Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection

Naveen N. Narisetty, Juan Shen, and Xuming He *

Abstract

We consider the computational and statistical issues for high dimensional Bayesian model selection under the Gaussian spike and slab priors. To avoid large matrix computations needed in a standard Gibbs sampler, we propose a novel Gibbs sampler called "Skinny Gibbs" which is much more scalable to high dimensional problems, both in memory and in computational efficiency. In particular, its computational complexity grows only linearly in p, the number of predictors, while retaining the property of strong model selection consistency even when p is much greater than the sample size n. The present paper focuses on logistic regression due to its broad applicability as a representative member of the generalized linear models. We compare our proposed method with several leading variable selection methods through a simulation study to show that Skinny Gibbs has a strong performance as indicated by our theoretical work.

*K*eywords: Variable Selection; Gibbs Sampling; Scalable Computation; High Dimensional Data; Bayesian Computation; Logistic Regression

*Naveen N. Narisetty is Assistant Professor, Department of Statistics, University of Illinois at Urbana-Champaign (email: naveen@illinois.edu); Juan Shen is Assistant Professor, Fudan University, Shanghai, China (email: shenjuan@fudan.edu.cn); and Xuming He is Professor, Department of Statistics, University of Michigan, Ann Arbor (email: xmhe@umich.edu). The research is partially supported by the NSF Awards DMS-1307566, DMS-1607840, DMS-1811768 and the Chinese National Natural Science Projects 11129101, 11501123 and 11690012. The authors would like to thank Professor Faming Liang for providing us the code to perform model selection based on the Bayesian Subset Regression.

1 Introduction

With the increased ability to collect and store large amounts of data, we have the opportunities and challenges to analyze data with a large number of covariates or features per subject. When the number of covariates in a regression model is greater than the sample size, the parameter estimation problem becomes ill posed, and variable selection is usually a natural first-step. There have been extensive studies on variable selection in high dimensional settings, especially since the advent of Lasso (Tibshirani 1996), an L_1 regularized regression method for variable selection. Other penalization methods for sparse model selection include smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), minimum concave penalty (MCP) (Zhang 2010), and many variations of such methods. Though many of these methods are first introduced in the context of linear regression, their theoretical properties and optimization methods for logistic regression and other generalized linear models (GLM) have also been studied. van de Geer S. A. (2008) proved oracle inequalities for L_1 penalized high dimensional GLM, whereas the oracle properties of Fan and Peng (2004) also hold for GLM. Friedman et al. (2008), Breheny and Huang (2011) and Huang and Zhang (2012) studied optimization approaches for penalized GLM methods. The computational complexity of these algorithms typically grows linearly in p.

The literature on high dimensional Bayesian variable selection has focused mostly on linear models, but most techniques generalize, with some efforts, to logistic regression and other GLMs. It has been understood that most penalization methods have Bayesian interpretations, because all the methods share the basic desire of shrinkage towards sparse models. We refer to Bhattacharya et al. (2015); Bondell and Reich (2012); Johnson and Rossell (2012); Park and Casella (2008); Ročková and George (2014) for some recent work on Bayesian shrinkage. An advantage of Bayesian methods for variable selection is that Markov Chain Monte Carlo (MCMC) techniques can be used to explore the posterior

distributions, which often offer a more informative approach to model selection than the corresponding penalization method with a highly non-convex optimization problem. For instance, the methods proposed by Liang et al. (2013), Narisetty and He (2014), and Shen et al. (2012) are similar to the L_0 penalty, which is generally considered to be desirable for model selection consistency.

Continuous spike and slab priors have been commonly used in practice and with appropriate prior choices have been shown to have desirable model selection properties (George and McCulloch 1993; Ishwaran and Rao 2005; Narisetty and He 2014). For Gaussian linear models, one has the option to use the point-mass distribution as the spike priors by integrating out the p-dimensional coefficient vector to reduce computation (Guan and Stephens (2011)). However, the integration step becomes less appealing in non-linear models. The standard Gibbs sampling algorithm for posterior computation in this framework requires sampling from a p-variate normal distribution with a non-sparse covariance matrix, which is not so scalable for large p. In this paper, we propose a new computational technique within the Gibbs sampling framework that replaces the high dimensional covariance matrix by a sparse one so that large matrix operations can be avoided. The resulting algorithm is called Skinny Gibbs, because it uses a "skinny" covariance matrix in the Gibbs algorithm. One might view Skinny Gibbs as an approximation to the original Gibbs sampler, but more importantly, we develop theoretical results which show that Skinny Gibbs has its own stationary distribution, and there is no sacrifice on the strong model selection consistency property desired for Bayesian model selection. In summary, the main contributions of our paper are the following:

- We propose a novel strategy to sparsify the conditional sampling steps of a Gibbs sampler to obtain a scalable Gibbs sampler suitable for high dimensional problems.

- In spite of the aforementioned sparsification of the conditional sampling steps within

the Gibbs sampler to make it scalable, we establish that the resultant Skinny Gibbs sampler has a stationary distribution (Theorem 3.1), which also exhibits the strong selection consistency property (Theorem 3.6).

– Skinny Gibbs has a computational complexity of np for each iteration and is more scalable than existing Gibbs sampling algorithms for variable selection.

– The techniques used in the paper provide a theoretical framework for studying high dimensional Bayesian model selection for cases when the likelihood-prior combination is not conjugate and when closed-form posterior expressions are not available.

The rest of the paper is organized as follows. In Section 2, we describe our model setup, including the prior distributions and the standard Gibbs sampler, and then propose Skinny Gibbs as a new model selection algorithm. In Section 3, we present the strong selection consistency results for the proposed method. In Section 4, we compare the proposed Skinny Gibbs approach to model selection with a number of leading penalization methods in simulated settings. In Section 5, we present empirical studies on two examples to demonstrate how the proposed methodology works with real data. We provide a conclusion in Section 6. In the supplementary materials, we provide proofs for all the theorems, several additional results, and discussions including a discussion about the connection between Skinny Gibbs and L_0 penalization and about an implementation of Skinny Gibbs using Polya-Gamma scale mixture of the logistic distribution (Polson et al. 2013). The supplementary materials also include discussions about marginal posterior plots for the empirical studies to demonstrate stability and convergence of the Skinny Gibbs chain, and additional simulation studies.

2 Variable selection with continuous spike-slab priors

In this section, we first describe the framework for Bayesian variable selection followed by our proposed Skinny Gibbs method. We focus on the logistic regression for the sake of simplicity in presentation, but the method is applicable to linear, probit, and other generalized linear models. Our data consist of an $n \times 1$ binary response vector denoted by $\mathbf{E} = (E_1, \cdots, E_n)^T$ and an $n \times p_n$ design matrix X. We use p_n for the model dimension in our methodological and theoretical development to emphasize its dependence on the sample size n. We assume that the columns of X are standardized to have zero mean and unit variance. We use x_i to denote the i^{th} row of X, which contains covariates for the i^{th} response E_i. Moreover, X_A will be used to denote the $n \times |A|$ dimensional submatrix of X containing the columns indexed by A, and $|A|$ is the cardinality of A. Logistic regression models the conditional distribution of E given X with the logit link, that is,

$$P[E_i = 1|x_i] = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}, \ i = 1, \cdots, n, \tag{1}$$

for some unknown parameter $\beta \in R^{p_n}$. The logistic model is one of the most widely used statistical models for binary outcomes. This paper attempts to address the problem of variable selection when the number of predictors p_n is large. If p_n is large relative to n, even the estimation problem is ill posed without any further assumptions on the model parameters. We work under the assumption that there is a true parameter vector β that is sparse in the sense that it has only a small number of non-zero components. Even under this assumption, it is a challenge to find the active predictors in the model.

In the Bayesian variable selection literature, spike and slab priors on β are commonly used. The idea is to introduce binary latent variables Z_j for the j-th component of β, which indicates whether the j^{th} covariate is active (i.e., having a nonzero coefficient). Then, priors

on β_j given Z_j are specified as

$$\beta_j \mid Z_j = 0 \sim \pi_0(\beta_j); \quad \beta_j \mid Z_j = 1 \sim \pi_1(\beta_j), \tag{2}$$

where π_0 and π_1 are called the spike and slab priors, respectively. We refer to George and McCulloch (1993), Ishwaran and Rao (2005) and Narisetty and He (2014) for further details. For linear regression with Gaussian errors, both the spike and slab priors are often taken to be Gaussian with a small and a large variance, respectively. An advantage of this approach for linear regression is that the conditionals of the Gibbs sampler are standard distributions due to conjugacy of those priors. Though they are not conjugate for the logistic model, the well-known normal scale mixture representation of the logistic distribution due to Stefanski (1991) enables us to derive the conditional distributions used in the Gibbs sampler. More specifically, let Y_i follow the logistic distribution with location parameter $x_i\beta$, and $E_i = \mathbb{1}_{\{Y_i>0\}}$ in distribution. Then, Y_i can be equivalently represented as

$$Y_i \mid s_i \sim N(x_i\beta, s_i^2), \;\; s_i/2 \sim F_{KS},$$

where F_{KS} is the Kolomogorov-Smirnov distribution whose CDF is given by $G(\sigma) = 1 - 2\sum_{n=1}^{\infty}(-1)^{n+1}\exp(-2n^2\sigma^2)$. By introducing the latent variables Y_i, we can implement the usual Gibbs sampler for logistic regression with simple conditionals. We will however need to draw s_i's from their conditional distributions, which we shall discuss in Section 2.2. An alternative sampler based on Polya-Gamma scale-mixture representation (Polson et al. 2013) is provided in Supplementary Materials.

2.1 Prior specification

We use the continuous spike and slab prior specification as proposed in George and McCulloch (1993). To achieve appropriate shrinkage and ensure model selection consistency, we consider the priors to be sample-size dependent (Narisetty and He 2014). The priors on the binary latent variables Z_j and the corresponding regression coefficients β_j are given by

$$\begin{aligned} &\beta_j \mid Z_j = 0 \sim N(0, \tau_{0,n}^2), \;\; \beta_j \mid Z_j = 1 \sim N(0, \tau_{1,n}^2) \\ &P(Z_j = 1) = 1 - P(Z_j = 0) = q_n, \end{aligned} \tag{3}$$

for $j = 1, \cdots, p_n$ (with independence across different j), where the constants $\tau_{0,n}^2, \tau_{1,n}^2$ and q_n are further specified below. Broadly speaking, we consider the settings where $\tau_{0,n}^2 \to 0$, and $\tau_{1,n}^2 \to \infty$ as $n \to \infty$. The specific rates for $\tau_{0,n}^2$, $\tau_{1,n}^2$ are given by Condition 3.5. The intuition behind such choices is that the inactive covariates will be identified with zero Z_j values, where small values of β_j relative to τ_{0n}^2 are truncated to zero. The diverging parameter $\tau_{1,n}^2$ forces the inactive covariates to be classified under $Z_j = 0$ because the prior probability around zero becomes negligible as $n \to \infty$. Finally, we shall use $q_n \sim p_n^{-1}$ to encourage the models to be sparse, i.e., it bounds the apriori size of $|Z| := \sum_{j=1}^{p_n} Z_j$ to be small, where Z denotes the vector of Z_j. The posterior probabilities of the binary variables Z_j will be used to select the active covariates.

In the linear regression case, Narisetty and He (2014) argued that the prior specification similar to (3), referred to as Bayesian shrinking and diffusing priors (BASAD), implies a posterior that is asymptotically similar to the L_0 penalized likelihood. More specifically, when $n\tau_{0,n}^2 = o(1)$, the prior parameters imply a penalty in the order of $\log(\sqrt{n}\tau_{1,n}q_n^{-1})$ for each additional covariate added in the model. In this paper, we propose a fast and scalable Gibbs sampler that preserves the similarity to the L_0 penalty and achieves the

strong selection consistency (see Section 3).

2.2 Gibbs sampler

As a prelude to our proposed Skinny Gibbs sampler, we first present the usual Gibbs sampler corresponding to (3), which will provide motivation for our proposal of Skinny Gibbs. **In the rest of the paper, all the distributions are conditional on X but we suppress it in the notations for convenience.** By considering

$$E_i = \begin{cases} 1 & \text{if } Y_i \geq 0 \\ 0 & \text{if } Y_i < 0 \end{cases} \tag{4}$$

$$Y_i \overset{\text{ind}}{\sim} N(x_i\beta, s_i^2), \qquad s_i/2 \overset{\text{ind}}{\sim} F_{KS},$$

together with the priors in (3), the joint posterior of β, Z, Y and $W = Diag\left(s_1^{-2}, ..., s_{p_n}^{-2}\right)$ is given by

$$\begin{aligned} f(\beta, W, Y, Z \mid \mathbf{E}) \;\propto\; & \textstyle\prod_{i=1}^{n} \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{E_i = \mathbb{1}\{Y_i \geq 0\}\}\; g(s_i) \\ & \times \textstyle\prod_{j=1}^{p_n} \left((1-q_n)\pi_0(\beta_j)\right)^{1-Z_j} \left(q_n\pi_1(\beta_j)\right)^{Z_j}, \end{aligned} \tag{5}$$

where $\pi_0(x) = \phi(x, 0, \tau_{0,n}^2)$, $\pi_1(x) = \phi(x, 0, \tau_{1,n}^2)$, $\phi(x, \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2 evaluated at x, and $g(\sigma) = (d/d\sigma)F_{KS}(\sigma/2)$ is the density function of two times the KS variable.

The conditional distribution of β is given by

$$\beta \mid (W, Y, Z, \mathbf{E}) \sim N\left((X'WX + D_z)^{-1}X'WY, (X'WX + D_z)^{-1}\right), \tag{6}$$

where $D_z = Diag(Z\tau_{1,n}^{-2} + (1-Z)\tau_{0,n}^{-2})$. The conditional distributions of Y_i are independent

with the marginals given by

$$f(Y_i \mid \beta, W, Z, \mathbf{E}) \propto \begin{cases} \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{Y_i > 0\} & \text{if } E_i = 1, \\ \phi\left(Y_i, x_i\beta, s_i^2\right) \mathbb{1}\{Y_i < 0\} & \text{if } E_i = 0, \end{cases} \tag{7}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. The conditional distributions of Z_j are independent (across j) and given by

$$P(Z_j = 1 \mid \beta, W, Y, \mathbf{E}) = \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1 - q_n)\phi(\beta_j, 0, \tau_{0,n}^2) + q_n \phi(\beta_j, 0, \tau_{1,n}^2)}. \tag{8}$$

The conditional distribution of W is described in terms of the independent distributions of s_i as

$$f(s_i \mid \beta, Y, Z, \mathbf{E}) \propto \phi(Y_i, x_i\beta, s_i^2)\ g(s_i). \tag{9}$$

In this Gibbs sampler, sampling from the distribution (9) is not as straightforward as the others. Holmes and Held (2006) proposed a rejection sampling algorithm. Albert and Chib (1993) noted that the univariate logistic density can be approximated well by a t-density. O'Brien and Dunson (2004) later used the t-approximation of the logistic density for multivariate logistic regression. We simply adopt those ideas to proceed as follows.

Let us denote the t-distribution that approximates the KS distribution by $T = w\ t_\nu$, i.e., t with ν degrees of freedom and scale parameter w. Due to the Gaussian scale mixture representation of the t-distribution, it can be equivalently represented as

$$T|\phi \sim N(0, \phi^2),\ \ \phi^2 \sim w^2 IG\left(\nu/2, \nu/2\right), \tag{10}$$

where IG is the inverse gamma distribution. Following O'Brien and Dunson (2004), we

take $w^2 = \pi^2(\nu - 2)/3\nu$ and $\nu = 7.3$ so that the resulting distribution of T is nearly indistinguishable from the KS distribution. Using this approximation, the sampling of (9) can be done using an inverse Gamma distribution. Polson et al. (2013) presented an alternative scale mixture representation of the logistic distribution with a Polya-Gamma distribution as the scale mixture, which does not require the introduction of the latent variables Y. As this is also a normal scale mixture representation, the corresponding Gibbs sampler obtained would have a conditional distribution for β which still requires sampling from a p_n dimensional multivariate normal distribution similar to Equation (6). The Skinny Gibbs algorithm we propose in the next section also generalizes if we use Polya-Gamma scale mixture representation and the details of this algorithm are provided in the Supplementary Materials.

When p_n is large, the real bottleneck with the usual Gibbs sampler lies in its need to sample from the p_n-variate normal distribution for β given by (6). For linear regression, Guan and Stephens (2011) avoided such sampling by integrating β out and devise an MCMC method that samples Z directly. However, this technique does not seem to generalize easily to logistic regression. A direct sampling scheme would require handling a $p_n \times p_n$ covariance matrix of general forms, which is expensive in both CPU and memory. Even if this task is decomposed into componentwise sampling by a further Gibbs iteration, it requires operations in the order of p_n^2, making Bayesian model selection algorithm less competitive with the penalty based optimization methods.

2.3 Skinny Gibbs algorithm

We propose the Skinny Gibbs algorithm as a simple yet effective modification of the Gibbs sampler to avoid the computational complexity in the case of large p_n. The idea is to split β into two parts in each Gibbs iteration, corresponding to the "active" (with the current

$Z_j = 1$) and "inactive" (with the current $Z_j = 0$) sub-vectors. The active part has a low dimension, and is sampled from the multivariate normal distribution. The inactive part has a high dimension, but we simply sample it from a normal distribution with independent marginals. More specifically, the Skinny Gibbs sampler proceeds as follows, after an initialization.

(a) Decompose $\beta = (\beta_A, \beta_I)$, where β_A and β_I contain the components of β corresponding to $Z_j = 1$ and $Z_j = 0$, respectively. Similarly let $X = [X_A, X_I]$. Then, generate

$$\beta_A \mid (W, Y, Z, \mathbf{E}) \sim N(m_A, V_A^{-1}), \qquad \beta_I \mid (W, Y, Z, \mathbf{E}) \sim N(0, V_I^{-1}), \tag{11}$$

where $V_A = (X_A' W X_A + \tau_{1n}^{-2} I)$, $m_A = V_A^{-1} X_A' W Y$, and $V_I = Diag\,(X_I' X_I + \tau_{0n}^{-2} I) = (n + \tau_{0n}^{-2}) I$. Note that the dimension of V_A is only $|Z|$.

(b) Generate Z_j $(j = 1, \cdots, p_n)$ sequentially based on

$$\begin{aligned} &\frac{P[Z_j = 1 \mid Z_{-j}, \beta, W, Y, \mathbf{E}]}{P[Z_j = 0 \mid Z_{-j}, \beta, W, Y, \mathbf{E}]} \\ &= \frac{q_n \phi(\beta_j, 0, \tau_{1,n}^2)}{(1 - q_n)\phi(\beta_j, 0, \tau_{0,n}^2)} \times \exp\left\{ \beta_j X_j' W (Y - X_{C_j} \beta_{C_j}) + \frac{1}{2} X_j' (I - W) X_j \beta_j^2) \right\}, \end{aligned} \tag{12}$$

where Z_{-j} is the Z vector without the jth component, and C_j is the index set corresponding to the active components of Z_{-j}, i.e., $C_j = \{k : k \neq j, Z_k = 1\}$.

(c) The conditional distribution of Y is changed to

$$f(Y_i \mid \beta, W, Z, \mathbf{E}) \propto \begin{cases} \phi\left(Y_i, x_{Ai}\beta_A, s_i^2\right) \mathbb{1}\{Y_i > 0\} & \text{if } E_i = 1, \\ \phi\left(Y_i, x_{Ai}\beta_A, s_i^2\right) \mathbb{1}\{Y_i < 0\} & \text{if } E_i = 0, \end{cases} \tag{13}$$

(d) The conditional distribution of s_i is

$$f(s_i \mid \beta, Y, Z, \mathbf{E}) \propto \phi(Y_i, x_{Ai}\beta_A, s_i^2)\ g(s_i). \tag{14}$$

In (a), the update of β is changed such that the coefficients corresponding to $Z_j = 1$ (denoted by β_A) and those corresponding to $Z_j = 0$ (denoted by β_I) are sampled independently. Furthermore, the components of β_I are updated independently. This is in contrast with the usual Gibbs, where the entire β is updated jointly. It is worth noting that the precision matrix of β_A is just the corresponding sub-matrix of the precision matrix of β, which is $V_z = (X'WX + D_z)$. Essentially, Skinny Gibbs sparsifies the precision matrix V_z as

$$V_z = \begin{pmatrix} X_A'WX_A + \tau_{1n}^{-2}I & X_A'WX_I \\ X_I'WX_A & X_I'WX_I + \tau_{0n}^{-2}I \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} X_A'WX_A + \tau_{1n}^{-2}I & 0 \\ 0 & (n + \tau_{0n}^{-2})I \end{pmatrix}.$$

This modification in step (a) alters the Gibbs sampler in such a non-trivial way that the correlation structure among the coefficients β_j is lost. Without any compensation, the modified sampler would not converge to a desirable stationary distribution. The step (b) of the proposed Skinny Gibbs is designed to compensate for the loss in step (a), but the computational complexity in step (b) is minimal.

Remark 1 (Computational complexity)**.** The computational complexity of Skinny Gibbs for each iteration is $n(p_n \vee |A|^2)$ where $|A|$ is the active model size. This is much faster than the typical complexity of $p_n^2(p_n \vee n)$ for BASAD. Bhattacharya et al. (2016) recently

proposed an alternative sampling algorithm for structured high dimensional multivariate Gaussian distributions whose complexity is $n^2 p_n$ if the weight matrix W is diagonal, whereas the complexity is np_n^2 if W is non-sparse.

Remark 2 (Prior choices)**.** Several other commonly used prior specifications on either β or Z can be incorporated in Skinny Gibbs. While we use the independent prior specification for the $Z's$, Scott and Berger (2010); Castillo and van der Vaart (2012) place a Beta prior on q_n instead of treating it as a hyperparameter. The Skinny Gibbs algorithm can be easily implemented with a Beta prior on q_n as the full conditional of q_n will also be a Beta distribution. Instead of Gaussian priors, Gelman et al. (2008) placed weakly informative t prior distributions on the regression coefficients. Implementation of Skinny Gibbs for the alternative prior choices on β and q_n can be derived in a similar way.

3 Theoretical results

In this section, we study the model selection properties of the proposed Skinny Gibbs algorithm. We show that Skinny Gibbs has a stationary posterior distribution that preserves the strong model selection consistency. We first introduce the following notations.

Notations: We use k (and s) to denote a generic model and t to denote the true model. A model is treated both as a $p_n \times 1$ binary vector similar to Z and as the set containing the active covariates, but this will be clear depending on the context. The size of the model k is denoted by $|k|$. For any $p_n \times 1$ vector v, $v(k)$ is used to denote the $|k| \times 1$ vector containing the components of v corresponding to model k. We denote the true regression vector as $\beta_0(t)$, and for any $k \supset t$, $\beta_0(k)$ denotes the $|k| \times 1$ vector having $\beta_0(t)$ for t and zeroes for $k \cap t^c$. For sequences a_n and b_n, $a_n \sim b_n$ means $\frac{a_n}{b_n} \to c$ for some $c > 0$, $b_n \succeq a_n$ (or $a_n \preceq b_n$) means $b_n = O(a_n)$, and $b_n \succ a_n$ (or $a_n \prec b_n$) means $b_n = o(a_n)$.

The log-likelihood for a model k is

$$L_n(\beta(k)) := \sum_{i=1}^{n} E_i \log F\left(x_i\beta(k)\right) + (1 - E_i) \log(1 - F\left(x_i\beta(k)\right)), \tag{15}$$

where $F(.)$ is the cdf of the logistic distribution. Let

$$s_n(\beta(k)) = \frac{\partial L_n(\beta(k))}{\partial \beta(k)} = \sum_{i=1}^{n} \left(E_i - \mu_i(\beta(k))\right) x_i, \tag{16}$$

with $\mu_i(\beta(k)) = \frac{\exp\{x_i\beta(k)\}}{1+\exp\{x_i\beta(k)\}}$. The negative Hessian of $L_n(\beta(k))$ is

$$H_n(\beta(k)) = -\frac{\partial^2 L_n(\beta(k))}{\partial \beta(k) \partial \beta(k)'} = \sum_{i=1}^{n} \sigma_i^2(\beta(k)) x_i x_i', \tag{17}$$

where $\sigma_i^2(\cdot) = \mu_i(\cdot)(1 - \mu_i(\cdot)$. Note that in our notations, x_i and X are restricted to the model under consideration, even though it is not explicitly displayed. That is, x_i in Equations (16) and (17) is a $|k| \times 1$ vector containing the components corresponding to model k. Therefore, the dimension of $s_n(\beta(k))$ is $|k| \times 1$ and that of $H_n(\beta(k))$ is $|k| \times |k|$. We shall also use μ_i and σ_i^2 in place of $\mu_i(\beta_0(t))$ and $\sigma_i^2(\beta_0(t))$, respectively, for the sake of simplicity.

We first prove the following to provide the posterior that corresponds to the Skinny Gibbs sampler.

Theorem 3.1. *The joint posterior of β, Z, Y and W corresponding to the Skinny Gibbs algorithm is given by*

(18)

$$\begin{aligned} f(\beta, W, Y, Z = k \mid \mathbf{E}) \propto \quad & |W|^{1/2} \exp\left\{-\tfrac{1}{2}(Y - X\beta(k))'W(Y - X\beta(k))\right\} \; v_n^{-|k|} \\ & \times \prod_i g(s_i) \exp\left\{-\tfrac{1}{2}\left(\beta' D_k \beta + n\beta(k^c)'\beta(k^c)\right)\right\} \mathbb{1}\{E_i = \mathbb{1}\{Y_i \geq 0\}\}, \end{aligned}$$

where $W = Diag\,(s_1^{-2}, ..., s_{p_n}^{-2})$, $D_k = Diag(k\tau_{1n}^{-2} + (1-k)\tau_{0n}^{-2})$ *and* $v_n = \tau_{1n}(1-q_n)/(q_n\tau_{0n})$.

Remark 3 (Connection to penalization)**.** The posterior (18) suggests that with everything else the same, a unit increase in the model size ($|k|$) reduces the posterior by a multiple of $v_n^{-1} = (q_n\tau_{0n})/\tau_{1n}(1-q_n)$. This hints at the following: (a) the similarity of the posterior to L_0 penalty as discussed in Supplementary Materials, and (b) the reason for allowing the prior parameters to depend on n (see Condition 3.5) so that the shrinkage implied by v_n^{-1} is at an appropriate level.

We now provide the conditions assumed for proving strong selection consistency property of Skinny Gibbs. By strong selection consistency, we mean that the posterior probability of the true model converges to one as sample size increases to infinity, as used in Johnson and Rossell (2012) and Narisetty and He (2014).

Condition 3.2 (On Dimension p_n)**.** $p_n \to \infty$ *and* $\log p_n = o(n)$ *as* $n \to \infty$.

Condition 3.3 (On Regularity of the Design)**.**

(a) The predictors are bounded, that is, $\max\{|x_{ij}|,\ 1 \le i \le n, 1 \le j \le p_n\} \le C$, *for some* $0 < C < \infty$;

(b) for some fixed $0 \le d < d' \le 1$,

$$0 < \lambda \le \min_{k:|k|\le m_n+|t|} \lambda_{min}\left(n^{-1}H_n(\beta_0(k))\right) \le \max_{k:|k|\le m_n+|t|} \lambda_{max}\left(n^{-1}X_k'X_k\right) \le C^2\left(\frac{n}{\log p_n}\right)^d,$$

where, $\lambda_{min}(\cdot), \lambda_{max}(\cdot)$ *are the minimum and maximum eigenvalues of their arguments respectively,*

$$m_n := \left(\left(\frac{n}{\log p_n}\right)^{\frac{1-d'}{2}} \wedge p_n\right),$$

and

(c) for any possible model k with $|k| \leq m_n + |t|$ and any $u \in R^n$ in the space spanned by the columns of $\Sigma^{1/2} X_k$, there exists $\delta^* > 0$ and $N(\delta^*)$ such that

$$\mathbb{E}\left[\exp\{u'\Sigma^{-\frac{1}{2}}(\mathbf{E}-\mu)\}\right] \leq \exp\left\{\frac{(1+\delta^*)u'u}{2}\right\},$$

for any $n \geq N(\delta^*)$, where $\mathbb{E}$ denotes expectation over $\mathbf{E}$ (conditional on the design).

Condition 3.4 (On True Model and Signal Strength)**.** *We assume that there exists constant $c > 1$ such that*

$$c|t| \leq m_n, \quad \text{and} \quad \min_{1\leq i\leq |t|} |\beta_{0i}(t)| \geq \sqrt{\frac{c|t|\Lambda_{c|t|}\log p_n}{n}},$$

where $\beta_0(t) = (\beta_{0i}(t))_{i=1}^{|t|}$ is the nonzero coefficients of β under the true model, and $\Lambda_{c|t|} := \max_{k:|k|\leq c|t|} \lambda_{max}\left(n^{-1}X_k'X_k\right)$.

Condition 3.5 (Prior Parameters)**.** *The prior parameters τ_{0n}^2, τ_{1n}^2 and q_n are such that for some $\delta > \delta^*$,*

$$n\tau_{0n}^2 = o(1), \quad n\tau_{1n}^2 \sim (n \vee p_n^{2+2\delta}), \quad q_n \sim p_n^{-1}.$$

The upper bound on the maximum eigenvalue in Condition 3.3 (b) is always satisfied if $1/3 < d < d'$. This is because, $\lambda_{max}\left(n^{-1}X_k'X_k\right) \leq Trace\left(n^{-1}X_k'X_k\right) \leq C^2|k| \leq C^2(n/\log p_n)^d$ holds for any $|k| \leq m_n + |t|$ when $1/3 < d < d'$. This condition is weaker than the bounded maximum eigenvalue condition as assumed in Bondell and Reich (2012). If the maximum eigenvalue here is bounded, we have the case $d = 0$, and m_n can be almost as large as $(n/\log p_n)^{1/2}$.

The lower bound in Condition 3.3 (b) is essentially a restricted eigenvalue condition for L_0-sparse vectors. Restricted eigenvalue (RE) conditions are routinely assumed in high-dimensional theory to guarantee some level of curvature of the objective function in lower dimensions. However, our restricted eigenvalue condition for L_0-sparse vectors is

weaker than irrepresentable conditions required for consistency of Lasso-type L_1 penalization methods (Zhao and Yu 2006; Liang et al. 2013). The lower bound in Condition 3.3 (b) is satisfied by sub-Gaussian random design matrices with high probability. A formal statement about this is stated and proved as Lemma A.4 in Supplementary Materials. The intuition behind the L_0-sparse eigenvalue condition for Skinny Gibbs is attributable to the similarity between Skinny Gibbs and the L_0 type penalization as discussed in Supplementary Materials.

Condition 3.3 (c) is not really a restriction, because such a $\delta^* > 0$ always exists due to the sub-Gaussianity of $\Sigma^{-\frac{1}{2}}(\mathbf{E} - \mu)$. Also note that for typical random designs, the variable $u'\Sigma^{-\frac{1}{2}}(\mathbf{E} - \mu)/||u||$ is asymptotically distributed as $N(0, 1)$, so Condition 3.3 (c) is expected to hold for a small positive constant δ^*. In Condition 3.4, the upper bound on the true model size, and the minimum signal strength match with those for penalized methods such as Lasso when $d = 0$, but impose slightly stronger conditions when $d > 0$. For the screening property of Lasso to hold, Corollary 7.6 of Bühlmann and van de Geer (2011) assumes the minimum signal to be at least in the order of $\sqrt{|t| \log p_n/n}$ and the true model size $|t| = O(\sqrt{n/\log p_n})$.

Theorem 3.6. *Under Conditions 3.2 – 3.5, we have*

$$P\left[Z = t \mid (\mathbf{E},\ \textit{and}\ |Z| \leq m_n)\right] \xrightarrow{\mathrm{P}} 1,\ \textit{as}\ n \to \infty.$$

Moreover, $\sum_{k \neq t; |k| \leq m_n} \frac{P[Z=k|\mathbf{E}]}{P[Z=t|\mathbf{E}]} \leq C \exp\{-\epsilon \log p_n\} \to 0$ *for some* $C,\ \epsilon > 0$.

The strong selection consistency (SSC) is a stronger property than the usual Bayes factor consistency as well as the model selection consistency considered for penalization methods. As Johnson and Rossell (2012) argued (see proof of Theorem 2 there) that for large $p_n > n^{1/2+\epsilon}$, the posterior of the true model relative to models of a fixed size may

also be very small, i.e., it is possible that $\sum_{k\neq t;|k|=|t|+1} \frac{P[Z=k|\mathbf{E}]}{P[Z=t|\mathbf{E}]} \to \infty$, even under the Bayes factor consistency. This will make it difficult to identify the active predictors based on a finite chain, because the posterior probability of the true model can be close to zero, so that the ratios $P[Z = k \mid \mathbf{E}]/P[Z = t \mid \mathbf{E}]$ are difficult to estimate. In a Bayesian context, the model selection consistency for penalization methods corresponds to showing that $P[Z = t \mid \mathbf{E}] > \sup_{k\neq t;|k|\leq m_n} P[Z = k \mid \mathbf{E}]$, with a large probability. On the other hand, SSC implies a much stronger statement since it guarantees a large "gap" between the posterior probabilities of the true model and the rest. This gap is practically useful for accurately identifying the true model.

Narisetty and He (2014) provided SSC for BASAD in the context of linear regression. Due to the conjugacy of likelihood-prior set-up of linear regression and the availability of closed-form posterior expressions, properties of Gaussian random vectors could be employed for the proofs of Narisetty and He (2014). On the other hand, a proof of Theorem 3.6 for showing SSC for Skinny Gibbs in the logistic regression context requires substantial technical developments with careful use of concentration inequalities since the posterior of Z is only available in the form of high dimensional integrals. For more details about the proof, we refer to the Supplementary Materials.

Remark 4 (True model)**.** For the sake of convenience, we assume that the true model representation t is unique. If multiple representations of the true model are available (due to the existence of linearly dependent predictors) the result of Theorem 3.6 holds if t represents the union of the true model's representations.

Remark 5 (Model size restriction)**.** Although Theorem 3.6 provides strong selection consistency conditional on $|Z| < m_n$, models of size larger than $|Z| \geq m_n$ can be completely avoided by restricting the prior distribution on $|Z| < m_n$ as commonly done in the literature (Liang et al. 2013). In practice, we restrict the model size apriori to $\max(30, \sqrt{n})$.

Remark 6 (Marginal posterior probabilities)**.** Theorem 3.6 justifies the use of marginal posterior probabilities $P[Z_j \mid \mathbf{E}]$ for selecting the variables. This is useful in practice because we only need to estimate and store p_n marginal posterior probabilities as opposed to dealing with posterior probabilities of $\binom{p_n}{m_n}$ models.

3.1 Comparisons with existing Bayesian methods

Chen and Chen (2012) proposed the extended Bayesian Information Criterion (EBIC)

$$EBIC(k) = -2L_n(\hat{\beta}(k)) + |k|(\log n + 2\gamma \log p_n) \tag{19}$$

for model selection, which is similar to the regular BIC except for an additional penalization term depending on p_n. The model selection consistency under EBIC is established by Chen and Chen (2012) for $\gamma > (1 - \frac{1}{2\kappa})$, where $p_n = O(n^\kappa)$. For high dimensional problems, such an objective function cannot be applied to all possible models. Even if we restrict ourselves to a model of size m for a relatively small m, the number of possible models $\binom{p_n}{m}$ could be too large. The EBIC is typically used to choose models among a much smaller number of candidate models.

An alternative approach to Gaussian spike priors used in this paper is to take point mass spike priors, i.e., $\beta_j \mid Z_j = 0 \sim \delta_0$, the point-mass distribution at zero. An apparent attraction of the point mass prior is that we no longer have to deal with $p_n \times p_n$ matrix computations, if we can sample from the posterior of Z_j without β_j. This can indeed be done in linear regression models, as shown in Guan and Stephens (2011). Unfortunately, the posterior $P(Z = k \mid \mathbf{E})$ does not have a closed-form for logistic regression, and approximations have to be used for sampling from the posterior. In this direction, Hans et al. (2007) proposed a shotgun stochastic search (SSS) algorithm based on a Laplace approximation to the posterior. Liang et al. (2013) proposed Bayesian subset regression (BSR)

modeling using a stochastic approximation Monte Carlo (SAMC) algorithm (Liang et al. 2007) that aims to avoid the potential local-trap problem for SSS by sampling a specified sub-regions of the model space uniformly. Like Skinny Gibbs, these algorithms avoid p_n^2 operations in each step of the iteration, but they are analogous to stepwise variable selection whereas Skinny Gibbs allows more general updates of the model in every iteration. More specifically, for stepwise algorithms using point mass priors if the current model is k, the model in the next step can only be a model that has at most one variable different from k restricting it to be one of $p_n(|k|+1)+|k|$ models when the total number of possible models is 2^{p_n}. Skinny Gibbs has the potential to update any number of Z components at each step, making the search more efficient. Also, for the SAMC algorithm to be competitive, the number of sub-regions used in the method needs to increase with p_n, making it less computationally competitive. Some empirical comparisons of various methods are given in Section 4.

For the strong selection consistency we established here, the spike prior variance τ_{0n}^2 can be arbitrarily close to zero making the limiting case of $\tau_{0n}^2 = 0$ the same as the point-mass prior for $\beta_j \mid Z_j = 0$. We note that Liang et al. (2013) used a point-mass spike prior and a slab prior whose variance depends on the size of the model, and showed strong selection consistency. However, the consistency result of Liang et al. (2013) relied on a condition on the posterior distribution itself, which makes their result indicative rather than confirmatory. In this sense, we hope that our theoretical treatment also completes the strong selection consistency theory on point-mass priors in high dimensional models.

Geweke (1996) proposed a sampling approach where $Z_i's$ are sampled using conditional Bayesian factors given the $\beta's$, followed by a conditional sampling of corresponding active $\beta's$ from univariate truncated normal distributions. This approach is similar to the block update of Ishwaran and Rao (2005) with as many blocks as the number of variables. Geweke (1996) and Ishwaran and Rao (2005) provide consistency of the posterior in

terms of convergence of Bayes factors whereas we show a stronger version of consistency for Skinny Gibbs in the form of strong selection consistency. Recently, Bhattacharya et al. (2016) proposed an alternative sampling approach whose typical complexity is $n^2 p_n$, which is slightly slower than Skinny Gibbs as discussed in Remark 1. Carbonetto and Stephens (2012) proposed a variational Bayesian algorithm which is also scalable. Skinny Gibbs has similarity with variational Bayes methods in the sense that both provided computationally motivated approximation to the posterior. The main difference is that Skinny Gibbs performs sampling rather than optimization and even though it was motivated as an approximation, Skinny Gibbs posterior itself can be shown theoretically to have strong selection consistency property.

4 Simulation study

In this section, we study the performance of the proposed method and compare them with several existing methods by simulation studies. Let X denote the design matrix whose first $|t|$ columns correspond to the active covariates for which we have nonzero coefficients, while the rest correspond to the inactive ones with zero coefficients. In this section and thereafter, we shall use $p := p_n$ to denote the number of covariates. In all the simulations, we generate each row of X independently from a normal distribution with a p-dimensional covariance matrix such that the correlation between any pair of active covariates is equal to ρ_1, the correlation between an active covariate and an inactive covariate is ρ_2, and the correlation between any pair of inactive covariates is ρ_3. Given X, we sample Y from a logistic model $P(Y_i = 1|x_i) = e^{x_i\beta}/(1 + e^{x_i\beta})$, for $i = 1, \ldots, n$. We will specify the number of observations n, the number of covariates p, the correlations ρ_1, ρ_2, and ρ_3 in the tables. We will also mention the number of active covariates $|t|$ and the corresponding active coefficients β_t in the tables.

We report the results from the usual Gibbs sampler described in Subsection 2.2 (BASAD), and Skinny Gibbs (as a simplified version of BASAD), Bayesian Subset Regression (BSR, Liang et al. (2013)), variational Bayes (Carbonetto and Stephens 2012), Adaptive Lasso, SCAD, as well as MCP. For BSR, we set its hyperparameter $\gamma = 1$. For variational Bayes algorithm, we choose the recent R package "varbvs" for implementation, and select the median probability model based on the variational posterior inclusion probabilities. We use the R package "glmnet" for Adaptive Lasso, and the package "ncvreg" for SCAD and MCP. For Adaptive Lasso, the initial estimate of β is obtained from Lasso tuned using BIC. For all the penalization methods, BIC is used to select the tuning parameters. More specifically, we select the model which corresponds to the first local minimum of BIC since it produces a sparser and more accurate model compared to the model that minimizes BIC globally. While it is possible to use EBIC for tuning the penalization methods instead of the classical BIC criterion, we use BIC because EBIC performed worse in our studies due to model under-fitting. For the BASAD and Skinny Gibbs, we have three parameters to choose: τ_{0n}^2, τ_{1n}^2 and q_n. In all our empirical work, we use

$$\tau_{0n}^2 = \tfrac{1}{n},\ \tau_{1n}^2 = \max\left(\tfrac{p^{2.1}}{100n}, 1\right),$$

and we choose $q_n = P[Z_i = 1]$ such that $P\left[\sum_{i=1}^p Z_i = 1 > K\right] = 0.1$, for a pre-specified value of K. Our default value is $K = \max(10, \log(n))$. These choices are very similar to the implementation of BASAD for linear models in Narisetty and He (2014). We would like to mention that instead of specifying the hyperparameters, a different approach is to tune one or more hyperparameters. For instance, a few different values for τ_{1n} around the default parameter may be used to obtain different models followed by using a criterion such as BIC to select the final model. In our empirical experience, such methods do not always outperform default choices.

For BASAD and Skinny Gibbs, we obtain the results based on averaging 10 Gibbs chains each with a burn-in of 2000 and a length of 5000. The maximum model size along the chain is restricted to be $\max(30, \sqrt{n})$. The models for BASAD and Skinny Gibbs are obtained by thresholding the marginal posterior probabilities at 0.5, which is referred to as the median probability model by Barbieri and Berger (2004).

We will present the following model selection performance measures using 100 randomly generated datasets. Average True Positive (TP) is the average number of active covariates chosen; TP_s is the average number of active covariates selected if the sparsity level is known, that is, if we select a model with exactly $|t|$ covariates; Average False Positive (FP) is the average number of inactive covariates chosen; The column $Z = t$ gives the proportion of choosing the true model exactly, while $Z \supset t$ is the proportion of times the true model is included in the chosen model; Finally, the column $Z_s = t$ gives the proportion of times the model chosen with size $|t|$ is the true model. We note that the measures TP_s and $Z_s = t$ indicate how well a method can order the active covariates ahead of the inactive ones, and do not depend much on the specific choice of the tuning parameters involved.

In Table 1, we have four cases corresponding to the number of covariates $p = 50$ or 250 and with a common correlation of $\rho_1 = \rho_2 = \rho_3 = 0$ or 0.25. The results show that BASAD and Skinny Gibbs, like other Bayesian model selection methods, have much smaller false positives than non-Bayesian methods, and do not lose much in terms of true positives. Overall, our proposed methods have higher exact identification rate ($Z = t$) but none of the methods dominate others in all the measures. Variational Bayes method also provides a competitive performance in terms of selecting the true model.

In our next simulation settings, we consider a less sparse case with $|t| = 8$ with all the other aspects of data generating model the same as that of Table 1. We observe that the performance of all the methods deteriorates in comparison to the sparser case (Table 1) as expected. However, the effects on our methods are less substantial than on the other

Table 1: Simulation results with high sparsity ($|t| = 4$ active variables): TP $\rightarrow$ True Positive; TP_s: true positives with known sparsity level, i.e., for the model selected with size equal to $|t| = 4$; FP $\rightarrow$ False Positive; $Z = t \rightarrow$ The proportion of times true model is selected; $Z \supset t \rightarrow$ The proportion of times true model is included; $Z_s = t \rightarrow$ The proportion of times the chosen model of size $|t| = 4$ is the true model. The active coefficients are $\beta_t = (-1.5, 2, -2.5, 3)$.

(a) $n = 100;\ p = 50;\ \rho_1 = \rho_2 = \rho_3 = 0; \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	3.71	3.78	0.19	0.59	0.71	0.78
Skinny	3.82	3.76	0.42	0.53	0.82	0.76
VBayes	3.61	3.79	0.07	0.59	0.62	0.79
BSR	3.45	3.85	0.15	0.47	0.54	0.85
Alasso	3.79	3.70	0.84	0.35	0.79	0.64
SCAD	3.83	3.63	1.24	0.25	0.83	0.60
MCP	3.87	3.69	2.11	0.12	0.87	0.64
(b) $n = 100;\ p = 50;\ \rho_1 = \rho_2 = \rho_3 = 0.25; \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	3.68	3.72	0.23	0.58	0.69	0.72
Skinny	3.74	3.72	0.46	0.51	0.75	0.72
VBayes	3.48	3.72	0.09	0.54	0.56	0.73
BSR	3.51	3.70	0.09	0.62	0.65	0.72
Alasso	3.70	3.66	0.65	0.42	0.72	0.59
SCAD	3.74	3.55	1.11	0.32	0.75	0.49
MCP	3.79	3.64	1.96	0.18	0.80	0.59
(c) $n = 100;\ p = 250\ \rho_1 = \rho_2 = \rho_3 = 0; \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	3.66	3.73	0.24	0.53	0.69	0.76
Skinny	3.72	3.68	0.54	0.50	0.75	0.70
VBayes	3.33	3.74	0.04	0.46	0.47	0.75
BSR	2.90	3.47	0.09	0.23	0.28	0.52
Alasso	3.69	3.54	1.23	0.24	0.70	0.52
SCAD	3.69	3.39	1.75	0.17	0.71	0.40
MCP	3.82	3.57	2.69	0.13	0.83	0.56
(d) $n = 100;\ p = 250;\ \rho_1 = \rho_2 = \rho_3 = 0.25; \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	3.55	3.63	0.29	0.45	0.60	0.65
Skinny	3.62	3.64	0.56	0.38	0.65	0.66
VBayes	3.07	3.57	0.03	0.33	0.33	0.60
BSR	2.82	3.38	0.14	0.23	0.26	0.46
Alasso	3.30	3.33	1.06	0.15	0.40	0.28
SCAD	3.33	3.16	1.65	0.09	0.47	0.24
MCP	3.72	3.48	3.22	0.03	0.74	0.47

methods. For example, in Table 2, exact selection proportions ($Z = t$ and $Z_s = t$) and true positive rates (TP and TP_s) are almost always higher for Skinny Gibbs than for the competing methods. The performance of variational Bayes suffers substantially compared to the results in Table 1. Overall, the results indicate that this is a challenging setting and it is difficult to detect all the active covariates. In Figure 1, we plot the proportion of active covariates (out of the eight active ones) that are selected as a function of model sizes. Note that this plot does not depend on tuning in the penalization methods and shows that Skinny Gibbs has the largest proportion of active covariates across settings for all model sizes. We exclude BASAD in Figure 1 simply because its performance is almost identical to that of Skinny Gibbs but with much more computational time involved.

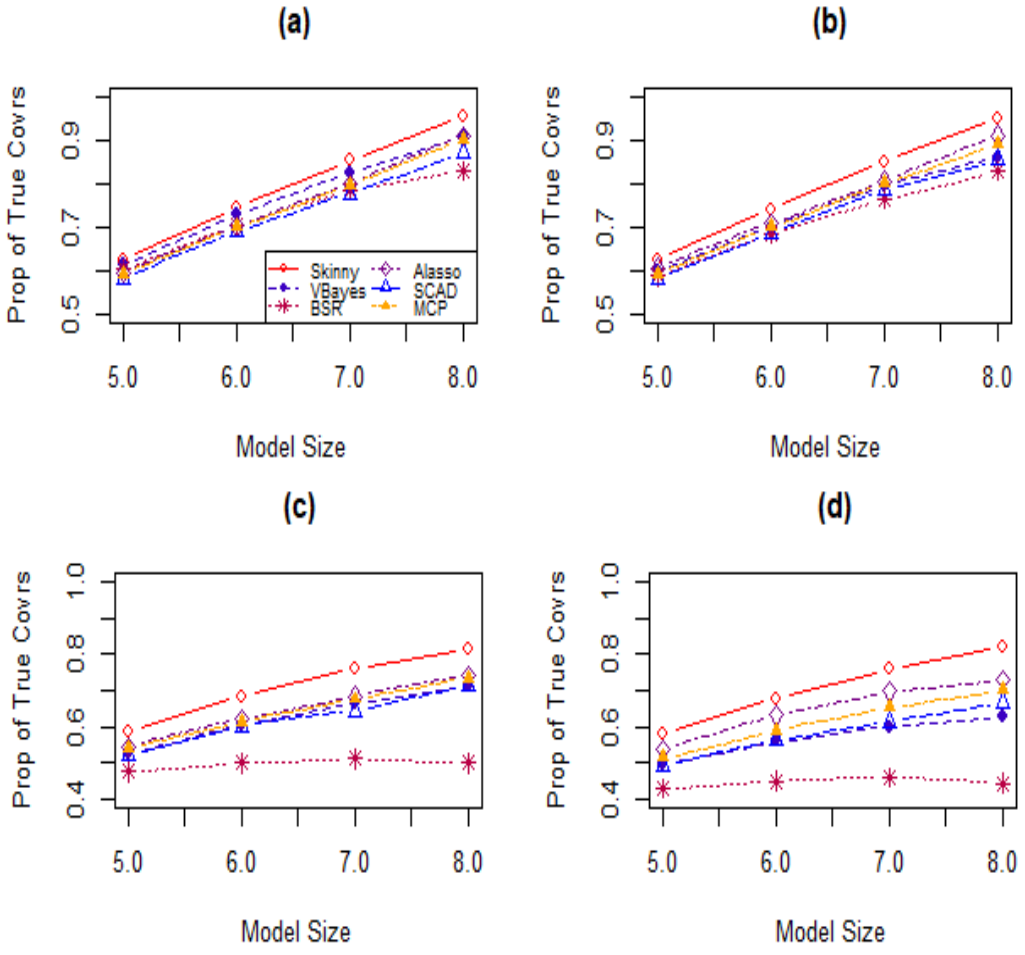
We will now consider a setting with a larger number of variables with $p = 1000$ and $n = 200$. In Table 3, we present the results for two sparsity levels $|t| = 4$ and $|t| = 8$ and two correlation settings $\rho_1 = \rho_2 = \rho_3 = 0$ and $\rho_1 = \rho_2 = \rho_3 = 0.25$. For this case, we do not report the results for BASAD and BSR as they are very time consuming and do not seem to offer significant performance gains compared to Skinny Gibbs based on the previous results. We use a burn-in of size 5000 and an iteration length of 5000 for each of the Skinny Gibbs chains. The results from Table 3 indicate that Skinny Gibbs and variational Bayes perform the best with Skinny Gibbs having a slight edge over variational Bayes for Table 3 (d) with $p = 1000$ and $|t| = 8$.

In the Supplementary Materials, we provide additional simulation results under high correlation and weak signal settings, and plots of the marginal posterior probabilities along the Skinny Gibbs iterates to demonstrate the stability in the convergence of the Skinny Gibbs chains. Based on the additional simulation results, we note that the performance of Skinny Gibbs is less affected by high correlations which can be attributed to its similarity with the L_0 penalty. It also demonstrates good performance in detecting weak signals relative to other existing methods. In summary, we conclude that Skinny Gibbs demonstrates

Table 2: Simulation results with a less sparse setting ($|t| = 8$ active variables): TP $\rightarrow$ True Positive; TP_s: true positives with known sparsity level, i.e., for the model selected with size equal to $|t| = 8$; FP $\rightarrow$ False Positive; $Z = t \rightarrow$ The proportion of times true model is selected; $Z \supset t \rightarrow$ The proportion of times true model is included; $Z_s = t \rightarrow$ The proportion of times the chosen model of size $|t| = 8$ is the true model. The active coefficients are $\beta_t = (-2, 2.143, -2.286, 2.429, -2.571, 2.714, -2.857, 3)_{8\times 1}$.

(a) $n = 100$; $p = 50$; $\rho_1 = \rho_2 = \rho_3 = 0$; $\lvert t\rvert = 8$.						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	6.95	7.62	0.10	0.30	0.34	0.65
Skinny	7.28	7.63	0.24	0.43	0.52	0.66
VBayes	5.61	7.29	0.02	0.10	0.11	0.46
BSR	4.86	6.64	0.34	0.15	0.26	0.27
Alasso	6.60	7.28	0.71	0.19	0.41	0.26
SCAD	6.82	6.97	1.07	0.15	0.49	0.20
MCP	7.43	7.23	1.18	0.24	0.69	0.39
(b) $n = 100$; $p = 50$; $\rho_1 = \rho_2 = \rho_3 = 0.25$; $\lvert t\rvert = 8$.						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	6.67	7.59	0.12	0.22	0.24	0.64
Skinny	7.23	7.61	0.23	0.40	0.51	0.66
VBayes	4.44	6.89	0.08	0.03	0.03	0.18
BSR	4.27	6.63	0.22	0.10	0.14	0.19
Alasso	5.57	7.29	0.61	0.06	0.29	0.17
SCAD	5.75	6.84	0.95	0.04	0.31	0.15
MCP	6.75	7.13	1.30	0.08	0.51	0.32
(c) $n = 100$; $p = 250$; $\rho_1 = \rho_2 = \rho_3 = 0$; $\lvert t\rvert = 8$.						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	5.69	6.47	0.43	0.12	0.12	0.21
Skinny	6.11	6.52	0.63	0.16	0.17	0.21
VBayes	2.64	5.71	0.12	0.00	0.00	0.06
BSR	1.97	4.00	0.06	0.00	0.00	0.00
Alasso	5.26	5.93	1.49	0.02	0.04	0.02
SCAD	5.33	5.69	2.11	0.02	0.05	0.02
MCP	6.06	5.88	2.37	0.04	0.16	0.03
(d) $n = 100$; $p = 250$; $\rho_1 = \rho_2 = \rho_3 = 0.25$; $\lvert t\rvert = 8$.						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
BASAD	5.37	6.42	0.34	0.06	0.08	0.19
Skinny	5.84	6.55	0.67	0.11	0.14	0.21
VBayes	1.93	5.01	0.07	0.00	0.00	0.00
BSR	2.06	3.56	0.13	0.00	0.00	0.00
Alasso	4.43	5.85	1.21	0.00	0.01	0.00
SCAD	4.57	5.33	1.41	0.00	0.01	0.00
MCP	5.21	5.60	1.96	0.00	0.10	0.00

Figure 1: Proportion of True Covariates included versus Model Size under the same settings of Table 2. The curve for Skinny Gibbs consistently stays above those for other methods indicating its better performance for variable selection. The below four plots are for the four different settings of Table 2 denoted by (a) - (d), respectively.



strong performance across a wide range of settings.

4.1 Time improvement of Skinny Gibbs

We perform a small study for time comparison between BASAD and Skinny Gibbs. We present the CPU time for different methods based on 10 data sets with sample size $n = 100$ and varying number of variables as $p = 50, 100, \ldots, 1000$. We generate data using one of our simulation settings (as in Table 1 of the paper with $\rho_1 = \rho_2 = \rho_3 = 0$). That is, we generate each row of X independently from a normal distribution with a p-dimensional identity matrix I_p. Given X, we sample E from a logistic model $P(E_i = 1|x_i) = e^{x_i\beta}/(1 + e^{x_i\beta})$, for $i = 1, \ldots, n$. We use $|t| = 4$, and $\beta_t = (1.5, 2, 2.5, 3)$. We use thinkpad s230u Twist with Intel(R) Core(TM) i7-3537U CPU@ 2.00GHz, 8.00GB memory, and Windows7 64bit. For both the methods, we use a burn-in of size 2000 and additional 5000 iterations.

In Figure 2, we plot the time (in seconds) for BASAD and Skinny Gibbs. It can be seen that the time for BASAD grows at a higher rate than Skinny Gibbs and look quadratic in p whereas the time for Skinny grows linearly in p. Computations for penalization methods are generally faster than those for Bayesian methods including Skinny Gibbs. However, in terms of computational complexity, Skinny Gibbs is competitive with penalization based methods since its complexity is also a linear function of p similar to the penalization based methods.

5 Real data examples

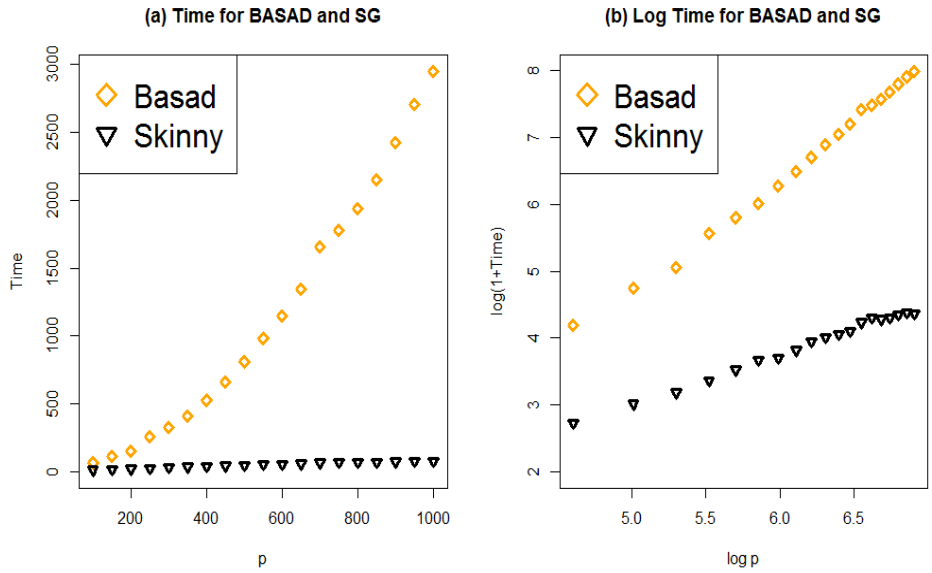
5.1 PCR dataset

We consider the data from an experiment by Lan et al. (2006) to study the genetics of two inbred mouse populations B6 and BTBR. The data include expression levels of 22,575

Table 3: Simulation results with a large number of variables ($p = 1000$):TP $\rightarrow$ True Positive; TP_s: true positives with known sparsity level, i.e., for the model selected with size equal to $|t|$; FP $\rightarrow$ False Positive; $Z = t \rightarrow$ The proportion of times true model is selected; $Z \supset t \rightarrow$ The proportion of times true model is included; $Z_s = t \rightarrow$ The proportion of times the chosen model of size equal to $|t|$ is the true model. The active coefficients are $\beta_t = (-1.5, 2, -2.5, 3)_{4\times 1}$ for $|t| = 4$ and $\beta_t = (-1.5, 1.714, -1.929, 2.143, -2.357, 2.571, -2.786, 3)_{8\times 1}$ for $|t| = 8$.

(a) $n = 200;\ p = 1000;\ \rho_1 = \rho_2 = \rho_3 = 0;\ \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
Skinny	3.95	3.97	0.18	0.82	0.95	0.97
VBayes	3.88	3.97	0.00	0.88	0.88	0.97
Alasso	3.98	3.90	1.93	0.23	0.98	0.89
SCAD	3.99	3.87	2.97	0.13	0.99	0.84
MCP	4.00	3.94	5.47	0.01	1.00	0.92
(b) $n = 200;\ p = 1000;\ \rho_1 = \rho_2 = \rho_3 = 0.25;\ \lvert t\rvert = 4.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
Skinny	3.93	3.98	0.13	0.81	0.93	0.98
VBayes	3.82	3.97	0.02	0.83	0.83	0.97
Alasso	3.87	3.87	1.30	0.24	0.87	0.80
SCAD	3.96	3.82	2.84	0.13	0.96	0.75
MCP	4.00	3.91	5.36	0.01	1.00	0.89
(c) $n = 200;\ p = 1000;\ \rho_1 = \rho_2 = \rho_3 = 0;\ \lvert t\rvert = 8.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
Skinny	7.37	7.70	0.08	0.44	0.47	0.71
VBayes	7.25	7.71	0.02	0.44	0.45	0.71
Alasso	7.06	7.11	1.19	0.11	0.32	0.19
SCAD	7.07	6.84	1.72	0.06	0.33	0.11
MCP	7.70	7.21	3.20	0.05	0.72	0.34
(d) $n = 200;\ p = 1000;\ \rho_1 = \rho_2 = \rho_3 = 0.25;\ \lvert t\rvert = 8.$						
	TP	TP_s	FP	$Z = t$	$Z \supset t$	$Z_s = t$
Skinny	7.04	7.46	0.20	0.29	0.35	0.52
VBayes	6.69	7.47	0.02	0.24	0.24	0.52
Alasso	6.68	7.00	1.04	0.11	0.23	0.20
SCAD	6.60	6.66	1.68	0.05	0.27	0.11
MCP	7.67	7.26	3.60	0.02	0.71	0.39

Figure 2: CPU time (in seconds) for BASAD and Skinny Gibbs (SG) on 10 data sets with $n = 100$ and p varies. Plot (a) shows Time as a function of p and (b) shows log (1+Time) as a function of $\log p$.



genes from 31 female and 29 male mice, resulting in a total of 60 arrays. The physiological phenotype glycerol-3-phosphate acyltransferase (GPAT) was also measured by quantitative real-time PCR. The gene expression data and the phenotypic data are publicly available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330). It is of importance to learn which genes are associated with low levels of GPAT as low levels of GPAT are found to diminish Hepatic Steatosis, a disease commonly caused by obesity (Wendel et al. 2010). For illustration, we obtain a binary response based on the variable GPAT as $\mathbf{E} = I(GPAT < Q(0.4))$, where $Q(0.4)$ is the 0.4th quantile of GPAT. The subsequent analysis will be made on the response variable $\mathbf{E}$. Due to the very large number of genes, we first perform a screening in this example but a larger p is considered in the Lymph dataset in Subsection 5.2. We use p-values obtained from the simple logistic regression of the response $\mathbf{E}$ against individual genes to select 99 marginally most significant genes, which along with the gender variable form $p = 100$ covariates. We apply Skinny Gibbs along

the Bayesian Subset Regression method (BSR) of Liang et al. (2013)), Lasso, SCAD and MCP for selecting the covariates. The results for Skinny Gibbs are based on a chain of length 4×10^4 obtained after a burn-in of length 2×10^4. The initial value for β is the zero vector and the initialization of Z contains ones for the $K = 10$ marginally most significant covariates. For BSR, the results are based on an MCMC chain of length 2×10^5 after a burn-in chain of length 5×10^4.

In the real data applications, we consider 10-fold cross-validated prediction errors as a measure of performance of the variable selection methods. For obtaining these cross-validated errors, we divide the data D into 10 folds $D_1, \cdots, D_{10}$. For each $D_k, \{k \in 1, 2, \cdots, 10\}$, we perform variable selection using the data from $D \setminus D_k$ to obtain predicted probabilities for the responses in D_k. The cross validation error for the fold k is defined as $CV_k = \sum_{i \in D_k} (\hat{\pi}_i - E_i)^2$, where $\hat{\pi}_i$ is the predicted probability for the ith observation. The overall CV error is $CV = \sum_{k=1}^{10} CV_k/n$, where $n = \sum_{k=1}^{10} |D_k|$.

Figure 3 shows the 10-fold cross validation errors for different methods given the number of covariates chosen. The X-axis represents different model sizes, and the Y-axis shows CV-errors for different methods considered. We note that Skinny Gibbs performs well along with MCP. In particular, the CV error is the smallest for Skinny Gibbs if we use smaller model sizes. In Figure 4, we plot the marginal posterior probabilities using the entire data D for two different Gibbs chains. We see that the top three genes from both the chains are the same and have higher inclusion probabilities than the rest. This is also consistent with Figure 3, which shows the largest decrease in CV error for the first three covariates. The Affymetrix IDs of the top genes in descending order of marginal posterior probabilities are 1432002-at, 1441569-at, and 1438936-s-at. The genes 1438936-s-at and 1438937-x-at (which is among the top five genes in both the Gibbs chains) belong to the Angiogenin gene family, which is previously found to be associated with obesity (Imai et al.

Figure 3: PCR Dataset: Cross Validated Prediction Error versus Model Size for several model selection methods

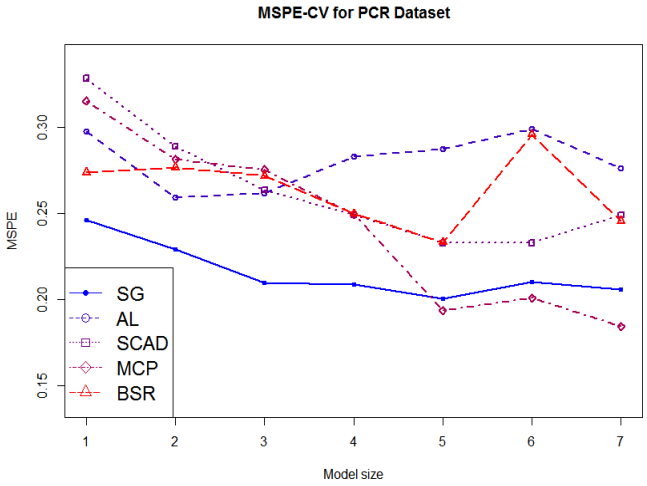
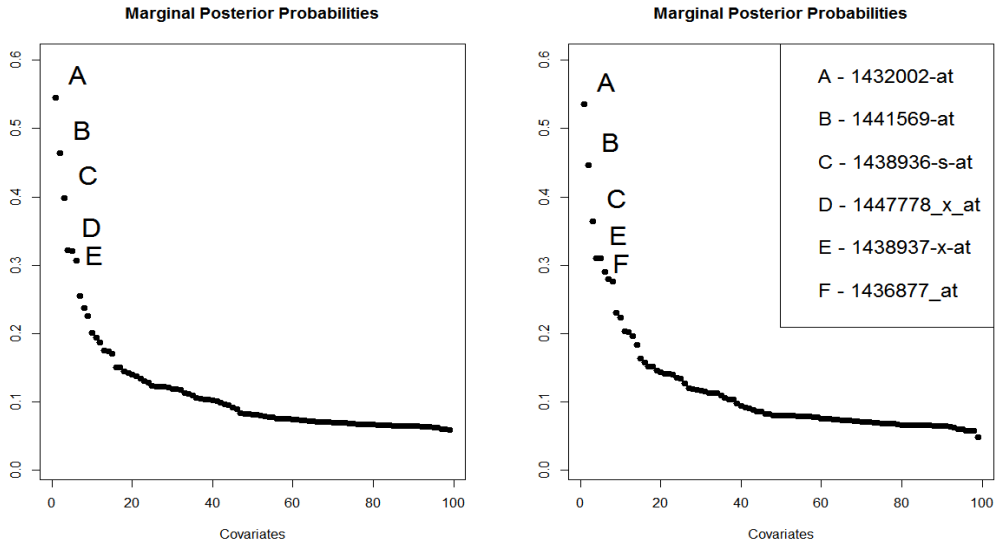


Figure 4: PCR Dataset: Marginal posterior probabilities from two different chains of Skinny Gibbs. The Affymetrix IDs of the top genes are given in the legend.



2008).

5.2 Lymph data

We now consider the gene expression data set considered in Hans et al. (2007), and Liang et al. (2013). The dataset contains gene expressions of $n = 148$ individuals. The response of interest is positive (high risk) or negative (low risk) status of the lymph node that is related to human breast cancer. There are 100 low risk cases and 48 high risk cases. After prescreening in Hans et al. (2007), a total of 4512 genes are selected showing a variation above the noise levels. In addition, there are two clinical variables, including the tumor size in centimeters as well as the protein assay-based estrogen receptor status (coded as binary). Hence we have $p = 4514$ candidate covariates with a sample size of $n = 148$. Due to the large p in this example, the results for Skinny Gibbs are based on combining ten different chains with the length and initialization described in Subsection 5.1.

As in Subsection 5.1, we present the 10-fold cross validated prediction errors for the methods considered, see Figure 5. We reported the errors for the models of size smaller than or equal to seven as the larger models often lead to complete separation when the model is fit to the estimation data leading to an unstable prediction for the testing data. All the methods considered have similar performance in terms of CV errors, with Skinny Gibbs having slightly lower errors. The CV errors from Skinny Gibbs suggest that the top six genes are important. Figure 6 shows the largest 100 posterior probabilities of $Z_j = 1$ from two different chains of Skinny Gibbs. The two chains lead to slightly different ordering of the top six genes, but there is only one non-overlapping gene in the two sets indicating the stability of the results. It is comforting to note that a few variables have substantially higher marginal probabilities than the rest in both the chains. However, some of the top variables do not have the marginal posterior probabilities close to 1, which

Figure 5: Lymph Dataset: Cross Validated Prediction Error versus Model Size for several model selection methods

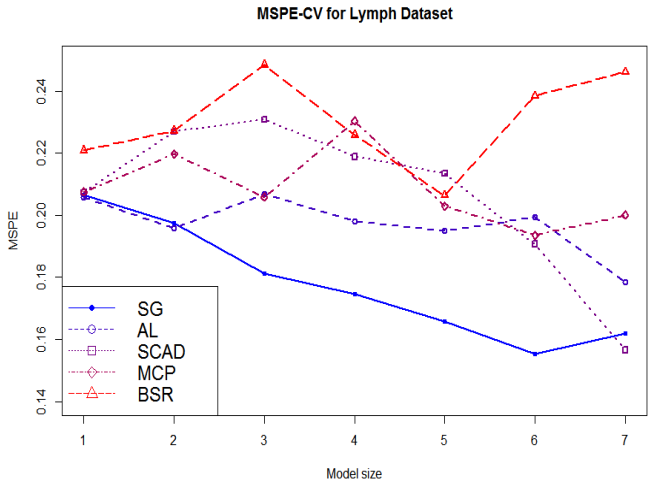
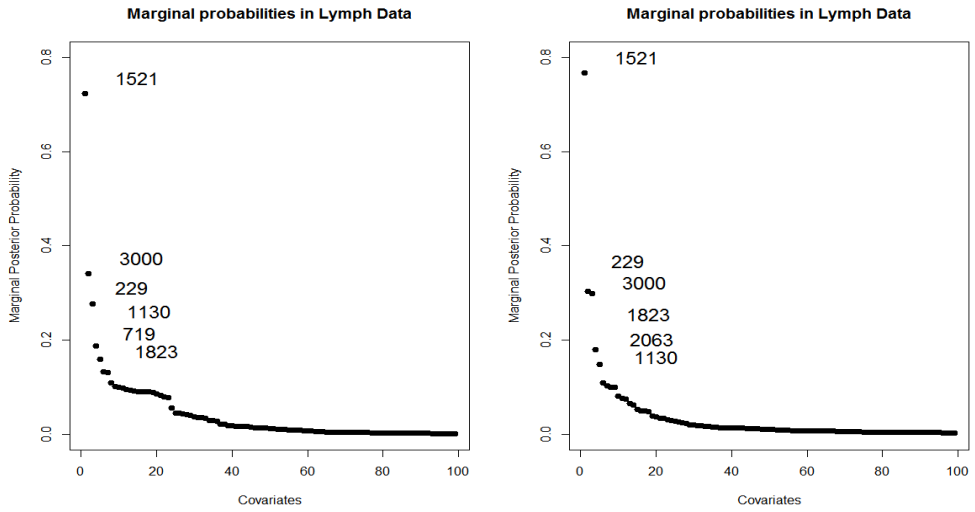


Figure 6: Lymph Dataset: Marginal Posterior Probabilities from two different chains of Skinny Gibbs. The labels on the top six points correspond to the column numbers of the genes.



can be attributed to the phenomenon that multiple sets of predictors in this problem can represent the model nearly equally well. In the supplementary materials, plots of the marginal posterior probabilities along the Skinny Gibbs iterates are provided.

6 Conclusion

In this paper, we propose a novel Gibbs sampler for variable selection in logistic regression. The proposed Skinny Gibbs has desired theoretical and computational properties. The strong selection consistency of the Gibbs sampler is established, which guarantees that the posterior probability of the true model goes to one. Computationally, each iteration of Skinny Gibbs requires complexity that is linear in p. Empirical results presented in the paper show the highly competitive performance of this approach for model selection. The theoretical and computational techniques developed in the paper can be extended to other models that have normal scale mixture representations. Skinny Gibbs can also be extended to the case where the prior distribution has a normal scale mixture representation such as the conjugate priors of Chen et al. (2008).

Supplementary materials

Part A: Proofs of Theorems 3.1 and 3.6.

Part B: A discussion about the connection between L_0 penalization and Skinny Gibbs.

Part C: A discussion about an unbiasedness property of Skinny Gibbs.

Part D: Skinny Gibbs algorithm using the Polya-Gamma scale mixture representation (Polson et al. 2013) of the logistic distribution.

Part E: A discussion on the stability and convergence of Skinny Gibbs chains for the empirical studies of Sections 4 and 5.

Part F: Additional simulation results for high correlation and weak signal settings that are not presented in the paper.

References

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal Predictive Model Selection," *Annals of Statistics*, 32, 870–897.

Bhattacharya, A., Chakraborty, A., and Mallick, B. (2016), "Fast sampling with Gaussian scale mixture priors in high-dimensional regression," *Biometrika*, 103, 985 – 991.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), "Dirichlet Laplace Priors for Optimal Shrinkage," *Journal of the American Statistical Association*, 110, 1479–1490.

Bondell, H. D. and Reich, B. J. (2012), "Consistent High Dimensional Bayesian Variable Selection via Penalized Credible Regions," *Journal of the American Statistical Association*, 107, 1610–1624.

Breheny, P. and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection," *Annals of Applied Statistics*, 5, 232–253.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer-Verlag Berlin Heidelberg.

Carbonetto, P. and Stephens, M. (2012), "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies," *Bayesian Analysis*, 7, 73–108.

Castillo, I. and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences," *Annals of Statistics*, 40, 2069–2101.

Chen, J. and Chen, Z. (2012), "Extended BIC for Small-n-large-P Sparse GLM," *Statistica Sinica*, 22, 555–574.

Chen, M. H., Huang, L., Ibrahim, J. G., and Kim, S. (2008), "Bayesian Variable Selection and Computation for Generalized Linear Models with Conjugate Priors," *Bayesian Analysis*, 3, 585–614.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Peng, H. (2004), "Nonconcave Penalized Likelihood with A Diverging Number of Parameters," *Annals of Statistics*, 32, 928–961.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2008), "Regularized Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33.

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008), "A Weakly Informative Default Prior Distribution for Logistic and other Regression Models," *Annals of Applied Statistics*, 4, 1360–1383.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford: Oxford University Press, chap. 5, pp. 609–620.

Guan, Y. and Stephens, M. (2011), "Bayesian Variable Selection Regression for Genome-wide Association Studies and Other Large-scale Problems," *Annals of Applied Statistics*, 5, 1780–1815.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for "Large p" Regression," *Journal of the American Statistical Association*, 102, 507–516.

Holmes, C. C. and Held, L. (2006), "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian analysis*, 1, 145–168.

Huang, J. and Zhang, C. H. (2012), "Estimation and Selection via Absolute Penalized Convex Minimization And its Multistage Adaptive Applications," *Journal of Machine Learning Research*, 13, 1839–1864.

Imai, Y., Patel, H. R., Doliba, N. M., Matschinsky, F. M., Tobias, J. W., and Ahima, R. S. (2008), "Analysis of Gene Expression in Pancreatic Islets from Diet-induced Obese Mice," *Physiol Genomics*, 36, 43–51.

Ishwaran, H. and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33, 730–773.

Johnson, V. E. and Rossell, D. (2012), "Bayesian Model Selection in High-dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski,

K., and Attie, A. D. (2006), "Combined Expression Trait Correlations and Expression Quantitative Trait Locus Mapping," *PLoS Genetics*, 2, e6.

Liang, F., Liu, C., and Carroll, R. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320.

Liang, F., Song, Q., and Yu, K. (2013), "Bayesian Subset Modeling for High Dimensional Generalized Linear Models," *Journal of the American Statistical Association*, 108, 589–606.

Narisetty, N. N. and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *Annals of Statistics*, 42, 789–817.

O'Brien, S. M. and Dunson, D. B. (2004), "Bayesian Multivariate Logistic Regression," *Biometrics*, 60, 739–746.

Park, T. and Casella, G. (2008), "The Bayesian LASSO," *Journal of the American Statistical Association*, 103, 681–686.

Polson, N., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models using Pólya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Ročková, V. and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846.

Scott, G. and Berger, J. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *Annals of Statistics*, 38, 2587–2619.

Shen, X., Pan, W., and Zhu, Y. (2012), "Likelihood-based Selection and Sharp Parameter Estimation," *Journal of the American Statistical Association*, 107, 223–232.

Stefanski, L. A. (1991), "A Normal Scale Mixture Representation of the Logistic Distribution," *Statistics and Probability Letters*, 69–70.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

van de Geer S. A. (2008), "High-dimensional Generalized Linear Models and the Lasso," *Annals of Statistics*, 36, 614–645.

Wendel, A. A., Li, L. O., Y., L., Cline, G. W., Shulman, G. I., and Coleman, R. A. (2010), "Glycerol-3-phosphate Acyltransferase 1 Deficiency in ob/ob Mice Diminishes Hepatic Steatosis but Does Not Protect against Insulin Resistance or Obesity." *Diabetes*, 59, 1321–1329.

Zhang, C. H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty." *The Annals of Statistics*, 38, 894–942.

Zhao, P. and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Efficient Empirical Bayes Variable Selection and Estimation in Linear Models*, 7, 2541–2563.