That is, for orthogonal design matrix, the marginal posterior probability of including an active covariate or excluding an inacive covariate converges to one under shrinking $\tau_{0,n}^2$ and fixed $\tau_{1,n}^2$ and q_n. However, this does not assure the consistency of overall model selection. Some arguments show that having $\tau_{1,n}^2$ and q_n fixed leads to inconsistency of selection if the number of covariates is much greater than $\sqrt{n}$. So, to get consistency of model selection, $\tau_{0,n}^2$ need to be shrinking, and $\tau_{1,n}^2$ need to be diffusing.

Under some conditions described in paper, we can get the consistency of model selection, $P(Z = t|Y, \sigma^2) \to^P 1$ as $n \to \infty$, that is, the posterior probability of the true model goes to 1 as the sample size increases to ∞. Here we do not need the true σ^2 to be known. Even for a misspecified $\tilde{\sigma}^2 \neq \sigma^2$, we can still have this consistency under some conditions.

Gibbs sampler can be obtained from the posterior distributions of parameters.

2.2.3 Coupled MH algorithm (CMH)

This part is based on paper on numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings([Johnson(2013)]). This paper works not only about how to get the marginal probability for each covariate included in model, but also the way to calculate the posterior probability of a specific model.

This method imposes nonlocal prior density on model parameters. Local prior density is positive at null parameter value, which is typically 0 in model selection settings. Nonlocal

prior density is function that are identically zero whenever a model parameter is equal to the null value. This paper shows model selection procedures based on nonlocal prior density assign a posterior probability of 1 to the true model as the sample size n increases when the number of possible covariates p is bounded by n and certain regularity conditions hold.

Linear model is of the form

$$y|\beta_k, \sigma^2 \sim N_n(X_k\beta_k, \sigma^2 I_n)$$

Here k denote a statistical model.

Two classes of nonlocal prior density was discussed. The first class of prior density for β is the product moment(pMOM) density, which is defined as

$$\pi(\beta|\tau, \sigma^2, r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2}|A_p|^{1/2}exp\left[-\frac{1}{2\tau\sigma^2}\beta' A_p\beta\right]\prod_{i=1}^{p}\beta_i^{2r}$$

Here d_p is normalizing constant, $\tau > 0$ is a scale parameter, which determines the dispersion of prior densities on β around 0. For a specific model k with number of covariates p, A_p is assumed to be the $p \times p$ identity matrix if no subjective information regarding the prior correlation between regression coefficients in model k, and r is called the order of density, can pick any positive integer.

For the second class of prior density, β is assumed to follow a product inverse moment (piMOM) density, which has the general form

$$\pi(\beta|\tau, \sigma^2, r) = \frac{(r\sigma^2)^{rp/2}}{\Gamma(r/2)^p}\prod_{i=1}^{p}|\beta_i|^{-(r+1)}exp\left(-\frac{\tau\sigma^2}{\beta_i^2}\right)$$

In piMOM density, $\tau > 0$ is a scale parameter explained the same as pMOM density, r

can take any positive integer.

These two density classes are nonlocal density at 0 because they are identically 0 when any component of β is 0. In model selection procedures, this is a good property, in the sense that it could efficiently eliminate regression models which contain any unnecessary explanatory variables.

For variance σ^2 known, the marginal density of the data can be expressed by

$$m_k(y_n) = \int p(y_n|\beta)p(\beta)d\beta.$$

If the variance σ^2 is not known, a common inverse gamma density is assumed for the value of σ^2. Then the marginal density of data under model k is

$$m_k(y_n) = \int\int p(y_n|\beta,\sigma^2)p(\beta)p(\sigma^2)d\beta d\sigma^2$$

In pMOM prior, the exact expressions for $m_k(y_n)$ can be obtained, see [Kan(2008)], even though the computational effort associated with the resulting expression increase exponentially with increasing model size. In piMOM prior, the analytic expression for $m_k(y_n)$ is not available. To fix these issues, Laplace approximation is recommended to approximate the marginal likelihood of the data $m_k(y_n)$ under each model.

Then, the posterior probability of a model t can be calculated by $p(t|y) = \frac{p(t)m_t(y)}{\sum_{k\in J} p(k)m_k(y)}$ based on the approximations of marginal density, assume the prior of a model $p(k)$ follows a beta function. The model space J has 2^p dimensions, which makes it impossible to compute the marginal density for all possible models when p is large. So a Markov chain Monte Carlo(MCMC) scheme is applied to obtain posterior samples of model from the model space. Metropolis-Hastings algorithm is implemented in the scheme to decide whether to update the

model for a new added covariate. Based on the posterior samples of model from model space, we can pick the model with highest posterior probability, also get the posterior probability for other probable models. After selecting the model, we can use ordinary least square estimation to estimate the coefficient.

2.2.4 Remark on methodology

First, we compare each method by their theoretical properties.

Three methods all focus on high-dimensional problems. While DL emphasize on estimating the coefficients of each covariate, and shrink some covariates into zero during estimation through shrinking prior. SD can calculate the posterior probability of each covariate included in the model, and set a threshold for the posterior probability to decide whether certain covariate should be in the true model. And CMH calculate the posterior probability for all possible models, then the model with highest posterior probability will be the resulted model.

There are some desired properties for these methods. SD has been proved with consistency in model selection, that is the posterior probability of the true model goes to 1 as the sample size increases to infinity. Minimax optimal rate of posterior contraction for DL has been established in [Bhattacharya et al.(2014)]. [Johnson and Rossell(2012)] also showed that CMH method can consistently select the true model when $p < n$.

SD allows dimension increases exponentially with sample size, that is $logp_n = o(n)$, and still has model selection consistency, this is a very desirable condition. Consistency on model selection for CMH does not hold when $p > n$, but the algorithm can be applied in settings $p \gg n$. Until now, DL only develop optimal minimax convergence rate for normal means problem, that is $p = n$, later we will show the convergence rate also holds for $p = O(n^{2-\epsilon})$.

Then, if we ponder over the process of each method, we can get Gibbs sampler from the process of DL and SD, if use the median of Gibbs sampler as the estimation of coefficient, we are applying the same prior for both model selection and parameter estimation. This is the one-stage method. While for CMH, we first select model based on pMOM(or piMOM) prior, then apply least square for estimation of coefficients. Least square estimation is the same as applying flat prior in Bayesian way. So CMH is a two-stage method in model selection and estimation. We may get the hint that it has some advantage due to its two-stage setting. And we will show this is the case in application results in Section 2.4.

In addition, we try to explain how to choose hyper-parameters in each model.

There are different parameters in prior setting for each method. For these parameters, some are default setting, some are recommended in paper. Here is a summary on how to choose the parameters.

In DL setting, the local variance $\psi_i \sim f(\psi_i)$, the default distribution for this f here is exponential distribution with mean 1/2. For vector $\phi \sim Dir(a, \ldots, a)$, $a = 1/n$. The global variance $\tau \sim g(\tau)$, the default setting for g is Gamma distribution with parameters $(na, 1/2)$.

In SD setting, for the spike prior variance term τ_{0n}^2, it is suggested use $\frac{\hat{\sigma}^2}{10n}$ to apply. While for the slab prior variance term τ_{1n}^2, $\hat{\sigma}^2 max\left(\frac{p_n^{2.1}}{100n}, logn\right)$ is proposed to plug in. Here $\hat{\sigma}^2$ is the sample variance of response vector Y, and choose $q_n = P[Z_i = 1]$ such that $P[\sum_{i=1}^{p}(Z_i = 1) > K] = 0.1$, and default value for K is $max(10, log(n))$. As stated in section 2.2.2, even for a misspecified $\tilde{\sigma}^2 \neq \sigma^2$, we can still have the consistency, so the choice of α_1 and α_2 is trivial.

In CMH setting, the prior for σ^2 is inverse Gamma distribution, $\sigma^2 \sim IG(10^{-3}, 10^{-3})$ is proposed. And for model k, $p(k) \sim B(k + a, p - k + b)$, here $B(.,.)$ is the beta function, and default value of a and b recommended by [Scott and Berger.(2010)] are $a =$

$b = 1$, p is the number of covariates. The recommended value of τ has been proposed in [Johnson and Rossell(2010)]. In practice, calculating inverse of $X_k'X_k + \frac{1}{\tau}A_k$ is involved, so we need to adjust the value of τ when $X_k'X_k$ is singular.

2.3 Extension of DL to linear model

In [Bhattacharya et al.(2014)], Dirichlet-Laplace priors are proposed for normal means problem. It said most of the ideas developed in this paper generalize directly to high-dimensional linear and generalized linear models. We try to extend the whole process to linear regression model. Following arguments also hold when $p > n$.

Linear model with Dirichlet-Laplace prior can be written as

$$
\begin{aligned}
& y = X\beta + \epsilon, \epsilon \sim N(0, I_n) \\
& \beta_i \sim N(0, \psi_i \phi_i^2 \tau^2), i = 1, \ldots, p \\
& \psi_i \sim exp(1/2), i = 1, \ldots, p \\
& \phi \sim Dir(a, \ldots, a) \\
& \tau \sim Gamma(pa, 1)
\end{aligned}
\tag{2.1}
$$

Denote $\Sigma = \begin{pmatrix} \psi_1\phi_1^2\tau^2 & 0 & \cdots & 0 \\ 0 & \psi_2\phi_2^2\tau^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p\phi_p^2\tau^2 \end{pmatrix}$

Then conditional distribution can be expressed by following:

(a)

$$\begin{aligned}\beta|\psi,\phi,\tau,y &\propto exp\{-\frac{1}{2}(y-X\beta)'(y-X\beta)\}exp\{-\frac{1}{2}\beta'\Sigma^{-1}\beta\}\\ &\sim MVNorm\left((\Sigma^{-1}+X'X)^{-1}X'Y,(\Sigma^{-1}+X'X)^{-1}\right)\end{aligned}$$

(b)

$$\begin{aligned}\psi_i|\phi_i,\tau,\beta_i &\propto p(\beta_i|\phi_i,\tau,\psi_i)*p(\psi_i)\\ &\propto (\psi_i)^{-1/2}exp\left\{-\frac{1}{2}\left(\frac{\beta_i^2}{\phi_i^2\tau^2}*\frac{1}{\psi_i}+\psi_i\right)\right\}\\ &\sim giG(\frac{1}{2},1,\frac{\beta_i^2}{\phi_i^2\tau^2})\end{aligned}$$

Note: generalized inverse Gaussian(giG) distribution $Y\sim giG(\lambda,\rho,\chi)$ if we have $f(y)\propto y^{\lambda-1}e^{-0.5(\rho y+\chi/y)}$ for $y>0$.

(c)

$$\begin{aligned}\tau|\phi,\beta &\propto p(\beta|\phi,\tau)*p(\tau)\\ &\propto \tau^{-p+pa-1}exp\left\{-\frac{1}{2}\left((2\sum_i\frac{|\beta_i|}{\phi_i})\frac{1}{\tau}+\tau\right)\right\}\\ &\sim giG(pa-p,1,2\sum_{i=1}^{p}\frac{|\beta_i|}{\phi_i})\end{aligned}$$

(d) $\phi|\beta$ has the same distribution with $T_1/T,\ldots,T_n/T$, where $T_i\sim giG(a-1,1,2|\beta_i|)$ independently and $T=\sum_{i=1}^{p}T_i$.

In before process, steps (b), (c) and (d) are exactly the same as normal means problem. But in step (a), we draw β from multi-variate normal distribution.

While in this multi-variate normal distribution, the covariance matrix is written as $\left(\Sigma^{-1} + X'X\right)^{-1}$. Rewrite $(\Sigma^{-1} + X'X)^{-1}$ as the form of $(I + \Sigma X'X)^{-1}\Sigma$, this matrix can work well even $X'X$ is singular.

In [Bhattacharya et al.(2014)], the property is developed for normal means problem. Under some restrictions on model means $\|\theta_0\|$, the posterior arising from the DL setting in [Bhattacharya et al.(2014)] contracts at the minimax rate of convergence for appropriate choice of Dirichlet concentration parameter a. Here, we can get similar result for linear regression model. If we put some conditions on design matrix, we can also get the mimimax optimal rate of posterior contraction, which means the posterior concentrates most of its mass on a ball around β_0 of squared radius of the order of $q_n log(p/q_n)$.

Theorem 1 *Consider model (2.1) where $a = p^{-(1+\beta)}$ for some $\beta > 0$ small. Assume $\beta_0 \in l_0[q_n; p]$ with $q_n = o(n)$ and $\|\beta_0\|_2^2 \leq q_n log^4 p$. Also, for design matrix X, suppose the elements in X are bounded, that is, there is a constant K, so that $max_{i,j} X_{i,j} \leq K$. And the dimension of β is within the order of $n^{2-\epsilon}$, i.e. $p = O(n^{2-\epsilon})$ for any $\epsilon > 0$. Then, with $s_n^2 = q_n log(p/q_n)$ and for some constant $M > 0$,*

$$lim_{n\to\infty} E_{\beta_0} P(\|\beta - \beta_0\|_2 < M s_n | y) = 1 \tag{2.2}$$

If $a = 1/p$ instead, then (2.2) holds when $q_n \succeq logn$.

2.4 Simulation and application

To compare the performance of these three methods, we show the results from some simulation study and application result on real data analysis.

2.4.1 Simulation on normal means problem

In this part, we investigate the performance of different methods when estimating the normal means problem. In each setting, model is $y_i = \theta_i + \epsilon_i$, $\epsilon_i \sim N(0,1), i = 1, \ldots, n$, suppose $n = 100$, and the true model size is 10. That is y sampled from a $N_{100}(\theta_0, I_{100})$ distribution, with θ_0 having 10 non-zero entries which are all set to be a constant $A > 0$. We choose different values of A, $A = 0.75, 1.5, 3, 4, 5, 6, 7, 8$. But for the simplicity of table, we only show the result for $A = 0.75, 1.5, 4, 7, 8$ in this part. We have 50 replicates for each case, and compare the average squared error and average absolute error in the table. The result is shown in table 2.1.

Table 2.1a shows the estimation result based on Gibbs sampler median. Squared error loss in the table is squared deviance between the posterior median and true value of coefficients, and take the average across the 50 replicates, the expression for squared error can be written as $\frac{\sum_{i=1}^{50}\left(\sum_{j=1}^{100}(\hat{\theta}_{ij}-\theta_{ij})^2\right)}{50}$. While the absolute error loss is the table is the absolute deviance between the estimator and the true parameters, also averaged across 50 replicates, absolute error can be expressed by $\frac{\sum_{i=1}^{50}\left(\sum_{j=1}^{100}|\hat{\theta}_{ij}-\theta_{ij}|\right)}{50}$. To better understand the source of error, we divide the squared error into two parts. The first part error comes from those parameters with nonzero true coefficients, that is $\frac{\sum_{i=1}^{50}\left(\sum_{j=1}^{10}|\hat{\theta}_{ij}-\theta_{ij}|\right)}{50}$. The second part error is from parameters with zero true coefficients, which is $\frac{\sum_{i=1}^{50}\left(\sum_{j=10}^{100}|\hat{\theta}_{ij}-\theta_{ij}|\right)}{50}$. We denote them respectively as sq.error1 and sq.error2 in the table. In the same way, we divide the absolute error in two parts, denote as abs.error1 and abs.err2 in table.

From the result in table 2.1a, we can see some interesting results. For the error calculated in DL, all the error comes from error1, that is the error for covarites with nonzero coefficients, and the error from zero coefficients are all zero. This implies DL could always shrink the

Table 2.1: Summary of error on estimation in normal means model. In each table, true model contains 10 nonzero constant entries, with value equals A. The number of total entry is 100.

(a) Summary of error on estimation based on Gibbs sampler median by each method. sq.error denotes for squared error of estimation. abs.error denotes for absolute error of estimation. sq.error1 represents for error from actually nonzero covariates, sq.error2 is error from actually zero covariates.

	A	0.75	1.5	4	7	8
DL	sq.error	5.625	22.500	122.659	26.276	14.233
	sq.error1	5.625	22.500	122.659	26.276	14.233
	sq.error2	0	0	0	0	0
	abs.error	7.500	15.000	32.593	11.289	9.376
	abs.error1	7.500	15.000	32.593	11.289	9.376
	abs.error2	0	0	0	0	0
SD	sq.error	5.752	24.883	24.389	8.733	8.131
	sq.error1	5.234	18.735	24.389	8.733	8.131
	sq.error2	0.518	6.1478	0	0	0
	abs.error	7.905	17.188	11.504	7.081	8.452
	abs.error1	7.127	13.090	11.504	7.081	8.452
	abs.error2	0.778	4.098	0	0	0
CMH	sq.error	5.625	22.500	18.411	5.643	5.540
	sq.error1	5.625	22.500	8.336	5.643	5.540
	sq.error2	0	0	10.075	0	0
	abs.error	7.500	15.000	11.120	5.890	5.584
	abs.error1	7.500	15.000	7.946	5.890	5.584
	abs.error2	0	0	3.174	0	0

(b) Squared error based on least square estimation by each method

A	0.75	1.5	4	7	8
DL	5.625	22.500	157.865	53.301	7.092
SD	13.591	28.501	12.451	5.643	5.540
CMH	5.625	22.5	18.411	5.643	5.540

(c) Variable selection performance for each method. FP represents for false positive rate. FN represents for false negative rate.

A	0.75		1.5		4		7		8	
Type	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
DL	0	10	0	10	0	9.86	0	0.98	0	0.12
SD	1.56	9.04	3.24	7.88	0.7	0	0	0	0	0
CMH	0	10	0	10	1	1	0	0	0	0

zero coefficients into exactly zero, which is a good property in high-dimension problem. While for SD, in any case, most of the error comes from error1. Compared to DL, SD yields smaller error1, which means SD gives better estimation for those nonzero coefficients. When θ is 4 or greater, CMH gives good estimate. Since it can pick up the correct covariates when θ is 7 or 8, its error all comes from error1. Obviously OLS gives good result with true nonzero covariates.

Note when the true parameters are around $A = 4$, the squared error is kind of large for all methods. It seems when the true parameter is around 4, this signal is not strong enough to be identified by this method, while the estimation result is still shrinking toward zero, makes the squared error large.

Since error1 and error2 here depend on which covarite is included in the model, we further see how each method performed when selecting the true model. When selecting variables, the ideal result should include all the nonzero coefficients in the model, while exclude all the zero coefficients. While in application each method inevitably make some mistakes. Table 2.1c summarize the variable selection result. In this table, FP represents for false positive rate, which means estimate a zero coefficient incorrectly as a nonzero one. FN stands for false negative rate, it is the mistake that estimate a nonzero coefficients into a zero one. Both FP and FN in this table is averaged across 50 replicates.

From table 2.1c, DL never estimates a zero coefficient as a nonzero one, while it shrinks all nonzero covariates into zero one when θ is small. In the case when θ is 0.75 or 1.5, it is quite weak signal compared to the noise variance which is 1. All three methods make obvious mistake when deciding which covariate should be included in the model. They just assume every covairate has zero coefficients. If θ is 7 or 8, any method could recognize the correct model in most cases, while CMH gives better estimation from the previous table.

We have explained the use of different priors for model selection and estimation in CMH, here we want to compare three methods accordingly, so we also applied least square estimate as in CMH method, and see how the estimation result will change. Table 2.1b gives the result.

From table 2.1b, the estimation does not improve, even get worse after applying flat prior for both DL and SD, since in most cases, for example when signal value is less than 7, either method has misspecified some predictors in the model. If the identified model is not the true model, then least square will give poor estimation on coefficient. In this case, DL and SD can give better estimation with the same prior as they applied in paper.

2.4.2 Simulation for linear regression case

As discussed in section 2.3, we can extend DL prior to linear model, even in the case when $p > n$. In this part, we first consider the linear regression case when $p < n$. The model is $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$, $\epsilon_i \sim N(0,1)$. Set $n = 100$, $p = 60$. For design matrix X, covariates X_i and X_j are standard normal with correlation given by $\rho^{|i-j|}$, $\rho = 0.5$. The true model size is 10. The true value of β contains 10 identical nonzero entries, and 50 zero entries. The value of non-zero entries are set to be $A = 0.75, 1.5, 4, 7, 8$. Table 2.2a summarized the result of squared error loss for estimation. For the full table with more information, see supplementary document.

From table 2.2a, CMH gives best estimation among the three. The result from SD is better than DL. If we look at the divided error, as long as signal coefficient greater than 1, the error of SD and CMH all comes from error1, that is the true parameters are not zero. All the error from zero coefficient part is zero, which means these two methods could always recognize the zero coefficients, or FP rate is zero. To check our speculation, we can see the variable selection performance, table 2.2c summarizes the result.

Table 2.2: Summary of error on estimation in linear regression model. In each table, the true model contains 10 non-zero constant entries, whose value equals A. The number of total possible covariates is 60.

(a) Squared error on estimation based on Gibbs sampler median by each method

A	0.75	1.5	4	7	8
DL	0.620	0.745	0.451	0.625	0.521
SD	0.204	0.176	0.332	0.252	0.389
CMH	0.201	0.117	0.077	0.085	0.187

(b) Squared error based on least square estimation by each method

A	0.75	1.5	4	7	8
DL	0.185	0.442	0.077	0.085	0.187
SD	0.132	0.117	0.077	0.085	0.187
CMH	0.201	0.117	0.077	0.085	0.187

(c) Variable selection performance for each method. FP represents for false positive rate. FN represents for false negative rate.

A	0.75		1.5		4		7		8	
Type	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
DL	0.6	0	1	0	0	0	0	0	0	0
SD	0	0	0	0	0	0	0	0	0	0
CMH	1	0	0	0	0	0	0	0	0	0

From table 2.2c, SD and CMH can recognize all the correct covariates even when $A = 1.5$, which is a desirable result. Combine the information in table 2.2a, all three methods give estimator pretty close to the true parameter value.

If we apply least square for all three methods, the squared error of estimation result is summarized in second part of table 2.2b.

From table 2.2b, the squared error loss is decreasing compared with the result based on Gibbs sampler, which means after applying flat prior, the estimation results improve toward the true value. Since in most cases, any method could pick up the correct model, then least square could give desirable results.

2.4.3 Simulation on high-dimension case

All these three methods are designed to solve high dimensional variable selection problems, so we want to check how they work in high-dimensional case. We compare the results of DL, SD and CMH when $p > n$. The basic settings are similar with before part. The model is $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$, $\epsilon_i \sim N(0, 1)$. In high-dimension case, p is greater than n. In simulation, we set two scenarios, $n = 100$, $p = 120$ and $n = 100$, $p = 200$. To construct design matrix X, covariates X_i and X_j are standard normal with correlation given by $\rho^{|i-j|}$, $\rho = 0.5$. For each case, the true model size is 10, and true value of β contains 10 identical nonzero entries, 110 zero entries in the case $p = 120$ and 190 zero entries in case $p = 200$. The value of non-zero entries are set to be $A = 0.75, 1.5, 4, 7, 8$. The result are summarized in table 2.3a. Each case has 50 replicates, and the squared error loss presented in table 2.3a are the average of squared error over 50 replicates.

In table 2.3a, whenever $p = 120$ or $p = 200$, estimation results are close to the true value, since all error in this table is quite small. If we look at the variable selection performance

Table 2.3: Summary of error on estimation in high-dimensional case. In each table, the true model contains 10 non-zero constant entries, whose value equals A. The number of total possible covariates is p.

(a) Squared error on estimation based on Gibbs sampler median by each method

$p = 120$	A	0.75	1.5	4	7	8
	DL	0.245	0.086	0.265	0.067	0.319
	SD	0.560	0.334	0.855	1.415	0.432
	CMH	0.132	0.147	0.258	0.088	0.225
$p = 200$	A	0.75	1.5	4	7	8
	DL	1.476	0.177	0.259	0.280	0.433
	SD	2.562	0.205	0.213	0.215	0.216
	CMH	0.246	0.070	0.141	0.266	0.189

(b) Squared error by least square estimation

$p = 120$	A	0.75	1.5	4	7	8
	DL	0.372	0.147	0.257	0.087	0.224
	SD	0.297	0.147	0.257	0.087	0.224
	CMH	0.132	0.147	0.257	0.087	0.224
$p = 200$	A	0.75	1.5	4	7	8
	DL	2.887	0.070	0.141	0.266	0.189
	SD	2.351	0.070	0.141	0.266	0.189
	CMH	0.246	0.070	0.141	0.266	0.189

result, all the FP and FN rate are zero for signal value $A \geq 1.5$, which means all three methods could recognize the true model.

While for $p = 200$, CMH can always pick out the correct predictors even when $A = 0.75$, this is an outstanding performance. Here DL and SD can identify the correct model in most cases when $A = 0.75$, the estimation is also comparable with the case for $p < n$. So all three methods give excellent performance in high-dimensional problems.

After applying least square estimation for estimation, table 2.3b are obtained for $n = 120$ and $n = 200$ respectively.

Compared table 2.3a with table 2.3b, the estimation improved slightly in some case, since the true model could always be identified.

2.4.4 Remark on simulation

Each method gives acceptable estimation and variable selection results. When the true nonzero signal is strong, like greater than 7 in simulation study, CMH will always be a good choice, since it can pick up the correct model, and give estimation with small error. If the nonzero signal is moderate, SD can be a good choice, it can correctly tell which covariate should be included in the model on most cases, also estimation is quite close to the true coefficients. If the signal is too weak, none of the methods could correctly estimate the model. DL works desirably in normal means problem and when we don't want too many covariates appeared in the model, since DL best shrinks all the zero parameters as zero, and it can be nicely explained in the context of relieving rely on one single global scale parameter.

Also, computation-consuming time is different for each method. If we make 8000 iterations, SD takes about 8 minumtes when n=100 and p=200, while it takes DL 16 minutes for the same setting, and 17 minutes for CMH. In the simulation study, SD always has advantage

in computing time.

2.4.5 Real data application

In this part, we apply these three variable selection method to a real data set to examine how they work in practice. We use the data from experiment to study the genetics of two inbred mouse populations. The data include expression levels of 22,575 genes of 31 female and 29 male mice resulting in a total of 60 arrays. Some physiological phenotypes are measured by quantitative real time PCR. The gene expression data and the phenotypic data are available at GEO(http://www.ncbi.nlm.nihgov/geo). Because this is an ultra-high dimensional problem with $p_n = 22,575$, we prefer to perform simple screenings of the genes first based on the magnitude of marginal correlations with the response. After the screening, the dataset for each of the responses consisted of $p = 200$ predictors (including the intercept and gender) by taking 198 genes based on marginal screening. We choose GPAT(glycerol-3-phosphate acyltransferase) as response. We performed variable selection with SD, DL and CMH.

We split the sample into a training set of 55 observations and a test set with the remaining five observations. We use the training set to get the fitted model, and predict the response in the test set.

By ordering the posterior inclusion probability for each method, we can list the highest 10 variables. For DL, we use the rank of Gibbs sampler median instead, since we are using Bayesian credible region to decide whether a covariate should be included in a model, the rank of Gibbs sampler median could provide information about the importance of a covariate. Table 2.4 lists 10 covariates with highest marginal inclusion probability for each method. The results are based on average rank of 30 replicates, the process is, first give the rank for each

Table 2.4: Order of covariates by highest posterior probability

DL	152, 149, 113, 25, 182, 183, 191, 139, 194, 125
SD	152, 149, 113, 191, 25, 183, 182, 34, 199, 56
CMH	152, 191, 194, 182, 199, 113,149, 25, 183, 69

replication, and then calculate the average rank among 30 replicates, which gives the result rank in table 2.4.

From table 2.4, variable with id 152 "1457715_*at*" has the highest posterior inclusion probability. Variables with id 25, 113, 149, 152, 182, 183, 191 are the common variables proposed by each method. The recommend model by CMH is size 2 model with covariates 152 and 194. If we see this as a variable selection problem, we can pick up the covariates included in model first, calculate the coefficients through LSE. We compare the mean squared prediction error for different model size with each method. We choose model size equals 2, 4, 6, 8 and 10. Since CMH gives the posterior probability of each model, for each model size, we pick out the model with highest posterior probability among certain model size.

It is an advantage for CMH to get the posterior probability of each model. To be comparable with other methods, we can also include those covariates with highest marginal posterior probability. So, we also calculate the MSPE including the covariates required by model size with highest marginal posterior probability.

Figure 2.1 compares MSPE obtained by each method. CMH1 represents model chosen from highest model posterior probability among certain model size. CMH2 represent for model including covariates with highest marginal posterior probability. From figure 2.1, we can see CMH gives best prediction if the model is chose based on model posterior probability. Instead of marginal inclusion probability for each covariate, it consider the probability of a whole model, this is advantageous since it combine all the covariates in a model.

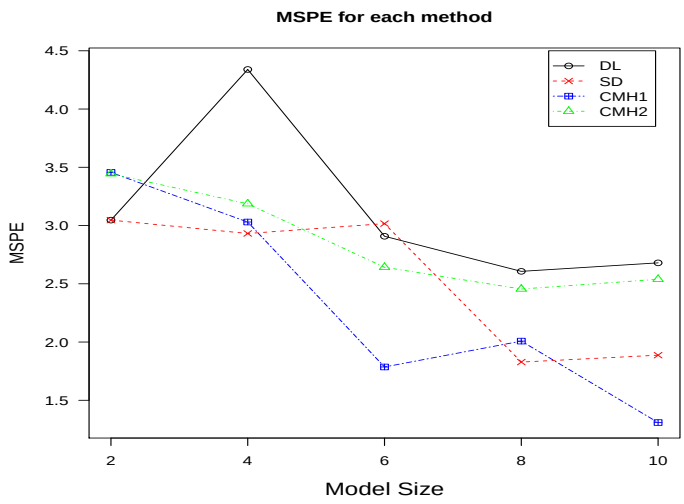

Figure 2.1: Plots of MSPE for each method. For the results shown in this figure, all the hyper-parameters are chosen by default value or recommended value as discussed in section 2.2.4. CMH1 is the MSPE resulted from choosing model with highest posterior probability among certain model size. CMH2 represents for model including covariates with highest marginal posterior probability.

On the other hand, we would like to know what will happen if we just use the median of posterior Gibbs sampler as the coefficient for SD and DL. The result is given in figure 2.2.

From figure 2.2, we can see prediction is not good if use the median of Gibbs sampler as estimation of coefficient, since either SD or DL is over shrinking the estimator, makes every coefficients close to zero. So from the application result, we may believe the model results from CMH method with highest posterior probability is more reliable according to the result of prediction.

2.4.6 Remark on application result

We can see that model with highest posterior probability based on CMH gives outstanding performance in most cases. And if SD or DL is employed, the recommendation is first choose the covariates, then apply least square estimation to get the model coefficients.

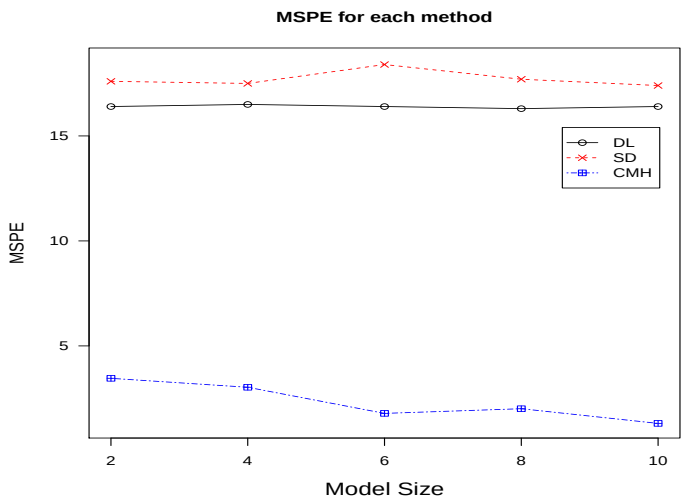

Figure 2.2: Plot of MSPE for each method. For the results shown in this figure, all the hyper-parameters are chosen by default value or recommended value as discussed in section 2.4.2. And estimation here use median of Gibbs sampler

To explain this further, in DL and SD, we assumed a prior for all the model parameters. And then we get the posterior distribution for each parameter, get Gibbs sample from each step, and obtain estimator from Gibbs sample. Since we draw the Gibbs sample under the assumption of prior, we are using the same prior for model selection also for parameter estimation. While this may be not the best idea, since for the aim of model selection, we are using shrinking prior. In CMH method, the Coupled MH algorithm is used for model selection, after we decide which model is selected, we use least square to estimate. Least square estimation is essentially we assume a flat prior for each parameter, this is more reasonable when we don't have evident information about parameters, and we know the selected model is the true model. From these argument, we can conjecture CMH has smaller estimation error as long as the correct model is selected.

Since CMH is using two-stage method to give estimation, while SD and DL just use one-stage for both model selection and estimation, this feature disadvantages SD and DL. If we also use two-stage like CMH method, suppose we use the same flat prior, that is the

least square result after we decide the model, then the estimation is comparable with CMH, like the result in first reveals.

2.5 Conclusion

In this chapter, we compared three different Bayesian methods for model selection, the comparison in both theory and application can give us some lights on how to work on high-dimensional problems in Bayesian perspective. Also, we extend the DL method into general linear regression case, and show the minimax convergence rate under some conditions. Each method has advantages in some sense, while how to develop a method that could be advantage in general may be an interesting topic.

2.6 Proof

This section shows proof sketch for Theorem.

The whole proof is based on the theorem of [Bhattacharya et al.(2014)]. First, we need to get a similar version of Theorem 3.2 [Bhattacharya et al.(2014)]. Follow the procedure in the paper, we can obtain the similar form of inequality (A.7). And then, by Lemma 5.2 in [Castillo and vander Vaart(2012)], we have $A'_n = \{D'_n \geq e^{-r_n^2} P(\|X\beta - X\beta_0\|_2 \leq r_n)\}$. If we have $max|X_{i,j}| \leq M$, with $t_n = r_n/M$, the same expression can be obtained as $A'_n = \{D'_n \geq e^{-r_n^2} P(\|\beta_0\|_2 \leq t_n)\}$ with $P_{\beta_0}\left(A_n^C\right) \leq e^{-r_n^2}$. Then follow the other steps, we can get the similar expression as in Theorem 3.2 [Bhattacharya et al.(2014)].

Then, follow the steps in proof of Theorem 3.1 in [Bhattacharya et al.(2014)]. Similar with previous argument, we can get $A_n = \{D_n \geq e^{-4r_n^2} P(\|\beta - \beta_0\|_2 \leq 2t_n)\}$ such that

$P_{\beta_0}(A_n^C) \leq e^{-r_n^2}$, here $t_n = r_n/M$. Then by constructing the net similarly, we can get

$$\|\beta^{S,j,i} - \beta\|_2^2 = \|\beta_S^{S,j,i} - \beta_S\|_2^2 + \|\beta_{S^C}\|_2^2 \leq (jr_n)^2 + (p - q_n)r_n^2/n^2 \leq 4j^2 r_n^2$$

, the last inequality holds if $p = O(n^{2-\epsilon})$ for any $\epsilon > 0$. Then we can finish the proof by similar argument in [Bhattacharya et al.(2014)] Proof of Theorem 3.1.

Chapter 3

Variable Selection by mixing spike and a nonlocal slab prior

3.1 Introduction

The literature of Bayesian variable selection is rapidly growing. Bayesian variable selection is equipped with natural measures of uncertainty, such as the posterior probability of each possible models and the marginal inclusion probabilities of each predictors. Given model with prior and likelihood, there are formal justifications for choosing a particular model.

Many Bayesian methods have been proposed for variable selection in recent years, including the stochastic search variable selection ([George and McCulloch (1993)]), empirical Bayes variable ([George and Foster(2000)]), penalized credible regions([Bondell and Reich(2012)]), nonlocal prior method ([Johnson and Rossell(2012)]), just to name a few. The spike and slab selection method proposed by [Mitchell and Beauchamp(1988)], then the method was further modified and developed by several authors, e.g., [Madigan and Raftery(1994)] and [George and McCulloch (1997)]. [Ishwaran and Rao (2005)] further generalized this model selection procedure with detail computational steps. Although the spike and slab prior has been surfacing in the literature for a while, very recently [Narisetty and He(2014)] developed model selection consistency under high-dimensional set up. Another notable development

of spike and slab prior was done by [Xu and Ghosh(2015)] in the context of bi-level selection who showed how to use spike and slab priors for selecting variables both at the group level as well as within a group.

Spike and slab prior assumed that the regression coefficients were mutually independent with a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). Several variations of spike and slab priors have been proposed in the literature. Zero inflated mixture priors have been utilized to a Bayesian approach for variable selection([Mitchell and Beauchamp(1988)]). [George and McCulloch (1997)] used zero inflated normal mixture priors in the hierarchical formulation for variable selection in linear model.

Spike and slab prior is an efficient method for variable selection. A latent binary variable is introduced for each of the covariates to be denoted by $Z = (Z_1, \ldots, Z_p)$. Z_i would indicate whether the i-th covariate is active in the model or not. Prior distribution on the regression coefficient β_i under $Z_i = 0$ is a point mass at zero, and a diffused prior is preferred under $Z_i = 1$. The concentrated prior of β_i under $Z_i = 0$ is called the spike prior, and the diffused prior under $Z_i = 1$ is called the slab prior. For slab prior, we would like to take a density which is flat and with heavy tails. Most commonly used slab prior is normal density with large standard deviation. While normal prior puts nontrivial density at point zero, this is overlapped with spike prior, which may cause some non-identifiablity issue.

In this chapter, a nonlocal prior is proposed as slab prior. Since nonlocal prior is zero when the parameter is zero, this could avoid overlap with spike prior, which is one natural desired property for spike and slab prior. Figure 3.1 are comparison between normal mixture spike and slab prior, with mixing spike and nonlocal slab prior. In extreme case, if the standard deviation for spike prior is closing to zero, then spike prior degenerates to a point mass

density at zero. In this chapter, we apply degenerated point mass density as spike prior, with nonlocal density as slab prior.

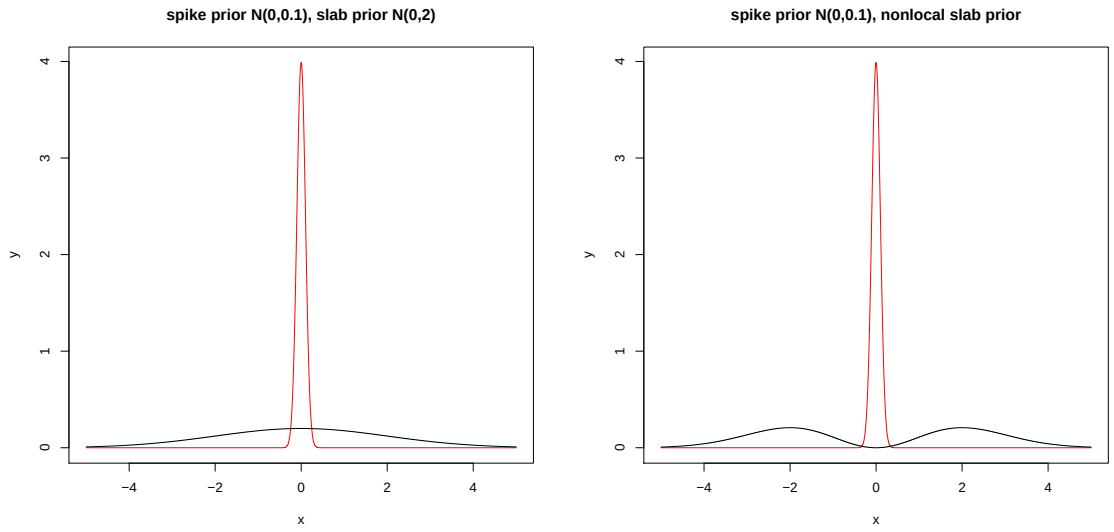

Figure 3.1: Comparison between normal-normal and normal-nonlocal mixture priors

3.2 Proposed model specification

We discussed the motivation for employing nonlocal prior as spike prior, which is symmetric bimodal density function, with function value closes to 0 whenever the parameter value approaches 0. Johnson (2010) proposed several forms of nonlocal densities, for example, product moment(pMOM) density, t-moment(tMOM) density, inverse moment(iMOM) density and so on. Among these, pMOM is derived from multivariate normal distribution with nonlocal properties. We use this pMOM density to explain nonlocal spike and slab prior model selection method. The functional form of pMOM density is

$$\pi(\beta|\tau,\sigma_1) = (2\pi)^{-1/2}(\tau\sigma_1^2)^{(-3/2)}exp\left(-\frac{\beta^2}{2\tau\sigma_1^2}\right)\beta^2 \tag{3.1}$$

Suppose τ and σ_1 are given hyperparameters. By denoting $(2\pi)^{-1/2}(\tau\sigma_1^2)^{(-3/2)}$ as C, the pMOM density can be expressed as

$$\pi(\beta|\tau,\sigma_1) = C * exp\left(-\frac{\beta^2}{2\tau\sigma_1^2}\right)\beta^2$$

Now, we will define the linear model specification. In linear model regression, we use P to denote the number of covariates, and response dimension is $n \times 1$, $n \times P$ design matrix corresponding to P covariates of interest. β stands for the regression coefficient vector. Also, we assume β is sparse in the sense that only a few components of β are non zero. The goal of variable selection in high dimensional data is dimension reduction, which is to identify the nonzero coefficients to explore the active covariates. The formal normal linear model is

$$Y_{n\times 1}|\beta_{P\times 1},\sigma^2 \sim N(X_{n\times P}\beta_{P\times 1},\sigma^2 I_n) \tag{3.2}$$

For each component in regression vector, that is for each j in $\{1,2,\ldots,P\}$, we assign a spike and slab prior for β_j, which is

$$\beta_j \sim Z_j\delta(\beta_j) + (1-Z_j)\pi(\beta_j) \tag{3.3}$$

where $\delta(\beta_j)$ is point mass when $\beta_j = 0$, and $\pi(\beta_j)$ is simplified notation for $\pi(\beta_j|\tau_j,\sigma_{1j})$, which has the nonlocal prior form in (3.1), written as

$$\pi(\beta_j|\tau_j,\sigma_{1j}) = (2\pi)^{-1/2}(\tau_j\sigma_{1j}^2)^{(-3/2)}exp\left(-\frac{\beta_j^2}{2\tau_j\sigma_{1j}^2}\right)\beta_j^2$$

In addition, we define the following prior distributions

$$Z_j \sim Bernoulli(p_0)$$

$$p_0 \sim Beta(a_1, b_1)$$

$$\sigma^2 \sim InverseGamma(\alpha_1, \alpha_2)$$

Denote β as the vector of $(\beta_1, \beta_2, \ldots, \beta_P)$ and Z as the vector of $(Z_1, Z_2, \ldots, Z_P)$. Under the above setting, we can write down the joint likelihood function as

$$\begin{aligned} p(\beta, Z, \sigma^2) \propto & \left(\sigma^2\right)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{p} X_{ip}\beta_p\right)^2\right\} \\ & \times \Pi_{j=1}^{P}\left[(Z_j)(2\pi)^{-1/2}(\tau_j\sigma_{1j}^2)^{(-3/2)} * exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 + (1-Z_j)\delta(\beta_j)\right] \\ & \times p_0^{a_1-1}(1-p_0)^{b_1-1} \times \left(\sigma^2\right)^{-\alpha_1-1} exp\left\{-\frac{\alpha_2}{\sigma^2}\right\} \end{aligned}$$

Here $Z_j = 0$ means β_j is excluded from model and $Z_j = 1$ means β_j included in the model.

3.2.1 Posterior median as thresholding estimator

In [Xu and Ghosh(2015)], a bi-level variable selection model with spike and slab prior is proposed, also the role of posterior median for thresholding is pointed out. Enlightened by them, I would like to propose similar posterior median with [Xu and Ghosh(2015)], and further analysis the property of proposed method based on posterior median estimator.

Consider the case when $p < n$ and orthogonal design matrix, with model defined by (3.2) and (3.3) with fixed $\tau_{j,n}$ and $\sigma_{1j,n}^2$, $j = 1, \ldots, P$. Here, we use subscript n in $\tau_{j,n}$ and $\sigma_{1j,n}^2$

to emphasize that τ_j and σ^2_{1j} depend on n for developing asymptotic theory. Under this model and assumptions, the marginal posterior distribution for β_j conditional on observed data is also a spike and slab distribution, which has the form of

$$\beta_j|Y,X \sim l_{j,n}\delta_0(\beta_j) + (1-l_{j,n})exp\left(\frac{\left(\beta_j - \left(1-B_{j,n}\right)|\hat{\beta}_j^{LS}|\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\beta_j^2 D_j \tag{3.4}$$

where $\hat{\beta}_j^{LS}$ is the least squares estimator of β_j, $B_{j,n} = \frac{\sigma^2}{\sigma^2+n\tau_{j,n}\sigma^2_{1j,n}}$, D_j is normalization constant and is calculated as

$$D_j = \frac{1}{\left(\left(\left(1-B_{j,n}\right)|\hat{\beta}_j^{LS}|\right)^2 + \frac{\sigma^2\left(1-B_{j,n}\right)}{n}\right)\sqrt{\frac{2\sigma^2\left(1-B_{j,n}\right)}{n}\pi}} \tag{3.5}$$

In addition, the posterior probability of $Z_j = 0$, which is the same as probability of $\beta_j = 0$ conditional on observed data can be calculated as

$$\begin{aligned} l_{j,n} &= P(\beta_j = 0|Y,X) \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)\left(1+n\tau_{j,n}\sigma^2_{1j,n}\right)^{-1/2} exp\left\{G_{j,n}\right\}\left(\frac{1}{2}+G_{j,n}\right)\left(\frac{2\sigma^2}{n\tau_{j,n}\sigma^2_{1j,n}+\sigma^2}\right)^{\frac{3}{2}}} \end{aligned} \tag{3.6}$$

where $G_{j,n} = \frac{\left(1-B_{j,n}\right)}{2\sigma^2} n\left(\hat{\beta}_j^{LS}\right)^2$. The specific expression and deduction for (3.4) and (3.6) can be seen in section 3.3.

Denote $F_{j,n}$ as cumulative function of $exp\left(\frac{\left(\beta_j - \left(1-B_{j,n}\right)|\hat{\beta}_j|^{LS}\right)^2}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\beta_j^2 D_j$, and quantile function of $F_{j,n}$ is defined as

$$Q_{j,n} = F_{j,n}^{-1}\left(max\left(0, \frac{0.5 - l_{j,n}}{1 - l_{j,n}}\right)\right) \tag{3.7}$$

Then, the resulting median of β_j, a soft thresholding estimator, can be given by

$$\hat{\beta}_j^{Med} = Med\left(\beta_j | Y, X\right) = sgn\left(\hat{\beta}_j^{LS}\right)\left(Q_{j,n}\right)_+ \tag{3.8}$$

3.2.2 Consistency

To further investigate the property of proposed thresholding estimator in (3.8), we assume orthogonal design matrix for the rest of this chapter, i.e., $X^T X = nI_P$. This assumption will simply the proof process in following theorem.

Let $\beta_1, \beta_2, \ldots, \beta_P$ denote the true coefficients value for P covarites respectively. Define the A as model index vector, with element value 1 if the corresponding covariate with nonzero coefficient, and with element value 0 if the corresponding covariate with zero coefficient. In other words, $A = \left(I\{\beta_j \neq 0\}\right)$ for $j = 1, 2, \ldots, P$. While selected model index vector by thresholding estimator $\hat{\beta}_j^{Med}$ in (3.8) is defined as $A_n^{Med} = \left(I\{\hat{\beta}_j^{Med} \neq 0\}\right)$ for $j = 1, 2, \ldots, P$. Model selection consistency is achieved if $lim_{n\to\infty} P\left(A_n^{Med} = A\right) = 1$. Following theorem states model selection consistency holds under very mild assumption.

Theorem 2 *Assume orthogonal design matrix, i.e., $X^T X = nI_P$. Suppose $\sqrt{n}\left(\tau_{j,n}\sigma_{1j,n}^2\right) \to \infty$ and $log\left(\tau_{j,n}\sigma_{1j,n}^2\right)/n \to 0$ as $n \to \infty$, for $j = 1, \ldots, P$, then the median thresholding estimator has variable selection consistency as*

$$lim_{n\to\infty} P\left(A_n^{Med} = A\right) = 1 \tag{3.9}$$

Theorem 2 states that we can select the true model based on threshold estimator with probability 1 for large enough sample size. Proof deduction is attached at section 3.6.

3.3 Gibbs samplers

In Bayesian variable selection methods, inference on parameter can be summarized from Gibbs sampler of posterior distribution. For special case, when prior distribution and posterior distribution are conjugate, which means they belong to the same density category, it is easier to update parameter value by analyzing posterior distribution. However, in many cases, posterior distribution is not conjugate with prior density, then we need to derive specific expression of posterior distribution.

In this part, we show the exact Gibbs samplers generating formula for each parameter. Specific deduction process is attached in last section.

3.3.1 Gibbs sampler for β_j

For each coefficient β_j, where j runs through $\{1, 2, \ldots, P\}$, posterior distribution also follows spike and slab distribution as a result of spike and slab prior. Then spike part would be point mass at zero, and slab part can be calculated from comprehensive integration. After calculation, the slab part of posterior distribution for β_j has following form:

$$\text{slab } \beta_j|rest \propto exp\left\{-A\left(\beta_j - \frac{B}{A}\right)^2\right\} \times \beta_j^2$$

where $A = \left(\frac{1}{2\tau_j\sigma_{1j}^2} + \frac{1}{2\sigma^2}\sum_{i=1}^n X_{ik}^2\right)$, $B = \frac{\sum_{i=1}^n X_{ij}\left(Y_i - \sum_{p\neq j} X_{ip}\beta_p\right)}{2\sigma^2}$, $C = \frac{1}{2\sigma^2}\sum_{i=1}^n \left(Y_i - \sum_{p\neq j} X_{ip}\beta_p\right)^2$

This is a non-symmetric and unknown density with a high peak around $\frac{B}{A}$. The normalization constant for this density can be obtained by doing integration. We can denote the density of $\left\{\text{slab } \beta_j | rest\right\}$ as DEN_j. Gibbs sampler from this density can be obtained by sampling scheme, similar to generate a random number from a distribution with cumulative distribution function information. Then the complete density for posterior distribution of β_k would be

$$\beta_j | rest \sim p_j\delta(\beta_j) + (1 - p_j)DEN_j$$

Here, p_j stands for the probability of β_j follows a point mass at zero density, which means the probability of $\beta_j = 0$.

3.3.2 Gibbs sampler for p_j

Since p_j is the probability of $\beta_j = 0$ for $j = \{1, 2, \ldots, P\}$, so it is critical when deciding which coefficient is significant in variable selection problem. In practice, we employed a latent variable Z_j corresponding to β_j, Z_j is generated from binomial distribution with success probability $\left(1 - p_j\right)$. So Z_j takes value either 0 or 1, $Z_j = 0$ implies $\beta_j = 0$, and $Z_j = 1$ implies β_j follows DEN_j. The calculation for p_j is as following. The specific

deduction process is attached in section 3.7.

$$p_j = P(Z_j = 0) = \frac{p_0}{p_0 + (1 - p_0) \times T}$$

$$T = K\,(2\pi)^{\frac{1}{2}} \left(\frac{\sigma^2}{\sum_{i=1}^{n} X_{ij}^2 + \frac{\sigma^2}{\tau_j \sigma_{1j}^2}} \right)^{\frac{1}{2}} \left[\left(\frac{\sigma^2}{\sum_{i=1}^{n} X_{ij}^2 + \frac{\sigma^2}{\tau_j \sigma_{1j}^2}} \right) + M^2 \right]$$

where $K = C\exp\left\{ \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} X_{ij}^2 + \frac{\sigma^2}{\tau\sigma_1^2} \right) M^2 \right\}$ and $M = \frac{\sum_{i=1}^{n} X_{ij} \left(Y_i - \sum_{p \neq j} X_{ip}\beta_p \right)}{\sum_{i=1}^{n} X_{ij}^2 + \frac{\sigma^2}{\tau_j \sigma_{1j}^2}}$.

3.3.3 Gibbs sampler for σ^2 and p_0

It is easier to obtain Gibbs sampler for error variance σ^2 and global proportion of nonzero coefficient p_0, since posterior are conjugate with prior densities for them. Posterior for σ^2 is also inverse gamma distribution

$$\sigma^2 | rest \sim InverseGamma \left(\frac{n}{2} + \alpha_1, \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \sum_{p} X_{ip}\beta_p \right)^2 + \alpha_2 \right)$$

While posterior for p_0 is still beta distribution

$$p_0 | rest \sim Beta \left(\#\left(\beta = 0\right) + a_1, \#\left(\beta \neq 0\right) + b_1 \right)$$

3.3.4 Comment on τ

Without any hands-on information, we can choose the same value of τ_j for $j \in 1, 2, \ldots, P$, that is use same prior for each coefficient. For simplicity, we use τ without subscript to denote this value is the same for every j. τ is an important tuning parameter for adjusting the prior

distributions. With large τ, the nonlocal slab prior is more flat, which puts majority density away from 0, and nonlocal density function approaches 0 when parameter value closes to zero, so large τ means strong penalty to small parameter value. This property helps avoiding selecting unnecessary covariates, however, it has the risk of missing important covaraites. On the other hand, with small τ, the prior distribution assigns more density to values around 0, this prior could easily detect small magnitude values, while too small τ value may result in over selection issue. It is critical to determine appropriate τ value such that we can balance the over selection issue and omit important covariates risk. From our empirical experiment, $\tau = 0.01$ could achieve satisfactory performance.

3.4 Simulation study

3.4.1 Eestimation performance by Mean Squared Errors

In linear regression case, model is $Y_{n\times 1} = X_{n\times P}\beta_{P\times 1} + \epsilon_{n\times 1}$, $\epsilon_i \sim N(0,1)$. Set $n = 100$, consider three different P value, $P = 60$, 120, and 200. For design matrix X, covariates X_i and X_j are standard normal with correlation given by $\rho^{|i-j|}$, $\rho = 0.3$. The true model size is 10. The true value of β contains 10 identical nonzero entries, and 50 zero entries. The value of non-zero entries are set to be $Coef = 1, 1.5, 4, 7$ respectively. For example, when $Coef = 1$, true nonzero coefficients are $(1, 1, \ldots, 1)_{10\times 1}$.

Table 3.1 summarized the result of mean squared error loss(MSE) for estimation. The simulation result also shows that there is no falsely selected covariate or omitted important covariate, so this mass-nonlocal prior could correctly identify the true model. Then MSE in table 3.1 are all caused by estimating the 10 nonzero coefficients, with definition $MSE = \frac{1}{10}\sum_{i=1}^{10}\left(\hat{\beta}_i - \beta_i\right)^2$. From table 3.1, we can notice that MSE are very small, which means

Table 3.1: MSE for n=100 with mass-nonlocal prior

Coef	1	1.5	4	7
dim=60	0.0635	0.0601	0.0766	0.0743
dim=120	0.0798	0.0767	0.0785	0.0824
dim=200	0.268	0.2475	0.2052	0.257

this method could give accurate estimation for coefficient value.

3.4.2 Selection performance

In this part, we report simulation results for different cases under several (n,p) combinations, signal strength and sparsity levels according to simulation setting at [Narisetty and He(2014)].

We will refer the proposed method as mass-nonlocal for the specification of spike and slab prior forms. Bayesian shrinking and diffusing prior setting in [Narisetty and He(2014)] is referred as BASAD. Other methods under comparison are piMOM with nonlocal prior of [Johnson and Rossell(2012)], SpikeSlab of [Ishwaran and Rao (2005)], and three penalization methods LASSO, elastic net(EN), and SCAD tuned by BIC, denote as LASSO.BIC, EN.BIC, SCAD.BIC respectively.

Case 1: We use sample size $n = 100$, and candidate dimension $P = n = 100$. The covariates are generated from multivariate normal distributions with zero mean and unit variance. The compound symmetric covariance with pairwise covariance $\rho = 0.25$ is used to represent correlation between covariates. Five covariates are taken active with coefficients $\beta = (0.6, 1.2, 1.8, 2.4, 3.0)$. Under this setting, covariats have moderate correlation and signal of coefficients are relatively strong.

Case 2: Consider scenario with $(n, P) = (100, 500)$, keep other parameters same as case 1.

Table 3.2: Case1: Performance of MASS-nonlocal for $n = P$. The other columns of the table are as follows: pp_0 and pp_1(when applicable) are the average posterior probabilities of inactive and active variables respectively; $Z = t$ is the proportion that the exact models is selected. $Z \supset t$ is the proportion that the selected model contains all the active covariates; FDR is the false discovery rate, and MSPE is the mean squared prediction error of the selected models.

	pp_0	pp_1	Z=t	$Z \supset t$	FDR	MSPE
(n,p)=(100,100), $\|t\|$ =5						
mass-nonlocal	**0.003**	**0.985**	**0.960**	**0.972**	**0.001**	**1.099**
BASAD	0.016	0.985	0.866	0.954	0.015	1.092
piMOM	0.012	0.991	0.836	0.982	0.030	1.083
SpikeSlab			0.005	0.216	0.502	1.660
LASSO.BIC			0.01	0.992	0.430	1.195
EN.BIC			0.398	0.982	0.154	1.134
SCAD.BIC			0.356	0.990	0.160	1.157
(n,p)=(200,200), $\|t\|$ =5						
mass-nonlocal	**0.001**	**1.000**	**0.996**	**1.000**	**0.001**	**1.028**
BASAD	0.002	1.000	0.944	1.000	0.009	1.087
piMOM	0.003	1.000	0.900	1.000	0.018	1.038
SpikeSlab			0.008	0.236	0.501	1.530
LASSO.BIC			0.014	1.000	0.422	1.101
EN.BIC			0.492	1.000	0.113	1.056
SCAD.BIC			0.844	1.000	0.029	1.040

Table 3.3: Case2: Performance of MASS-nonlocal for high-dimensional.

	pp0	pp1	Z=t	$Z \supset t$	FDR	MSPE
(n,p)=(100,500), $\|t\|$ =5						
mass-nonlocal	**0.007**	**0.967**	**0.682**	**0.876**	**0.054**	**1.152**
BASAD	0.001	0.948	0.730	0.775	0.011	1.130
SpikeSlab			0.000	0.040	0.626	3.351
LASSO.BIC			0.005	0.845	0.4661	1.280
EN.BIC			0.135	0.835	0.283	1.223
SCAD.BIC			0.045	0.980	0.328	1.260

Table 3.4: Case3: Performance of MASS-nonlocal for low signal.

	pp0	pp1	Z=t	$Z \supset t$	FDR	MSPE
(n,p)=(100,500), $\|t\|$ =5						
mass-nonlocal	**0.014**	**0.975**	**0.572**	**0.956**	**0.010**	**1.143**
BASAD	0.002	0.622	0.185	0.195	0.066	2.319
SpikeSlab			0.000	0.000	0.857	2.466
LASSO.BIC			0.000	0.520	0.561	1.555
EN.BIC			0.040	0.345	0.478	1.552
SCAD.BIC			0.045	0.340	0.464	1.561

Case 3: Keep design matrix covariance matrix and $|t| = 5$, with low signals

$\beta_t = (0.6, 0.6, 0.6, 0.6, 0.6)$.

MSPE stands for mean squared prediction error based on n (which equals 100 in case 2, case 3 and top part of case 1, 200 in bottom part of case 1) new observations as testing data. Simulation results are summarized in table 3.2, table 3.3 and table 3.4. Measurement index, include pp_0, pp_1, $Z = t$, $Z \supset t$, FDR and MSPE are defined in caption of table 3.2.

3.4.3 Observation from simulation

Table 3.2 shows result for case 1. We can see that three Bayesian methods mass-nonlocal, BASAD and piMOM perform better than other methods. Still mass-nonlocal prior is outperforming in the sense of $Z = t$, $Z \supset t$, FDR and MSPE. Result for case 2 can is at table 3.3, mass-nonlocal method still outperform among these methods in most of the index.

When signal is low, for example in case 3, the performance of mass-nonlocal is even impressive. Table 3.4 reveals simulation performance for case 3. All the other 5 methods have difficulty to identify the true model, the proportion of selecting the exact model for BASAD is 18.5%, while for the other four penalization method less than 5%.

Mass-nonlocal method is superior with 57.2%, which is more than three times reliable than BASAD. Also, mass-nolocal provides smallest FDR and MSPE.

3.4.4 Real data application

In this section, we apply the proposed variable selection method to a real data set to examine how it woks in practice. We consider the data from an experiment to study the genetics of two inbred mouse populations. The data include expression levels of 22,575 genes of 31 female and 29 male mice resulting in a total of 60 arrays. The numbers of phosphoenopyruvate carboxykinase(PEPCK) is measured by quantitative real-time PCR. The gene expression data and the phenotypic data are available at GEO(http://www.ncbi.nlm.nih.gov/geo;accession number GSE3330). [Narisetty and He(2014)] used this data for illustrating their method.

Following [Narisetty and He(2014)], we first performed simple screenings of the genes based on magnitude of marginal correlations with the response. [Fan and Lv(2008)] explained the power of marginal screening. After screening, the data set consist of $P = 118$ predictors (including the intercept and gender). We performed variable selection with mass-nonlocal along with BASAD. Following [Narisetty and He(2014)], the samples are randomly split into a training set of 55 observations and a test set with the remaining five observations. The fitted model using the training set were used to predict the response in the test set. This process was repeated 100 times to estimate the prediction power.

In Figure 3.2, we plot the average mean square prediction error(MSPE) for models of various size chosen by BASAD and mass-nonlocal. We can observe that MSPE of mass-nonlocal is mostly smaller than BASAD. About half the genes chosen by mass-nonlocal are overlapped with chosen result by BASAD.

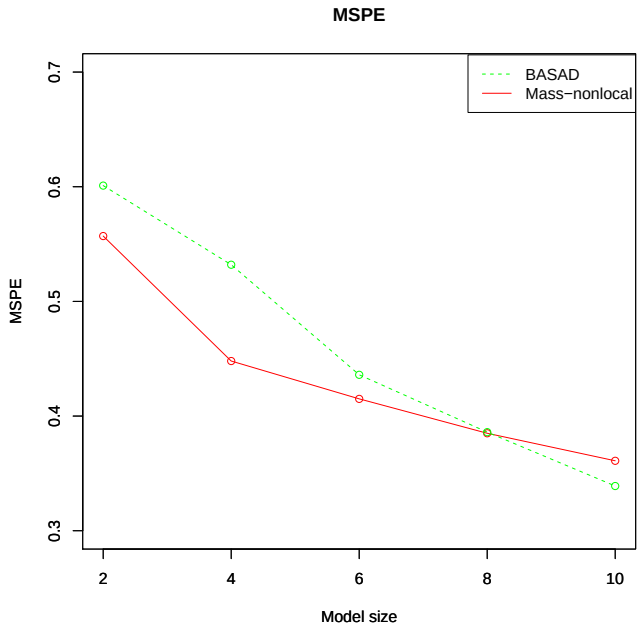

Figure 3.2: Mean squared prediction error versus model size for analyzing PEPCK.

3.5 Discussion

In this chapter, we proposed a mass-nonlocal form of spike and slab prior setting for variable selection in high dimensional data. This mass-nonlocal prior put point mass at 0 for spike prior, while employs nonlocal prior which avoid 0 point for slab prior, this property guaranteed superior performance on variable selection, also gives more accurate estimation compared with most other variable selection methods, which can be seen from simulation result.

However, variable selection performance is sensitive with respect to tuning parameter τ. We recommend tuning parameter value based on practical experiment. A rigorous method for proposing valid tuning parameter value could be further investigate in a future study. For example a variation of empirical Bayes can be developed.

In addition, from simulation results, we can see the superior performance of mass-nonlocal

prior in the sense of prediction since it yields much smaller mean prediction squared error. This could be explained by precise estimation of parameter value. In this chapter, we checked variable selection consistency property. We strongly believe that more strict property like oracle property should hold.

3.6 Proof of theorem

In this part, proof for theorem 2 will be provided. List of notations will be used in proof:

$\hat{\beta}_j^{Med}$ is defined in (3.8);

$F_{j,n}$ as cumulative function of $exp\left(\frac{\left(\beta_j-\left(1-B_{j,n}\right)|\hat{\beta}_j|^{LS}\right)^2}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\beta_j^2 D_j$;

$Q_{j,n}$ is defined in (3.7);

$F_{j,n}^{-1}$ is inverse function of $F_{j,n}$;

$l_{j,n}$ defined in (3.6);

Prove:

For j such that $|\beta_j| = 0$, since $\sqrt{n}\hat{\beta}_j^{LS} = O_p(1)$, and from assumption $n\left(\tau_{j,n}\sigma_{1j,n}^2\right) \to \infty$, we have $l_{j,n} \to 1$ as $n \to \infty$. The probability of correctly classifying this factor is

$$\begin{aligned} P\left(|\hat{\beta}_j^{Med}| = 0\right) &= P(Q_{j,n} \leq 0) \\ &\to 1 \end{aligned} \tag{3.10}$$

as $n \to \infty$, $l_{j,n} \to 1$, so $Q_{j,n} \to F_j^{-1}(0)$, which is negative with probability 1. Then (3.10) holds.

For j such that $|\beta_j| \neq 0$, since $\hat{\beta}_j^{LS} \to^p \beta_j^0$ and assumption $log\left(\tau_{j,n}\sigma_{1j,n}^2\right)/n \to 0$, we have

$l_{j,n} \to^p 0$ as $n \to \infty$. The probability of correctly identifying this factor is

$$P\left(|\hat{\beta}_j^{Med}| \neq 0\right) = P(Q_{j,n} > 0) \tag{3.11}$$

needs to show

$$Q_{j,n} = F_{j,n}^{-1}\left(max(0, \frac{0.5 - l_{j,n}}{1 - l_{j,n}})\right) > 0 \tag{3.12}$$

With $l_{j,n} \to 0$, we have $max\left(0, \frac{0.5-l_{j,n}}{1-l_{j,n}}\right) = \frac{0.5-l_{j,n}}{1-l_{j,n}}$, and $lim_{n\to\infty}\frac{0.5-l_{j,n}}{1-l_{j,n}} = \frac{1}{2}$, which implies $lim_{n\to\infty}F_{j,n}^{-1}\left(max(0, \frac{0.5-l_{j,n}}{1-l_{j,n}})\right) = F_j^{-1}\left(\frac{1}{2}\right)$, (3.12) is satisfied if we can show $t = F_j^{-1}\left(\frac{1}{2}\right) > 0$, which can be derived from

$$D_j \int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - \left(1 - B_{j,n}\right)|\beta_j|^{LS}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right)\beta_j^2 d\beta_j < \frac{1}{2} \tag{3.13}$$

where $D_j = \frac{1}{\left(\left(\left(1-B_{j,n}\right)|\beta_j|^{LS}\right)^2 + \frac{\sigma^2\left(1-B_{j,n}\right)}{n}\right)\sqrt{\frac{2\sigma^2\left(1-B_{j,n}\right)}{n}\pi}}$ from expression in (3.5).

Next, we show how (3.13) holds. Denote $H_{j,n} = \left(1 - B_{j,n}\right) |\beta_j|^{LS}$

$$
\begin{aligned}
&\int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) \beta_j^2 d\beta_j \\
&= \int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) \left(\beta_j - H_{j,n} + H_{j,n}\right)^2 d\beta_j \\
&= \int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) \left(\beta_j - H_{j,n}\right)^2 d\beta_j \\
&\quad + 2\left(H_{j,n}\right) \int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) \left(\beta_j - H_{j,n}\right) d\beta_j \\
&\quad + \left(H_{j,n}\right)^2 \int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) d\beta_j \\
&= \int_{-\infty}^{-H_{j,n}} exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) (u)^2 \, du \\
&\quad + 2\left(H_{j,n}\right) \int_{-\infty}^{-H_{j,n}} exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) (u) \, du \\
&\quad + \left(H_{j,n}\right)^2 \int_{-\infty}^{-H_{j,n}} exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1 - B_{j,n}\right)}\right) du
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{\sigma^2}{n}\left(1-B_{j,n}\right) exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) u\Big|_{-\infty}^{-H_{j,n}} \\
&\quad + \frac{\sigma^2}{n}\left(1-B_{j,n}\right)\int_{-\infty}^{-H_{j,n}} exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) du \\
&\quad - 2\left(H_{j,n}\right)\frac{\sigma^2}{n}\left(1-B_{j,n}\right) exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\Big|_{-\infty}^{-H_{j,n}} \\
&\quad + \left(H_{j,n}\right)^2\int_{-\infty}^{-H_{j,n}} exp\left(\frac{u^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) du \\
&= -\frac{\sigma^2}{n}\left(1-B_{j,n}\right) exp\left(\frac{\left(-H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\left(-H_{j,n}\right) \\
&\quad + \frac{\sigma^2}{n}\left(1-B_{j,n}\right)\sqrt{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)\pi} * \Phi\left(\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}}\right) \\
&\quad - 2\left(H_{j,n}\right)\frac{\sigma^2}{n}\left(1-B_{j,n}\right) exp\left(\frac{\left(-H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) \\
&\quad + \left(H_{j,n}\right)^2\sqrt{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)\pi} * \Phi\left(\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}}\right) \\
&= -\frac{\sigma^2}{n}\left(1-B_{j,n}\right)\left(H_{j,n}\right) exp\left(\frac{\left(H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) \\
&\quad + \left(\frac{\sigma^2}{n}\left(1-B_{j,n}\right) + \left(H_{j,n}\right)^2\right)\sqrt{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)\pi} \\
&\quad * \Phi\left(\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}}\right)
\end{aligned}
$$

Also,

$$
\begin{aligned}
& D\int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\beta_j^2 d\beta_j \\
=& \frac{\left(1-B_{j,n}\right)\left(H_{j,n}\right)}{\left(\left(H_{j,n}\right)^2 + \frac{\sigma^2\left(1-B_{j,n}\right)}{n}\right)\sqrt{\frac{2\sigma^2\left(1-B_{j,n}\right)}{n}\pi}}\left(-\frac{\sigma^2}{n}\right) exp\left(\frac{\left(H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right) \\
& + \Phi\left(\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}}\right)
\end{aligned}
$$

Since $\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}} < 0$, so $\Phi\left(\left(-H_{j,n}\right)\sqrt{\frac{1}{\frac{\sigma^2}{n}\left(1-B_{j,n}\right)}}\right) < \Phi\left(0\right) = \frac{1}{2}$, the first term is negative, which implies $D\int_{-\infty}^{0} exp\left(\frac{\left(\beta_j - H_{j,n}\right)^2}{\frac{2\sigma^2}{n}\left(1-B_{j,n}\right)}\right)\beta_j^2 d\beta_j$ always less than $\frac{1}{2}$, so (3.13) holds, then (3.12) holds as a result of (3.13). This proves theorem 2.

3.7 Additional: calculation of Gibbs samplers

3.7.1 Gibbs sampler for β_k

$$
\begin{aligned}
\text{slab } \beta_j|rest &\propto exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 \times exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\sum_p X_{ip}\beta_p\right)^2\right\} \\
&= exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\sum_p X_{ip}\beta_p\right)^2\right\}\beta_j^2 \\
&= exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\sum_{p\neq j} X_{ip}\beta_j-X_{ik}\beta_j\right)^2\right\}\beta_j^2 \\
&= exp\Big\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{n}\left(Y_i-\sum_{p\neq j} X_{ip}\beta_p\right)^2+\sum_{i=1}^{n}\left(X_{ij}^2\right)\beta_j^2 \\
&\quad -2\sum_{i=1}^{n}X_{ij}\left(Y_i-\sum_{p\neq j}X_{ip}\beta_p\right)\beta_j\Big)\Big\}\beta_j^2 \\
&= exp\Big\{-\left(\frac{1}{2\tau_j\sigma_{1j}^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{n}X_{ij}^2\right)\beta_j^2+2\frac{\sum_{i=1}^{n}X_{ij}\left(Y_i-\sum_{p\neq j}X_{ip}\beta_p\right)}{2\sigma^2}\beta_j \\
&\quad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\sum_{p\neq j}X_{ip}\beta_p\right)^2\Big\}\beta_j^2
\end{aligned}
$$

Denote $A = \left(\frac{1}{2\tau_j\sigma_{1j}^2}+\frac{1}{2\sigma^2}\sum_{i=1}^{n}X_{ij}^2\right)$, $B = \frac{\sum_{i=1}^{n}X_{ij}\left(Y_i-\sum_{p\neq j}X_{ip}\beta_p\right)}{2\sigma^2}$, $C = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\sum_{p\neq j}X_{ip}\beta_p\right)^2$

$$
\begin{aligned}
\text{slab } \beta_j|rest &\propto exp\left\{-A\beta_j^2+2B\beta_j-C\right\}\beta_j^2 \\
&= exp\left\{-A\left(\beta_j-\frac{B}{A}\right)^2\right\}\times\beta_j^2\times exp\left\{\frac{B^2}{A}-C\right\}
\end{aligned}
$$

Denote slab $\beta_j|rest$ as DEN_j,so

$$\beta_j|rest \sim p_j\delta(\beta_j) + (1-p_j)DEN_j$$

3.7.2 Gibbs sampler for p_j

$$\begin{aligned}
p_j &= P(Z_j = 0) = \frac{p_0}{p_0 + (1-p_0)\times T} \\
T &= \frac{C}{exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{p\neq j} X_{ip}\beta_p\right)^2\right\}} \int exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{p} X_{ip}\beta_p\right)^2\right\} \\
&\qquad exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 \\
&= M\int exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{p\neq j} X_{ip}\beta_p - X_{ij}\beta_j\right)^2\right\} exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 \mathrm{d}\beta_j \\
&= C\int exp\left\{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{n} X_{ij}^2\right)\beta_j^2 - 2\left(R_j\right)\beta_j\right]\right\} exp\left\{-\frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 d\beta_j \\
&= C\int exp\left\{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{n} X_{ij}^2\right)\beta_j^2 - 2\left(R_j\right)\beta_j\right] - \frac{1}{2\tau_j\sigma_{1j}^2}\beta_j^2\right\}\beta_j^2 d\beta_j
\end{aligned}$$

where $R_j = \sum_{i=1}^{n} X_{ij}\left(Y_i - \sum_{p\neq j} X_{ip}\beta_p\right)$.

Denote $N_j = \sum_{i=1}^{n} X_{ij}^2 + \frac{\sigma^2}{\tau_j \sigma_{1j}^2}$, $M = \frac{\sum_{i=1}^{n} X_{ij}\left(Y_i - \sum_{p \neq j} X_{ip}\beta_p\right)}{N_j}$, then

$$\begin{aligned}
T &= C \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left[\beta_j^2 - 2M\beta_j\right]\right\}\beta_j^2 d\beta_j \\
&= C \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left[\left(\beta_j - M\right)^2 - M^2\right]\right\}\beta_j^2 d\beta_j \\
&= C\exp\left\{\frac{1}{2\sigma^2}\left(N_j\right)M^2\right\} \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\beta_j^2 d\beta_j
\end{aligned}$$

Now denote $K = C\exp\left\{\frac{1}{2\sigma^2}\left(N_j\right)M^2\right\}$,

$$\begin{aligned}
T &= K \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left(\beta_j - M + M\right)^2 d\beta_j \\
&= K \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left[\left(\beta_j - M\right)^2 + 2M\beta_j - M^2\right] d\beta_j \\
&= K * (part1 + part2 + part3)
\end{aligned}$$

$$\begin{aligned}
part1 &= \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left[\left(\beta_j - M\right)^2\right] d\beta_j \\
&= (2\pi)^{\frac{1}{2}}\left(\sigma^2\frac{1}{N_j}\right)^{\frac{3}{2}}
\end{aligned} \tag{3.14}$$

Equation in (3.14) is obtained by integrating density of nonlocal MOM prior.

$$
\begin{aligned}
part2 &= \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left[2M\beta_j\right]d\beta_j \\
&= 2M\sqrt{\frac{2\pi\sigma^2}{N_j}} \times \frac{1}{\sqrt{\frac{2\pi\sigma^2}{N_j}}} \\
&\qquad \times \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left(\beta_j\right)d\beta_j \\
&= 2M^2\sqrt{\frac{2\pi\sigma^2}{N_j}}
\end{aligned}
$$

$$
\begin{aligned}
part3 &= \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}\left[-M^2\right]d\beta_j \\
&= -M^2\int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}d\beta_j \\
&= -M^2\sqrt{\frac{2\pi\sigma^2}{N_j}} \times \frac{1}{\sqrt{\frac{2\pi\sigma^2}{N_j}}} \\
&\qquad \times \int exp\left\{-\frac{1}{2\sigma^2}\left(N_j\right)\left(\beta_j - M\right)^2\right\}d\beta_j \\
&= -M^2\sqrt{\frac{2\pi\sigma^2}{N_j}}
\end{aligned}
$$

Then,

$$
\begin{aligned}
T &= K \times (part1 + part2 + part3) \\
&= K \times \left[(2\pi)^{\frac{1}{2}} \left(\sigma^2 \frac{1}{N_j} \right)^{\frac{3}{2}} + 2M^2 \sqrt{\frac{2\pi\sigma^2}{N_j}} - M^2 \sqrt{\frac{2\pi\sigma^2}{N_j}} \right] \\
&= K \times \left[(2\pi)^{\frac{1}{2}} \left(\sigma^2 \frac{1}{N_j} \right)^{\frac{3}{2}} + M^2 \sqrt{\frac{2\pi\sigma^2}{N_j}} \right] \\
&= K \, (2\pi)^{\frac{1}{2}} \left(\frac{\sigma^2}{N_j} \right)^{\frac{1}{2}} \left[\left(\frac{\sigma^2}{N_j} \right) + M^2 \right]
\end{aligned}
$$

Chapter 4

Model selection for generalized linear model using nonlocal priors

4.1 Introduction

In modern statistical practice, variable selection is one of the most commonly used technique, especially in clinical and genetic research, due to complex and high dimensional nature of data. A good amount of work has been done recently based on both frequentist and Bayesian perspective. This chapter concentrates on Bayesian model selection for generalized linear models, in particular the logistic regression based on a nonlocal prior distribution.

In classical statistics, many approaches have been proposed to deal with model selection problems. Some popular methods include F tests, Akaike information criterion ([Akaike (1973)]), Bayes information criterion ([Schwarz (1978)]), Mallows C_p, exhaustive search, stepwise, backward and forward selection procedures. Also, many frequentist methods based on penalization have been developed with good properties. Among these, least absolute shrinkage and selection operator (LASSO) method ([Tibshirani(1996)]) based on L_1 norm penalty is one of the most popular and proven to be effective model selection procedures. Elastic net ([Zou and Hastie (2005)]) is derived from the linear combination of L_1 and L_2 norm penalties. Smoothly clipped absolute deviation (SCAD) ([Fan and Li (2001)]) uses nonconcave penalty, which leads oracle property. Adaptive LASSO ([Zou (2006)]) using adaptive

weights for penalizing different coefficients in L_1 norm shows consistency under some conditions. Dantzig selector ([Candes and Tao (2007)]) is a solution to a L_1 regularization problem which also shows efficient convergence.

In generalized linear regression, calculation can be tedious because of the associated likelihood function is complicated. Earlier literature on variable selection in logistic regression used criterion based methods such as AIC or BIC. However, with some approximation technique, penalized methods can also be applied in generalized linear regression ([Van de Geer (2008)], [Huang et al. (2008)]). With the popularity of LASSO, fitting the generalized linear model with LASSO or elastic-net regularization path is proposed by [Friedman, Hastie and Tibshirani (2010)].

Classical statistical methods have some limitations, for example, lack of explanation, consistency or uncertainty estimation, which motivated the employment of Bayesian methods. Bayes factors and posterior probabilities are easy to understand, and Bayesian model selection is consistent if one of the entertained model is actually the true model and if enough data are observed. Also, Bayesian approach can account for model uncertainty. The solution for Bayesian methods can be equivalent to frequentist penalized approach under appropriate priors. All these advantages make Bayesian methods popular, which also benefit from the development of efficient computational algorithms. Many Bayesian methods have been proposed, like Bayesian LASSO ([Park and Casella (2008)]) using Laplacian shrinkage, Bayesian model average technique ([George (1999)]), spike and slab prior ([Ishwaran and Rao (2005)]) by introducing a latent variable, Gibbs variable selection ([Dellaportas et al. (1997)]) with a mixture prior assumed for each variable, stochastic search variable selection ([George and McCulloch (1993)]; [George and McCulloch (1997)]) with spike prior centered around zero and small variance.

When it comes to a generalized linear model, the computation problem arises again because of the intrinsically complicated likelihood function. The computation requirement remains with the Bayesian methods due to no-conjugacy. With the development of modern computational platforms and efficient algorithms, fully Bayes approach can be applied. For example, informative prior is proposed by [Chen, Ibrahim and Yiannoutsos (1999)]. [Nott and Leonte (2012)] discussed an efficient sampling algorithm. [Jiang (2006)] considered the logistic regression in which prior under some conditions leads to consistent convergence toward the true model. Then, the theory was extended for generalized linear models in [Jiang (2007)]. [Sha et al.(2004)] applied probit regression to classify binary responses based on microarray data. [Zhou, Liu and Wong(2004)] used Bayesian variable selection for logistic regression to achieve excellent cross-validated classification errors. On the other hand, Bayesian subset modeling is proposed by [Liang, Song and Yu (2013)]. Most of these Bayesian methods put a great mass on the density of a null parameter value to reach the result of shrinkage and go to variable selection results which is referred as model selection based on local prior densities by [Johnson and Rossell(2012)].

In this chapter we are interested in nonlocal prior ([Johnson and Rossell(2010)]). Nonlocal prior density is a density function that is identically zero whenever a model parameter is equal to its null value, typically 0 in model selection settings. Most current Bayesian model selection procedures employ local prior density, which is positive at null parameter values. Since nonlocal prior density would be zero if any component of parameter is zero, this property could bring a parsimonious model selection.

Bayesian model selection procedures for linear regression model by imposing a nonlocal prior density is proposed by [Johnson and Rossell(2012)]. They further summarized the comparison with popular local prior densities as well as with frequentist approaches. In

this chapter, we extend their approach to a generalized linear model, specifically logistic regression model. Main contribution of our work is employing nonlocal prior density for a logistic regression model and studying its validity. We prove the posterior convergence in high-dimensional setting along the line of [Jiang (2007)]. We specifically find the convergence rate for a logistic regression model with nonlocal prior density. The numerical results are promising.

The rest of the chapter is organized as follows: The section 4.2 discusses the proposed methodology. The section 4.3 discusses the algorithm for implementation and the section 4.4 provides the numerical results along with a real data example. The proofs are given in section 4.6.

4.2 Methodology

4.2.1 Bayesian logistic regression

Let μ be the success probability for a binary random variable Y, then the logistic regression is defined as $logit(\mu) = X\beta$, where X is the design matrix with dimension $n \times p_n$, β is the $p_n \times 1$ regression coefficient vector. The function $logit(z) = log\left(\frac{z}{1-z}\right)$ for $z \in (0,1)$. With n data points $\{(X_i, y_i)\}$ for $i = 1, \cdots, n$, the likelihood function is

$$p(y|\beta) = \prod_{i=1}^{n} \left(\frac{1}{1+\mathrm{e}^{-\sum_s X_{is}\beta_s}}\right)^{y_i} \left(\frac{1}{1+\mathrm{e}^{\sum_s X_{is}\beta_s}}\right)^{1-y_i}.$$

In Bayesian settings, we suppose β has a prior density $p(\beta)$, then the joint density of the data and β is $p(y,\beta) = p(y|\beta)p(\beta)$, and the marginal density of the data y is $p(y) = \int p(y|\beta)p(\beta)d\beta$. Since this is independent of β, we can denote the marginal density as Z.

The Bayesian inference is made from the posterior distribution of β which is given by

$$p(\beta|y) = \frac{1}{p(y)}p(y|\beta)p(\beta) = \frac{1}{Z}p(y|\beta)p(\beta).$$

Denote $F(\beta) = -\log p(y, \beta) = -\log(p(y|\beta)p(\beta))$. Then the posterior of β can be written as

$$p(\beta|y) = \frac{1}{Z}e^{-F(\beta)}.$$

Assume there exists a posterior mode β^*. Expanding $F(\beta)$ around β^* gives

$$F(\beta) \approx F(\beta^*) + (\beta - \beta^*)^T g(\beta^*) + \frac{1}{2}(\beta - \beta^*)^T H(\beta^*)(\beta - \beta^*), \tag{4.1}$$

where $g(\beta) = \frac{\partial F(\beta)}{\partial \beta}$ and $H(\beta) = \frac{\partial^2 F(\beta)}{\partial \beta \partial \beta^T}$. Both g and H are evaluated at the posterior mode β^*. Here β^* maximizes $-F(\beta)$, so the gradient $g(\beta^*)$ equals 0 and Hessian matrix $H(\beta^*)$ will be positive definite. Now we have

$$p(\beta|y) \approx \frac{1}{Z}e^{-F(\beta^*)}\exp\left\{-\frac{1}{2}(\beta - \beta^*)^T H(\beta - \beta^*)\right\}$$

so that the posterior of β is approximated by a normal distribution, $\mathrm{N}(\beta^*, H^{-1})$. This could be a general scheme for Bayesian analysis in logistic regression model to avoid complicated computational schemes such as Metropolis-Hastings algorithm ([Hoff (2009)]). Since we obtained the posterior density of regression vector β in conjugate family, we can simply perform Gibbs sampler for its evaluation. Newton-Raphson algorithm may be used to find β^* and H. To achieve dimension reduction, we use prior density with some characteristic to shrink each coefficient.

4.2.2 Model selection via nonlocal prior

We consider the nonlocal prior ([Johnson and Rossell(2012)]) as the prior specification. In particular, product moment density (pMOM) proposed by [Johnson and Rossell(2012)]:

$$p(\beta|\tau,\sigma^2,r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2}|M_p|^{1/2}\exp\left[-\frac{1}{2\tau\sigma^2}\beta' M_p\beta\right]\prod_{i=1}^{p}\beta_i^{2r}. \tag{4.2}$$

Here, σ^2 is a dispersion parameter. In the case of no over-dispersion logistic regression, we set $\sigma^2 = 1$. Positive value τ is a scale parameter that determines dispersion of the prior density on β around 0. M_p is a $p \times p$ nonsingular scale matrix. In the case of no subjective information regarding the prior correlation between regression coefficients, we can set $M_p = I_p$. r takes value from positive integers and it is called the order of density. We set $r = 1$ for simplicity. d_p is the normalizing constant.

In logistic regression with the pMOM prior, we continue using the notations as introduced before. Then the expression of $F(\beta)$ is

$$\begin{aligned} F(\beta) = -\log p(y,\beta) = &\sum y_i \log(1+e^{-X_i\beta}) + \sum (1-y_i)\log(1+e^{X_i\beta}) \\ &- \log(d_p) + \frac{p}{2}\log(2\pi) + \left(\frac{p}{2}+rp\right)\log(\tau\sigma^2) - \frac{1}{2}\log|M_p| \\ &+ \frac{1}{2\tau\sigma^2}\beta' M_p\beta - 2r\sum\log(\beta_i). \end{aligned} \tag{4.3}$$

With numerical optimization method, we can get the maximizer of $\log p(y,\beta)$, equivalently, a minimizer of $F(\beta)$, which is β^*. Then the value of joint density evaluated at β^* is easy to calculate. Also, we can get the Hessian matrix of $F(\beta)$ in the closed form expression: The

(i,j)th element of the Hessian matrix for $i \neq j$ is expressed by

$$\frac{\partial F}{\partial \beta_i \partial \beta_j} = \sum_m y_m X_{mi} X_{mj} \frac{e^{\sum_s X_{ms}\beta_s}}{(1+e^{\sum_s X_{ms}\beta_s})^2} + \sum_m (1-y_m) X_{mi} X_{mj} \frac{e^{-\sum_s X_{ms}\beta_s}}{(1+e^{-\sum_s X_{ms}\beta_s})^2},$$

while the ith diagonal element equals to

$$\frac{\partial F}{\partial^2 \beta_i} = \sum_m y_m X_{mi}^2 \frac{e^{\sum_s X_{ms}\beta_s}}{(1+e^{\sum_s X_{ms}\beta_s})^2} + \sum_m (1-y_m) X_{mi}^2 \frac{e^{-\sum_s X_{ms}\beta_s}}{(1+e^{-\sum_s X_{ms}\beta_s})^2} + \frac{1}{\tau} + \frac{2}{\beta_i^2}.$$

With these expressions, we can easily obtain the determinant of the Hessian matrix evaluated at β^*. Then the marginal density of y can be approximated by Laplace approximation. This gives us the marginal density of the data, $p(y)$, in a closed form:

$$\begin{aligned} Z = p(y) &= \int p(y,\beta) d\beta = \int e^{-F(\beta)} d\beta \\ &\approx det \left(\frac{F''(\beta^*)}{2\pi} \right)^{-\frac{1}{2}} e^{-F(\beta^*)} \\ &= e^{-F(\beta^*)} (2\pi)^{k/2} |H|^{-1/2}. \end{aligned} \tag{4.4}$$

Now for any model k, we can assign a prior probability $p(k)$, where k can be any subset of the full model. A reasonable and simple prior density can be a uniform prior, that is, $p(k)$ is same for all k. Another model prior can be a binomial prior. In this case, each covariate has probability π to be included in the model. For example we can set $\pi = 0.5$. Another candidate prior is a beta-binomial prior. In this prior, each covariate is included in the model with probability π and π follows a beta density with parameters a and b, that is, $p(\pi|a,b) \sim \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)}$, where $B(a,b)$ is the beta function.

If we choose a certain model prior $p(k)$, combined with marginal density of the data, $p(y)$,

posterior probability for any model k can be expressed by $p(k|y) = \frac{p(k)p_k(y)}{\sum_t p(t)p_t(y)}$. When the number of covariates is large, it is impossible to list all candidate models which makes challenge in calculating the denominator of $p(k|y)$. But we can compare the posterior of any two models k_1 and k_2, since $p(k_1|y)$ and $p(k_2|y)$ share the same denominator so that we only need to compare the numerators, $p(k_1)p_{k_1}(y)$ and $p(k_2)p_{k_2}(y)$, and calculation of these two terms is not difficult. Expression (4.4) is used to approximate $p(y)$ and $p(k)$ is from the prior specification. Since comparison of posterior probabilities of models can be done, Metropolis-Hasting algorithm can be employed to obtain a sequence of sampled models. With sampled models, we can identify the maximum a posteriori (MAP) model and estimate the posterior probabilities of the MAP and other high-probability models.

4.2.3 Extension to GLM

We have explained logistic regression model selection in the previous section. In this part, we extend the whole theory to a generalized linear model(GLM). Suppose that the data can be modeled by GLM with the density function given by

$$p(y|\theta) = \exp\{a(\theta)y + b(\theta) + c(y)\}, \tag{4.5}$$

where $a(\theta)$ and $b(\theta)$ are continuously differentiable functions of θ, $c(y)$ is a constant function of y, $a(\theta)$ has nonzero derivative, and θ is called the natural parameter that relates y to the predictors through a linear function

$$\theta = \beta_1 x_1 + \cdots + \beta_p x_p, \tag{4.6}$$

where $\beta_1, \ldots, \beta_p$ are regression coefficients. The mean function is $\mu = E(y|x_1, \ldots, x_p) = -\frac{b'(\theta)}{a'(\theta)} \equiv \psi(\theta)$.

In the following, we will show the sketch of GLM model selection procedure. Joint density is $p(y, \beta) = p(y|\beta)p(\beta) = \exp\{\sum_{i=1}^{n} \big(a(\theta_i)y_i + b(\theta_i) + c(y_i)\big)\} \cdot p(\beta)$, where (4.2) is used for $p(\beta)$. With the specific GLM definition, we can get corresponding function of $a(\theta)$, $b(\theta)$ and $c(\theta)$. Then we can write down the log likelihood function $F(\beta)$ in a similar way as in (4.3):

$$
\begin{aligned}
F(\beta) = & -\log p(y, \beta) \qquad (4.7)\\
= & \sum (a(\theta_i(\beta))y_i + b(\theta_i(\beta)) + c(y_i)) \\
& - \log(d_p) + \frac{p}{2}\log(2\pi) + \left(\frac{p}{2} + rp\right)\log(\tau\sigma^2) - \frac{1}{2}\log|M_p| \\
& + \frac{1}{2\tau\sigma^2}\beta' M_p \beta - 2r\sum \log(\beta_i).
\end{aligned}
$$

In this expression, $\theta_i(\beta)$ is used to emphasize a linear relationship between θ and β. With specific forms of $a(\theta)$, $b(\theta)$ and $c(y)$, calculation of a minimum point β^* for $F(\beta)$ will be feasible and computing the Hessian matrix is doable, even though the procedure can be tedious and complicated depending on the form of $a(\theta)$ and $b(\theta)$. The (i, j)th element of the Hessian matrix is expressed by $\sum_m (a''(\theta_m)y_m + b''(\theta_m))X_{mi}X_{mj}$ while the ith diagonal element equals to $\sum_m (a''(\theta_m)y_m + b''(\theta_m))X_{mi}^2 + \frac{1}{\tau} + \frac{2}{\beta_i^2}$.

After obtaining minimum point β^* of $F(\beta)$ and Hessian matrix $\mathrm{H}(\beta^*)$, marginal density $p(y) \approx e^{-F(\beta^*)}(2\pi)^{k/2}|H(\beta^*)|^{-1/2}$ can be easily computed. With a certain model prior $p(k)$ and marginal density $p_k(y)$, for any two models k_1 and k_2, we can compare the posterior probabilities $p(k_1|y)$ and $p(k_2|y)$. So a sequence of candidate models can be generated by the Metropolis-Hastings algorithm. The MAP model and posterior probability of highly-likely candidate models can be derived from the sequence of sampled models.

4.2.4 Theoretical properties

We would like to have model t obtained from the maximum posterior probability converge to the true model t^* as sample size n is increasing. Define f^* as the true density under the model t^* and f as the proposed density under the model t. To investigate the convergence rate, we follow the results of [Jiang (2007)]. We assume the nonlocal prior specification is used and $\lim_{n\to\infty}\sum_1^{p_n}|\beta_j^*| < \infty$, where p_n is the number of covariates that allows to increase with sample size n. γ is a subset of covariate indices for which $|\beta_j| > 0$ and $|\gamma|$ is the cardinality of γ. Also, let $ch_1(M)$ be the largest eigenvalue of a matrix M. For two positive sequences a_n and b_n, $a_n \prec b_n$ means $\lim_{n\to\infty} a_n/b_n = 0$.

For the nonlocal prior with density (2), we consider a diagonal matrix M_p. That is, no prior correlation between regression coefficients is assumed, since the prior normalization constant d_p can be difficult to evaluate when M_p is not a diagonal matrix. However, following theory holds for any matrix M_p that satisfies certain assumptions. When M_p is a diagonal matrix, it is proportional to a covariance matrix A_γ. In the following, we put some assumptions on M_γ.

We first introduce some notations. r_n is prior expectation of the model size. $\triangle(r_n) = \inf_{\gamma:|\gamma|=r_n}\sum_{j:j\notin\gamma}|\beta_j^*|$, $B(r_n) = \sup_{\gamma:|\gamma|=r_n} ch_1(M_\gamma)$, $\bar{B}(r_n) = \sup_{\gamma:|\gamma|=r_n} ch_1(M_\gamma^{-1})$, $\tilde{B}_n = \sup_\gamma ch_1(M_\gamma^{-1})$, $D(R) = 1 + R\cdot\sup_{|h|\le R}|a'(h)|\cdot\sup_{|h|\le R}|\psi(h)|$. Without loss of generality, we assume $B(r_n), \bar{B}(r_n)$ and $\tilde{B}_n$ are bounded. Let $\epsilon_n \in (0,1]$ for each n, $n\epsilon_n^2 \succ 1$ and assume the following conditions hold:

Assumption 1

(C 1) $p_n \cdot \log(1/\epsilon_n^2) \prec n\epsilon_n^2$,

(C 2) $p_n \cdot \log(p_n) \prec n\epsilon_n^2$,

(C 3) $p_n \cdot \log\left(D\left(2p_n\sqrt{n\epsilon_n^2\tilde{B}_n}\right)\right) \prec n\epsilon_n^2$,

(C 4) $r_n \prec p_n$,

(C 5) $r_n \log \bar{B}(r_n) \prec n\epsilon_n^2$ *and* $\triangle(r_n) \prec \epsilon_n^2$,

(C 6) $\log\left(\frac{r_n}{p_n}\right) \leq -\frac{4n\epsilon_n^2}{p_n}$.

Then we prove the following theorem to show models selection consistency and also derive the convergence rate.

Theorem 3 *Suppose that the prior given in (4.2) is employed and the conditions in Assumption 1 hold. Let* $P\{\cdot\}$ *denote the probability measure for the data* D^n*. Then, we have*
(a) for some $c_0 > 0$, $\lim_{n\to\infty} P\{\pi[d(f, f^*) \leq \epsilon_n | D^n] \geq 1 - e^{-c_0 n\epsilon_n^2}\} = 1$, *where* $d(f, f^*) = \sqrt{\int\int(\sqrt{f} - \sqrt{f^*})^2 \nu_y(dy)\nu_x(dx)}$ *is the Hellinger distance between* f *and* f^*.
(b) for some $c_1 > 0$, *and for all sufficiently large* n, $P\{\pi[d(f, f^*) > \epsilon_n | D^n] \geq e^{-0.5c_1 n\epsilon_n^2}\} \leq e^{-0.5c_1 n\epsilon_n^2}$.

The proof is given in section 4.6.

4.3 Algorithm

When the number of covariates p is increasing with the sample size, the number of candidate models is exponentially increasing with 2^p. So it is nearly impossible to calculate the posterior probability for each individual candidate model. Following [Johnson and Rossell(2012)], a Methropolis-Hastings algorithm is employed to generate MCMC samples from the model space, which is described below.

Step 1: Choose an initial model k^{curr}.

Step 2: For $i = 1, 2, \ldots, p$,

(a) Define a model k^{cand} by excluding or including β_i from the model k^{curr}, according to whether β_i is currently included or excluded from k^{curr}. Specifically, if model k^{curr} includes β_i, then deleting β_i from k^{curr} gives model k^{cand}. Conversely, if model k^{curr} does not include β_i, adding β_i into k^{curr} gives model k^{cand}.

(b) Compute

$$\alpha = \frac{p_{k^{cand}}(y)p(k^{cand})}{p_{k^{cand}}(y)p(k^{cand}) + p_{k^{curr}}(y)p(k^{curr})}, \tag{4.8}$$

where $p_{k^{curr}}(y)$ and $p_{k^{cand}}(y)$ are calculated from the equation (4.4). Note that $p(k^{curr})$ and $p(k^{cand})$ depend on our prior assumption.

(c) Draw $u \sim U(0, 1)$. If $\alpha > u$, update $k^{curr} = k^{cand}$.

Step 3: Repeat step 2 until a sufficiently long chain is acquired. After the sequence of sampled models is obtained, we can use this sampler chain to identify the maximum a posteriori(MAP) model, the posterior probability of the MAP and other high-probability models. Also, the marginal information including probability for each covariate can be computed based on sampled models.

4.4 Simulation and real data application

In this section, we evaluate the performance of the proposed nonlocal prior in high-dimensional logistic regression models. We call this approach nonlocal prior method. We compare the nonlocal prior method with LASSO ([Friedman, Hastie and Tibshirani (2010)]) and Empirical Bayesian LASSO in generalized linear models, denoted as gLASSO and EBLASSO, respectively, where empirical Bayesian LASSO proposed an efficient algorithm

to solve Bayesian LASSO models ([Huang, Xu and Cai(2013)]).

In the simulation study, we adopt the null model ($k = 0$) as an initial model k^{curr} to avoid bias. For the prior density of a model $p(k)$, we use a beta-binomial model on the model space. Suppose ρ is a value between 0 and 1, which represent inclusion probability for each covariate. This prior is obtained by assuming that prior probability assigned to a model k is specified as

$$p(k|\rho) = \rho^{|k|}(1-\rho)^{(p-|k|)}, \rho \sim Beta(a, b).$$

We further assume that $a = 1$ and $b = 1/\sqrt{n}$ so that the prior expectation of the model size satisfies conditions in Assumption 1. For the prior on regression coefficients, we use first order pMOM density. A hyper-parameter τ need to be settled and we follow the recommendation of [Johnson and Rossell(2012)], that is, $\tau = 0.384$.

4.4.1 Simulation study

We first investigate how our nonlocal prior method perform in some basic settings. In this part, we generate a design matrix from a multivariate normal distribution, with moderate correlation with correlation coefficient 0.3.

In the first simulation setting, we generate 100 observations with various dimensions for covariates ($p = 60, 120, 200$, respectively). So for the design matrix X, 100 samples are drawn from a multivariate normal distribution with mean 0 and the covariance matrix whose ijth element is $\delta^{|i-j|}$ with $\delta = 0.3$. First, we consider the true regression parameter β contains only 2 nonzero values, and 0 for the rest of them. We set different values for this 2 nonzero regression coefficients (e.g. $(2, -2)$,$(1.5, -1.5)$,$(1, -1)$). The response variable is generated from the binomial distribution with success probability $\frac{e^{X\beta}}{1+e^{X\beta}}$. Summary of variable selection results are given in Table 4.1.

Table 4.1: Summary of variable selection for different parameter values. In each panel, results of gLASSO and EBLASSO are based on the tuning parameter chosen as the average of 20 cross-validation values. All results are averaged among 100 replications.

	Nonlocal Prior		**gLASSO**		**EBLASSO**	
dim	True select	False select	True select	False select	True select	False select
60	2	0.88	2	10.8	2	6.04
120	2	0.2	2	12.91	2	8.74
200	1.99	0.2	1.99	15.22	1.99	10.5

(a) Variable selection result for the parameter value (2,-2)

	Nonlocal Prior		**gLASSO**		**EBLASSO**	
dim	True select	False select	True select	False select	True select	False select
60	1.97	1.86	1.98	9.32	1.99	6.25
120	1.98	0.35	1.97	11.48	1.98	8.18
200	1.96	0.26	1.9	12.28	1.99	9.08

(b) Variable selection result for the parameter value (1.5,-1.5)

	Nonlocal Prior		**gLASSO**		**EBLASSO**	
dim	True select	False select	True select	False select	True select	False select
60	2	0.88	1.83	6.94	1.85	5.27
120	1.71	0.66	1.59	8.58	1.73	7.49
200	1.72	0.54	1.49	7.52	1.76	8.26

(c) Variable selection result for the parameter value (1,-1)

We use true selection and false selection as criterion. Our definition for true selection is the number of selected true nonzero coefficients. False selection is defined as the number of falsely selected coefficients, that is coefficients selected in model with actually zero value. A method with good performance on variable selection should have true selection close to 2 and false selection close to 0.

When the true parameter value is $(2, -2)$ or $(1.5, -1.5)$, which is moderately strong signal, nonlocal prior method recognizes the true model even for relatively high dimensional cases (e.g. $p = 200$ and $n = 100$).

Table 4.2: Summary of variable selection for parameter value (2,-2,2,-2). Results of gLASSO and EBLASSO are based on tuning parameter chosen as the average of 20 cross-validation value. All results are averaged among 100 replications.

	Nonlocal Prior		**gLASSO**		**EBLASSO**	
dim	True select	False select	True select	False select	True select	False select
60	4	1.13	4	14.92	4	6.4
120	3.89	0.18	3.97	19.32	3.97	10.84
200	3.88	0.24	3.9	20.62	3.81	11.77

However, gLASSO selects largest models. EBLASSO selects larger model than nonlocal prior method, but smaller model compared with gLASSO. When coefficient value is large, it is easier to identify the significant variable for most methods. When coefficient value is relatively small (Part (c) in the Table 4.1), i.e. coefficient signal is relatively weak, all methods select some noise covariates. However, the model size under gLASSO is much higher than proposed method, while model size under EBLASSO is intermediate among these three. Now we extend the model with regression coefficients β contains 4 nonzero components. We set these 4 nonzero values as $(2, -2, 2, -2)$, and 0 for the rest. Summary results are shown in Table 4.2. Targeted value for true selection is 4 and false selection is 0. From the result in Table 4.2, nonlocal prior method works well with 4 nonzero coefficients while gLASSO includes some extra covariates again. For EBLASSO, falsely selected covariates number is higher than the proposed nonlocal method, but lower than gLASSO.

We also tested variable selection performance on models with 6 nonzero coefficients. 6 nonzero values are set as $(2, -2, 2, -2, 2, -2)$ and 0 for the rest. Model selection results are displayed in Table 4.3. The nonlocal prior method tends to include some noise coefficients or omit some important covariates. This may be caused by some potential correlation between these 6 covariates which may weaken the effect of important variables and make some noise variables. On the other hand, gLASSO contains more noise covariates.

Table 4.3: Summary of variable selection for parameter value (2,-2,2,-2,2,-2). Results of gLASSO and EBLASSO are based on tuning parameter chosen as the average of 20 cross-validation value. All results are based on 100 replications.

	Nonlocal Prior		**gLASSO**		**EBLASSO**	
dim	True select	False select	True select	False select	True select	False select
60	5.82	1.05	5.99	16.94	5.95	7.58
120	5.19	0.48	5.73	20.81	5.71	9.83
200	4.26	0.62	5.31	22.13	5.27	11.26

4.4.2 SNPs study

This simulation study mimics case-control genetic association study. Response variable y represents disease status of a subject. It takes value 1 for the case and 0 for the control. Explanatory variables are generated as SNPs in human genome. So x_{ij} is genotype of SNP j of the subject i which takes value 0, 1 or 2. x_i represents for all genotype expression for subject i. Following [Chen and Chen (2012)], the data were generated by following procedure.

Let n_1 and n_2 denote the numbers of cases and controls, respectively. Let $s = \{1, 2, \ldots, k\}$ denote the causal SNPs for the disease. Here $x_i(s)$ stands for the causal genotype set $\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$. Thus, there are 3^k possible genotype profiles for the k SNPs. For the SNPs belonging to s, the disease risk model is given by

$$\text{logit}P(y_i = 1|x_i(s)) = \sum_{j=1}^{k} \beta_j x_{ij}$$

for the prespecified values of $\beta_1, \ldots, \beta_k$. For the noncausal SNPs $x_{k+1}, \ldots, x_p$, each x_{ij} is generated from a binomial distribution with parameters $(2, p_j)$, where p_j represents the frequency of one allele and is generated from $Beta(2, 2)$. This example consists of 10 simulated datasets. Each was generated with $n_1 = n_2 = 500$, $p = 10,000$, $k = 8$, and $(\beta_1, \ldots, \beta_8) = (0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2.1.3)$.

Table 4.4: Comparison of Nonlocal prior with gLASSO and BSR. Results are averaged among 10 replications.

Methods	**Nonlocal prior**	**gLASSO**	**BSR**
size	5.75	44.9	8.2
fsr(%)	0	81.8	3.66
nsr(%)	28.13	0	12.22

This setting is the same as Bayesian subset regression (BSR) method in [Liang, Song and Yu (2013)]. The numerical results are summarized in Table 4.4.

To measure the performance of each method, we calculate the false selection rate (fsr) and negative selection rate (nsr) among 10 replications. Let s_i^* denote the set of selected features in the dataset i. Then,

$$fsr = \frac{\sum_{i=1}^{10} |s_i^* \backslash s|}{\sum_{i=1}^{10} |s_i^*|},\ nsr = \frac{\sum_{i=1}^{10} |s \backslash s_i^*|}{\sum_{i=1}^{10} |s|}.$$

Variable selection method with low fsr and nsr implies good performance. Nonlocal prior has the smallest fsr and highest nsr. Also nonlocal prior chooses the smallest model size among three methods, which implies that nonlocal prior tends to choose a simpler model. By choosing the model as simple as possible, nonlocal prior may omit some causal variables, which results in high nsr in Table 4.4. This indicates the proposed method is definitely effective in regression setting, but should be carefully adopted if there are causal variables.

4.4.3 Real data application

In this part, we test nonlocal prior Bayesian variable selection method on colon gene expression data ([Alon et al. (1999)]) and compare our result with gLASSO. Colon data set studies on 62 samples, which are composed of 40 colon tumor samples and 22 normal colon

tissue samples, analyzed with 2000 human genes. The response variable has two level: 0 for normal colon tissue and 1 for colon tumor. To see which genes are closely correlated with colon tumor, we use a logistic regression model and select variables with our nonlocal prior. To evaluate our method, we divide data into two groups, use 52 samples as a training data set, and 10 samples as a test data set. To reduce potential bias in sample observations, 20 repetitions are considered, in which 10 samples are randomly selected as a test set for each time. In each replication, with 52 samples, we first apply nonlocal prior Bayesian variable selection method and use the maximum likelihood estimator for the estimation of regression coefficients for the selected covariates. Then we apply this fitted logistic regression model to test classification using a data set with 10 samples for checking performance of this method in terms of prediction. Since gLASSO is a renowned method for variable selection, we compare nonlocal prior prediction result with gLASSO. To make this comparison consistent, we keep the same training data set for gLASSO and nonlocal prior for each replication. The summarized prediction results are listed in Table 4.5. All results in Table 4.5 are averaged among 20 replications.

We consider True positive and False positive for the performance measures. True positive (TP) represents those samples observed as colon tumor is predicted as colon tumor (response value 1). False positive (FP) is samples predicted as colon tumor while observed as colon normal tissue (response value 0). True negative (TN) is for those predicted as colon normal tissue but observed as colon normal tissue. False negative (FN) means samples predicted as colon normal tissue while observed as colon tumor. True positive rate (TPR), also called as sensitivity, measures the test's ability to correctly identify patients who do have the condition.

Table 4.5: Compare on prediction results based on Nonlocal prior and gLASSO. Result from gLASSO is based on tuning parameter chosen as the average of 20 cross-validation value. All results are averaged among 20 replications.

	Sensitivity	**FNR**	**Specificity**	**FPR**
Nonlocal prior	86.96%	13.04%	65.85%	34.15%
gLASSO	86.96%	13.04%	70.73%	29.27%

(a) Prediction accuracy result summary

	Average model size	**TR**	**FR**
Nonlocal prior	1.25	79.09%	20.91%
gLASSO	12.82	80.91%	19.09%

(b) Overall prediction performance comparison. TR represents for total correct prediction rate, FR represents for total incorrect prediction rate. Detailed definition explained in chapter.

	Deviance
Nonlocal prior	-23.97
gLASSO	-21.15

(c) Average deviance by Nonlocal prior and gLASSO.

It is defined as

$$TPR(sensitivity) = \frac{TP}{TP + FN},$$

and *false negative rate (FNR)=1-sensitivity*. True negative rate (TNR), also called specificity is related to the test's ability to correctly delete patients without the condition. Definition is expressed as

$$TNR(specificity) = \frac{TN}{TN + FP},$$

and *false positive rate (FPR)=1-specificity*. To measure prediction power of the model, higher sensitivity and specificity means better prediction power. In Table 4.5 (b), true rate (TR) is the percentage for all correct prediction, expressed by $TR = \frac{TP+TN}{TP+FP+TN+FN}$ and *false rate(FR)=1-TR*. Again, we would expect high TR for the model with good prediction performance.

From the prediction result in Table 4.5, we can see that nonlocal prior and gLASSO has the same sensitivity, so they have the same performance in predicting the positive result. And gLASSO shows more accurate result when predicting negative result since gLASSO has higher specificity. When comparing overall prediction performance in Table 4.5 (b), gLASSO has higher correct prediction rate than nonlocal prior (in terms of TR in the table), with 2% difference. In most cases, nonlocal prior recommends a simpler model. The averaged model size for nonlocal prior is 1.25 and gLASSO contains 12.82 covariates on average. Nonlocal prior is only slightly underperformed than gLASSO but with a much simpler model.

We also compare the goodness of fit in model fitting, measured by deviance. Here, deviance is defined as $-2(logLike - logLike\_sat)$, where $logLike$ is the log-likelihood for the fitted model and $logLike\_sat$ is the log-likelihood for the saturated model. Result is listed in Table 4.5 (c). Similarly as before, numbers are averaged among 20 replications. Nonlocal prior shows smaller deviance, which means better fitting.

We now apply our model with full data set, the recommended model would contain two genes, gene 493 and gene 1884. gLASSO with tuning parameter λ given by cross-validation agrees with model size 8. Compared with gLASSO, nonlocal prior Bayesian method recommends a simpler model.

4.5 Conclusion

The unique characteristic of nonlocal prior provides us with a new model selection method in generalized linear model, which could efficiently eliminate unnecessary covaraites and lead to a parsimonious model. Convergence rate is derived under some attainable assumptions. Laplace approximation is applied to overcome calculation difficulty.

Based on application and simulation result, our proposed nonlocal prior method leads to a simpler interpretation of the model and coefficients without added issue of over selection. Based on application result, our method is comparable with gLASSO in terms of prediction rate, but results in a simpler model. Simulation result shows our proposed method could identify the true model with less non significant covariates included in the model, compared with gLASSO and empirical Bayesian LASSO. Also, our method shows better goodness of fit in model fitting.

The proposed method is well defined in theory along with a clear algorithm. However, one limitation for the method is intensive computing time. In application, the proposed method is much slower than gLASSO especially when candidate dimension is large. In future study, an efficient algorithm needs to be developed to improve this intensive computing issue.

4.6 Proof of theorem 3

[Jiang (2007)] provide general conditions for the prior to give a convergence rate of the probability regarding the Hellinger distance between the posterior model and the true model. We check the conditions are satisfied in our setting. In particular, it is enough to show conditions (O) and (N) of [Jiang (2007)] are satisfied. Condition (O) limits the tail densities of prior and Condition (N) defines the prior density on an approximation neighborhood.

In the following description of Condition (O), K_n is the same as the candidate dimension p_n. To be consistent with the notation with [Jiang (2007)], we use K_n to denote p_n.

4.6.1 Condition (O)

Let $D(R) = 1 + R \cdot \sup_{|h|\leq R}|a'(h)| \cdot \sup_{|h|\leq R}|\psi(h)|$ for any $R > 0$. Define $\bar{r}_n$ as maximal model size, which satisfy $1 \leq \bar{r}_n < K_n$. There exist some $C_n > 0$ such that

$$\bar{r}_n \log(1/\epsilon_n^2) \prec n\epsilon_n^2 \tag{4.9}$$

$$\bar{r}_n \log K_n \prec n\epsilon_n^2 \tag{4.10}$$

$$\bar{r}_n \log D(\bar{r}_n C_n) \prec n\epsilon_n^2 \tag{4.11}$$

Furthermore, for all large enough n, the following two equations hold:

$$\pi(|\gamma| > \bar{r}_n) \leq e^{-4n\epsilon_n^2} \tag{4.12}$$

and for all γ such that $|\gamma| \leq \bar{r}_n$, for all $j \in \gamma$,

$$\pi(|\beta_j| > C_n|\gamma) \leq e^{-4n\epsilon_n^2} \tag{4.13}$$

4.6.2 Condition (N)

Assume that a sequence of models γ_n exists such that, as n increases,

$$\sum_{j\notin\gamma_n} |\beta_j^*| \prec \epsilon_n^2 \tag{4.14}$$

and for any sufficiently small $\eta > 0$, there exists N_η such that, for all $n > N_\eta$, we have

$$\pi(\gamma = \gamma_n) \geq e^{-n\epsilon_n^2/8} \tag{4.15}$$

and

$$\pi\left(\beta_\gamma \in M\left(\gamma_n, \eta\right) | \gamma = \gamma_n\right) \geq e^{-n\epsilon_n^2/8} \tag{4.16}$$

where $M\left(\gamma_n, \eta\right) = \left(\beta_j^* \pm \eta\epsilon_n^2/|\gamma_n|\right)_{j\in\gamma_n}$.

4.6.3 Verification of Conditions (O) and (N) for nonlocal prior

We first check condition (O).

(4.9) and (4.10) are satisfied automatically by (C 1) and (C 2) in Assumption 1. Recall $K_n = p_n$. By setting $\bar{r}_n = p_n - 1$, $1 \leq \bar{r}_n < K_n$. r_n is the prior expectation of model size.

$$\pi(|\gamma| > \bar{r}_n) = \pi(|\gamma| = p_n) = \left(\frac{r_n}{K_n}\right)^{K_n}, \tag{4.17}$$

$$\log\pi(|\gamma| = p_n) = K_n \log\left(\frac{r_n}{K_n}\right) \leq -4n\epsilon_n^2 \tag{4.18}$$

by (C 6) in Assumption 1 so that (4.12) holds.

Next we need to show $\pi(|\beta_j| > C_n|\gamma) \leq e^{-4n\epsilon_n^2}$ holds.

$$\pi(|\beta_j| > c|\gamma) \leq M\int_c^\infty \beta^2 \exp\left\{-\frac{\beta^2}{2\tilde{B}_n}\right\} \mathrm{d}\beta \tag{4.19}$$

$$= \left(M\tilde{B}_n\right)\left((-\beta)\exp\left\{-\frac{\beta^2}{2\tilde{B}_n}\right\}\Big|_c^\infty + \int_c^\infty \exp\left\{-\frac{\beta^2}{2\tilde{B}_n}\right\}\mathrm{d}\beta\right)$$

$$< M'\left(c\,\exp\left\{-\frac{c^2}{2\tilde{B}_n}\right\} + \frac{1}{c}\exp\left\{-\frac{c^2}{2\tilde{B}_n}\right\}\right) \tag{4.20}$$

$$= M'\left(c + \frac{1}{c}\right)\exp\left\{-\frac{c^2}{2\tilde{B}_n}\right\}.$$

In (4.19), M is the normalization constant when taking marginal distribution from the joint distribution. M' in (4.20) is $M\tilde{B}_n$. If $c = C_n = 2\sqrt{\tilde{B}_n n\epsilon_n^2}$ and $n\epsilon_n^2 \succ 1$ with $\tilde{B}_n$ bounded,

$$M'\left(c + \frac{1}{c}\right)\exp\left\{-\frac{c^2}{2\tilde{B}_n}\right\} \leq \exp\left\{-n\epsilon_n^2\right\}.$$

Take $\epsilon_n' = \epsilon_n/2$, we can get

$$\pi(|\beta_j| > C_n|\gamma) \leq e^{-4n\epsilon_n'^2} \tag{4.21}$$

so that (4.13) is satisfied. Also condition (4.11) is satisfied with $C_n = 2\sqrt{\tilde{B}_n n\epsilon_n^2}$ by assumption (C 3). Thus, Condition (O) is checked.

Now, we verify Condition (N) for nonlocal prior.

Take the sequence of models γ_n such that, for each n, $\gamma = \gamma_n$ reaches its infimum in $\triangle(r_n) = inf_{\gamma:|\gamma|=r_n}\sum_{j:j\notin\gamma}|\beta_j^*|$. Then $\sum_{j\notin\gamma_n}|\beta_j^*| = \triangle(r_n) \prec \epsilon_n^2$. For the condition on prior $\pi[\beta \in (\beta_j^* \pm \eta\epsilon_n^2/r_n)_{j\in\gamma_n}|\gamma_n]$:

(a) if for any $j \in \gamma_n$, 0 is not covered by interval $(\beta_j^* \pm \eta\epsilon_n^2/r_n)$,

$$\begin{aligned}\pi[\beta \in (\beta_j^* \pm \eta\epsilon_n^2/r_n)_{j\in\gamma_n}|\gamma_n] &\geq |2\pi M_{\gamma_n}^{-1}|^{-\frac{1}{2}}e^{-0.5\cdot\frac{1}{\tau}\bar{\beta}^T M_{\gamma_n}\bar{\beta}}(\eta\epsilon_n^2/r_n)^{r_n}\prod_i \bar{\beta}_i^2 \\ &= T_1 \cdot \prod_i \bar{\beta}_i^2.\end{aligned}$$

Here $\bar{\beta}$ is some intermediate value making the density achieving its minimum over $(\beta_j^* \pm \eta\epsilon_n^2/r_n)_{j\in\gamma_n}$. Also, we denote $|2\pi M_{\gamma_n}^{-1}|^{-\frac{1}{2}}e^{-0.5\cdot\frac{1}{\tau}\bar{\beta}^T A_{\gamma_n}\bar{\beta}}(\eta\epsilon_n^2/r_n)^{r_n}$ as T_1. Define c_1, c_2, c_3 as positive constants, and $c_2 > c_1$. We need to show $T_1\prod_i \bar{\beta}_i^2 \succ e^{-c_2 n\epsilon_n^2}$. Since

[Jiang (2007)] showed $T_1 \succ e^{-c_1 n\epsilon_n^2}$ with (C 5) holds, it is sufficient to show

$$e^{-c_1 n\epsilon_n^2} \prod_i \bar{\beta}_i^2 \succ e^{-c_2 n\epsilon_n^2}$$

which implies $\prod_i \bar{\beta}_i^2 \succ e^{-(c_2-c_1)n\epsilon_n^2}$. Consequently, we need to show $\prod_i \bar{\beta}_i^2 \succ e^{-c_3 n\epsilon_n^2}$. Without loss of generality, suppose β_i is positive then the minimum of β_i is $\beta_i^* - \eta\frac{\epsilon_n^2}{r_n}$, so $\prod_i \bar{\beta}_i^2 > \prod_i(\beta_i^* - \eta\frac{\epsilon_n^2}{r_n})^2$. Now our question is to show $\prod_i(\beta_i^* - \eta\frac{\epsilon_n^2}{r_n})^2 \succ e^{-cn\epsilon_n^2}$. For this, it is enough to show the order for $\left(\eta\frac{\epsilon_n^2}{r_n}\right)^2$ since β_i^* is constant. That is, we want to show

$$\prod \left(\eta\frac{\epsilon_n^2}{r_n}\right)^2 = \left(\eta\frac{\epsilon_n^2}{r_n}\right)^{2r_n} \succ e^{-cn\epsilon_n^2} \tag{4.22}$$

Since $r_n \log\frac{1}{\epsilon_n^2} \leq p_n \log\frac{1}{\epsilon_n^2} \prec n\epsilon_n^2$ and $r_n \log r_n \leq p_n \log p_n \prec n\epsilon_n^2$ holds by (C 1) and (C 2), we can derive $-2r_n \log\left(\eta\frac{\epsilon_n^2}{r_n}\right) \prec cn\epsilon_n^2$, which implies $e^{2r_n \log\left(\eta\frac{\epsilon_n^2}{r_n}\right)} \succ e^{-cn\epsilon_n^2}$. This is equivalent to (4.22).

(b) if there is at least one j, such that 0 is covered by interval $(\beta_j^* \pm \eta\epsilon_n^2/r_n)$ where $j \in \gamma_n$. We can separate the index set into two groups, the first group corresponding with intervals not cover 0, denote as I_1, all other index belongs to group I_2. Consider ϵ close to 0, denote $\boldsymbol{\epsilon} = (\epsilon, \epsilon, \ldots,)_{|\gamma_n|}$ such that

$$
\begin{aligned}
&\pi[\beta_j \in (\beta_j^* \pm \eta\epsilon_n^2/r_n)_{j\in\gamma_n}|\gamma_n] \\
\geq&\pi[\cap_{j\in I_1}\left(\beta_j \in (\beta_j^* \pm \eta\epsilon_n^2/r_n)\right)\cap \\
&\left(\cap_{j\in I_2}\left(\beta_j \in (\beta_j^* - \eta\epsilon_n^2/r_n, -\epsilon)\cup\left(\epsilon, \beta_j^* + \eta\epsilon_n^2/r_n\right)\right)\right)|\gamma_n] \qquad (4.23) \\
\geq&|2\pi M_{\gamma_n}^{-1}|^{-1/2}e^{-0.5\boldsymbol{\epsilon}^T M_{\gamma_n}\boldsymbol{\epsilon}}\left(\eta\frac{\epsilon_n^2}{r_n} - 2\epsilon\right)^{r_n}\epsilon^{2r_n} \qquad (4.24)
\end{aligned}
$$

The inequality in (4.24) holds, since density function of β goes to 0 as β is close to 0, so $f(\epsilon)$ can be very small when ϵ is very small. With the constraint $lim_{n\to\infty}\sum_{j=1}^{p_n}|\beta_j^*| < \infty$, we know that β_j is bounded for any j. In this case, we can always find very small ϵ such that $f(\epsilon)$ is smaller than any $f(\beta)$ in the stated interval.

$\boldsymbol{\epsilon}^T M_{\gamma_n}\boldsymbol{\epsilon} \leq r_n\epsilon^2 B(r_n)$, so $e^{-0.5\boldsymbol{\epsilon}^T M_{\gamma_n}\boldsymbol{\epsilon}} \geq e^{-0.5r_n\epsilon^2 B(r_n)}$, with bounded $B(r_n)$ and $\epsilon \to 0$, this term can be bounded away from below. For ϵ^{2r_n}, need to show ϵ^{2r_n} is not smaller than $e^{-cn\epsilon_n^2}$, that is

$$
\begin{aligned}
\epsilon^{2r_n} &\succ e^{-cn\epsilon_n^2}, \\
\epsilon &\succ e^{-\frac{cn\epsilon_n^2}{2r_n}}.
\end{aligned}
$$

Also, we need ϵ is smaller than $\frac{\epsilon_n^2}{r_n}$ in order, since $\frac{\epsilon_n^2}{r_n}$ is the interval width. Following is how

to show this may hold. By (C 1) and (C 2), we have

$$r_n \log \frac{r_n}{\epsilon_n^2} \prec n\epsilon_n^2$$

$$\log \frac{r_n}{\epsilon_n^2} \prec \frac{n\epsilon_n^2}{r_n} \tag{4.25}$$

$$e^{-\frac{n\epsilon_n^2}{r_n}} \prec \frac{\epsilon_n^2}{r_n} \tag{4.26}$$

(4.26) can be derived from (4.25) since we have $\frac{n\epsilon_n^2}{r_n} \to \infty$ from (C 2) in Assumption 1. The logic here is, if $\frac{A}{B} \to 0$ and $B \to \infty$, then $(B - A) \to \infty$ and $e^{A-B} \to 0$, which is $e^A \prec e^B$. Since $\frac{n\epsilon_n^2}{r_n} \to \infty$, we can find some small ϵ such that $\epsilon \geq e^{-\frac{cn\epsilon_n^2}{2r_n}}$. Also, we consider such ϵ is small enough so that $\epsilon \leq \frac{\epsilon_n^2}{r_n}$.

(4.26) also implies $\left(\frac{\epsilon_n^2}{r_n}\right)^{r_n} \succ e^{-cn\epsilon_n^2}$. Also, $\epsilon \leq \frac{\epsilon_n^2}{r_n}$ implies at least 2ϵ is an order of $\frac{\epsilon_n^2}{r_n}$ so that $\left(\eta\frac{\epsilon_n^2}{r_n} - 2\epsilon\right) = O\left(\frac{\epsilon_n^2}{r_n}\right)$, then $\left(\eta\frac{\epsilon_n^2}{r_n} - 2\epsilon\right)^{r_n} \succ e^{-cn\epsilon_n^2}$ holds as a result of $\left(\frac{\epsilon_n^2}{r_n}\right)^{r_n} \succ e^{-cn\epsilon_n^2}$. Thus, condition $\pi(\beta_\gamma \in M(\gamma_n, \eta)|\gamma = \gamma_n) \geq e^{-n\epsilon_n^2}$ in (4.16) is checked.

Next check inequality $\pi(\gamma = \gamma_n) \geq e^{-n\epsilon_n^2/8}$ in (4.15). Notice that γ_n is chosen such that $|\gamma_n| = r_n$, so $\pi(\gamma = \gamma_n) = (\frac{r_n}{K_n})^{r_n}(1 - \frac{r_n}{K_n})^{K_n - r_n}$, since $\frac{r_n}{K_n} \prec 1$ by (C 4), we have $\log\pi(\gamma = \gamma_n) \sim r_n \log\frac{r_n}{K_n} \geq -r_n \log K_n$ since $r_n \log r_n$ is positive, and $r_n \log K_n \prec n\epsilon_n^2$ by (C 4) and (C 2), so $\pi(\gamma = \gamma_n) \geq e^{-n\tilde{\epsilon}_n^2/8}$ by taking $\tilde{\epsilon}_n = \epsilon_n/\sqrt{8}$.

If we take $\epsilon_n' = \epsilon_n/\sqrt{8}$, the inequality in (4.21) still holds. So we can apply ϵ_n replaced by ϵ_n' in Theorem 1, so that the Hellinger neighborhood will take a radius $\sqrt{8}\epsilon_n'$. Condition (N) is checked.

4.6.4 Proof of theorem 1

By checking conditions at Verification of Conditions (O) and (N) for nonlocal prior, Theorem 1 holds as a result of Theorem 4 from [Jiang (2007)].

BIBLIOGRAPHY

BIBLIOGRAPHY

[Akaike (1973)] Akaike, H., "Information Theory and an Extension of the Maximum Likelihood Principle," in 2nd International Symposium on Information Theory, B. N. Petrov and F. Csaki, eds., Akademiai Kiado, Budapest, 1973, pp. 267-281.

[Alon et al. (1999)] Alon, Uri, et al. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." Proceedings of the National Academy of Sciences 96.12 (1999): 6745-6750.

[Bhattacharya et al.(2014)] Bhattacharya, Anirban, et al. "Dirichlet-Laplace priors for optimal shrinkage." Journal of the American Statistical Association just-accepted (2014): 00-00.

[Bondell and Reich(2012)] Bondell, Howard D., and Brian J. Reich. "Consistent high-dimensional Bayesian variable selection via penalized credible regions." Journal of the American Statistical Association 107.500 (2012): 1610-1624.

[Candes and Tao (2007)] Candes, Emmanuel, and Terence Tao. "The Dantzig selector: Statistical estimation when p is much larger than n." The Annals of Statistics (2007): 2313-2351.

[Carvalho, Polson, and Scott(2009)] Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. "Handling sparsity via the horseshoe." International Conference on Artificial Intelligence and Statistics. 2009.

[Carvalho, Polson, and Scott(2010)] Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. "The horseshoe estimator for sparse signals." Biometrika (2010): asq017.

[Castillo and vander Vaart(2012)] Castillo, Ismael, and Aad van der Vaart. "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." The Annals of Statistics 40.4 (2012): 2069-2101.

[Chen and Chen (2012)] Chen, Jiahua, and Zehua Chen. "Extended BIC for small-n-large-P sparse GLM." Statistica Sinica (2012): 555-574.

[Chen, Ibrahim and Yiannoutsos (1999)] Chen, M-H., Joseph G. Ibrahim, and Constantin Yiannoutsos. "Prior elicitation, variable selection and Bayesian computation for logistic regression models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61.1 (1999): 223-242.

[Dellaportas et al. (1997)] Dellaportas, Petros, Dimitris Karlis, and Evdokia Xekalaki. "Bayesian analysis of finite poisson mixtures." Manuscript (1997).

[Fan and Lv(2008)] Fan, Jianqing, and Jinchi Lv. "Sure independence screening for ultra-high dimensional feature space." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70.5 (2008): 849-911.

[Fan and Li (2001)] Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association 96.456 (2001): 1348-1360.

[Friedman, Hastie and Tibshirani (2010)] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

[George (1999)] George, Edward I. "Bayesian model selection." Encyclopedia of Statistical Sciences Update 3 (1999): 39-46.

[George and Foster(2000)] George, Edward I., and Dean P. Foster. "Calibration and empirical Bayes variable selection." Biometrika (2000): 731-747.

[George and McCulloch (1993)] George, Edward I., and Robert E. McCulloch. "Variable selection via Gibbs sampling." Journal of the American Statistical Association 88.423 (1993): 881-889.

[George and McCulloch (1997)] George, Edward I., and Robert E. McCulloch. "Approaches for Bayesian variable selection." Statistica sinica (1997): 339-373.

[Hoff (2009)] Hoff, Peter D. "Nonconjugate priors and Metropolis-Hastings algorithms." A First Course in Bayesian Statistical Methods. Springer New York, 2009. 171-193.

[Huang, Xu and Cai(2013)] Huang, Anhui, Shizhong Xu, and Xiaodong Cai. "Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping." BMC genetics 14.1 (2013): 5.

[Huang et al. (2008)] Huang, J., Horowitz, J. L., Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. The Annals of Statistics, 587-613.

[Ishwaran and Rao (2005)] Ishwaran, Hemant, and J. Sunil Rao. "Spike and slab variable selection: frequentist and Bayesian strategies." Annals of Statistics (2005): 730-773.

[Ishwaran and Rao (2011)] Ishwaran, Hemant, and J. Sunil Rao. "Consistency of spike and slab regression." Statistics and Probability Letters 81.12 (2011): 1920-1928.

[Jiang (2006)] Jiang, Wenxin. "On the consistency of Bayesian variable selection for high dimensional binary regression and classification." Neural computation 18.11 (2006): 2762-2776.

[Jiang (2007)] Jiang, Wenxin. "Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities." The Annals of Statistics 35.4 (2007): 1487-1511.

[Johnson(2013)] Johnson, Valen E. "On Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings." Bayesian analysis (Online) 8.4 (2013): 741-758

[Johnson and Rossell(2010)] Johnson, Valen E., and David Rossell. "On the use of nonlocal prior densities in Bayesian hypothesis tests." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72.2 (2010): 143-170.

[Johnson and Rossell(2012)] Johnson, Valen E., and David Rossell. "Bayesian model selection in high-dimensional settings." Journal of the American Statistical Association 107.498 (2012): 649-660.

[Kan(2008)] Kan, Raymond. "From moments of sum to moments of product." Journal of Multivariate Analysis 99.3 (2008): 542-554.

[Liang, Song and Yu (2013)] Liang, Faming, Qifan Song, and Kai Yu. "Bayesian subset modeling for high-dimensional generalized linear models." Journal of the American Statistical Association 108.502 (2013): 589-606.

[Liang, Truong and Wong(2001)] Liang, Faming, Young K. Truong, and Wing Hung Wong. "Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting." Statistica Sinica 11.4 (2001): 1005-1030.

[Madigan and Raftery(1994)] Madigan, David, and Adrian E. Raftery. "Model selection and accounting for model uncertainty in graphical models using Occam's window." Journal of the American Statistical Association 89.428 (1994): 1535-1546.

[Mitchell and Beauchamp(1988)] Mitchell, Toby J., and John J. Beauchamp. "Bayesian variable selection in linear regression." Journal of the American Statistical Association 83.404 (1988): 1023-1032.

[Narisetty and He(2014)] Narisetty, Naveen Naidu, and Xuming He. "Bayesian variable selection with shrinking and diffusing priors." The Annals of Statistics 42.2 (2014): 789-817.

[Nott and Leonte (2012)] Nott, David J., and Daniela Leonte. "Sampling schemes for Bayesian variable selection in generalized linear models." Journal of Computational and Graphical Statistics (2012).

[Park and Casella (2008)] Park, Trevor, and George Casella. "The bayesian lasso." Journal of the American Statistical Association 103.482 (2008): 681-686.

[Polson and Scott(2010)] Polson, Nicholas G., and James G. Scott. "Shrink globally, act locally: sparse Bayesian regularization and prediction." Bayesian Statistics 9 (2010): 501-538.

[Rossell et al. (2013)] Rossell, David, Donatello Telesca, and Valen E. Johnson. "High-dimensional Bayesian classifiers using non-local priors." Statistical Models for Data Analysis. Springer, Heidelberg, 2013. 305-313.

[Schwarz (1978)] Schwarz, Gideon. "Estimating the dimension of a model." The annals of statistics 6.2 (1978): 461-464.

[Scott and Berger.(2010)] Scott, James G., and James O. Berger. "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." The Annals of Statistics 38.5 (2010): 2587-2619.

[Sha et al.(2004)] Sha, Naijun, et al. "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage." Biometrics 60.3 (2004): 812-819.

[Tibshirani(1996)] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

[Van de Geer (2008)] Van de Geer, Sara A. "High-dimensional generalized linear models and the lasso." The Annals of Statistics (2008): 614-645.

[Xu and Ghosh(2015)] Xu, Xiaofan, and Malay Ghosh. "Bayesian variable selection and estimation for group lasso." Bayesian Analysis 10.4 (2015): 909-936.

[Zhou, Liu and Wong(2004)] Zhou, Xiaobo, Kuang-Yu Liu, and Stephen TC Wong. "Cancer classification and prediction using logistic regression with Bayesian gene selection." Journal of Biomedical Informatics 37.4 (2004): 249-259.

[Zou (2006)] Zou, Hui. "The adaptive lasso and its oracle properties." Journal of the American statistical association 101.476 (2006): 1418-1429.

[Zou and Hastie (2005)] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.