Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Bayesian Subset Modeling for High-Dimensional Generalized Linear Models

Faming Liang , Qifan Song & Kai Yu

To cite this article: Faming Liang , Qifan Song & Kai Yu (2013) Bayesian Subset Modeling for High-Dimensional Generalized Linear Models, Journal of the American Statistical Association, 108:502, 589-606, DOI: 10.1080/01621459.2012.761942

To link to this article: https://doi.org/10.1080/01621459.2012.761942

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Bayesian Subset Modeling for High-Dimensional Generalized Linear Models

Faming LIANG, Qifan SONG, and Kai YU

This article presents a new prior setting for high-dimensional generalized linear models, which leads to a Bayesian subset regression (BSR) with the maximum a posteriori model approximately equivalent to the minimum extended Bayesian information criterion model. The consistency of the resulting posterior is established under mild conditions. Further, a variable screening procedure is proposed based on the marginal inclusion probability, which shares the same properties of sure screening and consistency with the existing sure independence screening (SIS) and iterative sure independence screening (ISIS) procedures. However, since the proposed procedure makes use of joint information from all predictors, it generally outperforms SIS and ISIS in real applications. This article also makes extensive comparisons of BSR with the popular penalized likelihood methods, including Lasso, elastic net, SIS, and ISIS. The numerical results indicate that BSR can generally outperform the penalized likelihood methods. The models selected by BSR tend to be sparser and, more importantly, of higher prediction ability. In addition, the performance of the penalized likelihood methods tends to deteriorate as the number of predictors increases, while this is not significant for BSR. Supplementary materials for this article are available online.

KEY WORDS: Bayesian classification; Posterior consistency; Stochastic approximation Monte Carlo; Sure variable screening; Variable selection.

1. INTRODUCTION

Let $D^n = \{(\boldsymbol{y}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{P_n}) : \boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{x}_i \in \mathbb{R}^n,\ i = 1, 2, \ldots, P_n\}$ denote a dataset of n observations each consisting of P_n predictors (also known as explanatory variables or features in this article) and a response, where P_n can increase with the sample size n. By showing the subject index, we can also rewrite the dataset as $D^n = (y^{(i)}; x_1^{(i)}, \ldots, x_{P_n}^{(i)})_{i=1}^n$. Let $(y, x_1, \ldots, x_{P_n})$ denote a generic observation of the dataset. Suppose that the data can be modeled by a generalized linear model (GLM) with the density function given by

$$f(y|\theta) = \exp\{a(\theta)y + b(\theta) + c(y)\}, \tag{1}$$

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

$$\theta = \beta_1 x_1 + \cdots + \beta_{P_n} x_{P_n}, \tag{2}$$

where $\beta_1, \ldots, \beta_{P_n}$ are regression coefficients. Here, the intercept term has been treated as a special predictor included in the set $\{x_1, \ldots, x_{P_n}\}$. The mean function $\mu = E(y|x_1, \ldots, x_{P_n}) = -b'(\theta)/a'(\theta) \equiv \varphi(\theta)$, where $\varphi(\cdot)$ is the inverse of a chosen link function. This class of GLMs includes regression models with responses that are binary, Poisson, and Gaussian (with known variance), whose link functions are logit, log, and linear, respectively. A dispersion parameter can also be included in (1), which can then include the Gaussian model with unknown variance.

The GLMs have been used in diverse fields, ranging from biomedical sciences to economics. In these studies, a problem of particular interest is variable selection, which is to identify a subset of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{P_n}\}$ that forms causal features of $\boldsymbol{y}$. Examples include identification of genes or single nucleotide polymorphisms (SNPs) that are responsible for certain diseases, identification of stocks that generate profits in investment portfolios, etc. Under the situation of small-n-large-P, variable selection can pose a great challenge for existing statistical methods.

The problem of variable selection for high-dimensional GLMs has been treated with both penalized likelihood and Bayesian approaches in the literature. The penalized likelihood approach is to find a model, that is, a subset of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{P_n}\}$, to minimize a penalized likelihood function of the form

$$-\sum_{i=1}^{n} \log f(y_i|\theta) + p_\lambda(\boldsymbol{\beta}), \tag{3}$$

where $p_\lambda(\cdot)$ is the penalty function, λ is a tunable scale parameter, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{P_n})'$ is the vector of regression coefficients.

Different choices of $p_\lambda(\cdot)$ lead to different methods. For example, the L_1 penalty norm leads to the Lasso method (Tibshirani 1996), a linear combination of the L_1 and L_2 penalty norms leads to the elastic net method (Zou and Hastie 2005), and particular choices of concave penalty norms lead to the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) and minimax concave penalty (MCP; Zhang 2009) methods. A nice property shared by these penalty norms is that they are singular at zero, and thus, can induce sparsity of the model by increasing the value of λ. In practice, the parameter λ can be determined through a cross-validation procedure. Refer to Fan and Lv (2010) for an overview of the penalized likelihood approach and the related sure independence screening (SIS;

Faming Liang is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *fliang@stat.tamu.edu*). Qifan Song is Graduate Student, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *qsong@stat.tamu.edu*). Kai Yu is Investigator, Division of Cancer Epidemiology & Genetics, National Cancer Institute, Rockville, MD 20892-7335 (E-mail: *yuka@mail.nih.gov*). Liang's research was partially supported by grants from the National Science Foundation (DMS-1007457 and DMS-1106494) and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST). The authors thank Dr. Chris Hans for sending us the lymph data, and thank the editor, associate editor, and two referees for their constructive comments that have led to significant improvement of this article.

Fan and Lv 2008; Fan and Song 2010) and iterative sure independence screening (ISIS) methods (Fan, Samworth, and Wu 2009). SIS first ranks predictors according to their marginal utility (that is, each predictor is used independently as a predictor to decide its usefulness for predicting the response), then selects a subset of predictors with the marginal utility exceeding a predefined threshold, and finally refines the model using Lasso or SCAD. Fan and Song (2010) suggested a way to measure the marginal utility for GLMs using marginal regression coefficients or marginal likelihood ratios (with respect to the null model). ISIS is an iterative version of SIS. It considers information from those predictors already chosen in the previous SIS steps when measuring the marginal utility of remaining predictors, and it has, in general, an improved performance over SIS.

Another class of penalty functions stems from information theory, including the famous Akaike information criterion (AIC) and Bayesian information criterion (BIC) and their recent extension EBIC (Chen and Chen 2008, 2012). It is known that, in the situation of small-n-large-P, AIC and BIC are overly liberal and are ineffective for variable selection (e.g., Broman and Speed 2002). To overcome this obstacle, Chen and Chen (2012) advocated the use of EBIC, which takes an L_0-penalty of the form

$$p_\lambda(\boldsymbol{\beta}) = \frac{|\boldsymbol{\beta}|}{2} \log n + \gamma |\boldsymbol{\beta}| \log P_n, \tag{4}$$

where $|\boldsymbol{\beta}|$ denotes the number of nonzero elements of $\boldsymbol{\beta}$ (i.e., the model size), and $\gamma > 0$ is a tunable parameter. Variable selection under the L_0-penalty is called the subset regression, which does not shrink the regression coefficients toward 0 in contrast to the L_a-penalty (with $a > 0$). Under regularity conditions, Chen and Chen (2012) showed that EBIC is consistent if $P_n = O(n^\kappa)$ and $\gamma > 1 - \frac{1}{2\kappa}$, where the consistency means that the minimum EBIC model will converge to the true model in probability as the sample size $n \to \infty$. However, how to choose γ for a given dataset is unclear. An overly large value of γ will certainly lead to the selection of a subset of the causal features. On the other hand, an overly small value of γ may lead to the selection of a large model including some irrelevant features.

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. Motivated by this observation, several authors subsequently proposed the Bayesian Lasso approach using Laplace-like priors, for example, Figueiredo (2003), Bae and Mallick (2004), and Park and Casella (2008). On the other hand, the traditional normal–inverse gamma prior seems also to work well for this problem, for example, Hans, Dobra, and West (2007), Bottolo and Richardson (2010), and Richardson, Bottolo, and Rosenthal (2010). To deal with the explosive sample space due to the increase in dimension, advanced Monte Carlo algorithms have been developed. For example, Hans, Dobra, and West (2007) proposed a shotgun stochastic search (SSS) algorithm aimed at finding the MAP model. A fully Bayesian treatment for this problem can be found in Bottolo and Richardson (2010), where the authors propose an adaptive Markov chain Monte Carlo (MCMC) algorithm to accelerate the simulation. From the perspective of Bayesian theory, a remarkable result was obtained by Jiang (2006, 2007), who showed that the true density (1) can be estimated consistently, based on the models sampled from the posterior distribution, through a Bayesian variable selection procedure.

Prior to the development of small-n-large-P regressions, a large number of Bayesian variable selection methods had been proposed for large-n-small-P regressions. A work of particularly interest to us is Liang, Truong, and Wong (2001), where the authors establish an explicit relationship between the Bayesian approach and the penalized likelihood approach for linear regression; they show empirically that Bayesian subset regression (BSR), that is, choosing priors such that the resulting negative log-posterior probability of the subset model can be approximately reduced to a frequentist subset model selection statistic (up to a multiplicative constant), for example, Mallows' C_p (Mallows 1973), AIC or BIC, when the sample size becomes large, often outperforms conventional Bayesian regression in prediction. Since the regression parameters are often estimated by their maximum likelihood estimation (MLE) in the frequentist subset model selection statistics, BSR encourages people to sample from the profile posterior (Hsu 1995) of subset models but with the regression coefficients being replaced by their MLE. This is attractive for GLMs, for which sampling regression coefficients from the full posterior are often quite complicated. Sampling from the profile posterior also allows us to table-list some models together with their energy (i.e., the negative log-posterior probability) and use the table to avoid computation for repeatedly sampled models. As discussed in Section 7, the CPU saving by the table-listing approach can be substantial for some problems.

Our contribution in this article is three-fold:

- Motivated by Liang, Truong, and Wong (2001), we propose a new prior setting for high-dimensional GLMs, which leads to a BSR with the negative log-posterior distribution approximately reduced to the EBIC when the sample size is large. Under mild conditions, we establish the consistency of the posterior; that is, the true density (1) can be estimated consistently by the density of the models sampled from the posterior as $n \to \infty$. In addition, we show that the posterior probability of the true model will converge to 1 as $n \to \infty$, and this directly leads to the consistency of the MAP model; that is, the MAP model will converge to the true model in probability as $n \to \infty$.
- We propose a variable screening procedure based on the marginal inclusion probability of predictors. We show that it has the same property of sure screening as SIS. However, since the marginal inclusion probability has incorporated the joint information of all predictors, the new procedure can generally outperform SIS and ISIS as shown by our numerical examples. As previously mentioned, SIS uses only the marginal information from each predictor, and ISIS tries to incorporate information from other predictors, but in an indirect way.
- We conduct extensive numerical comparisons of BSR with the penalized likelihood methods, including Lasso, elastic net, SIS, and ISIS. The numerical results show that BSR significantly and consistently outperforms the penalized likelihood methods. In particular, through a subset versus full data comparison study (Section 4), we find that the penalized likelihood methods can scale badly,

whose performance deteriorates as the dimension P_n increases. Due to their forward variable selection nature, Lasso and elastic net are sensitive to noise predictors and tend to be adversely affected by the dimension P_n. Since SIS and ISIS use only the marginal information or limited joint information of all predictors, they are also sensitive to noise predictors and tend to be adversely affected by the dimension P_n. However, BSR conducts a global search over the entire model space and makes use of joint information from all predictors; therefore, it is insensitive to noise predictors and less affected by increasing P_n. This suggests that BSR is more suitable for high-dimensional variable selection than the penalized likelihood methods, although the latter are computationally more attractive.

The remainder of this article is organized as follows. Section 2 describes the new prior setting and establishes the consistency of the resulting posterior. Section 3 describes a variable selection procedure for high-dimensional GLMs and proves its properties of sure screening and consistency. Based on one numerical example, Section 4 illustrates why BSR is preferable to the penalized likelihood methods. Sections 5 and 6 compare BSR with the penalized likelihood methods on a simulation study and some gene expression data, respectively. Section 7 evaluates the computational time of BSR. Section 8 concludes the article with a brief discussion.

2. BAYESIAN SUBSET MODELING

2.1 The Prior and Posterior

Consider the GLM specified by Equations (1) and (2). We assume the following condition holds:

(A$_1$) $P_n \succ n^\delta$ for some positive constant δ,

where $b_n \succ a_n$ means $\lim_{n\to\infty} a_n/b_n = 0$. Let ξ_n denote a subset model of the GLM, and let $|\xi_n|$ denote the size of ξ_n. To emphasize that the model size can increase with n, the subscript n is placed on the subset model.

To conduct Bayesian analysis for the GLM, we consider the following priors. Let $\{\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_{|\xi_n|}\}(\subset \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{P_n}\})$ denote the predictors included in ξ_n, and let $\boldsymbol{\beta}_{\xi_n} = (\beta^*_1, \ldots, \beta^*_{|\xi_n|})$ denote the vector of true regression coefficients of ξ_n. Assume that $\boldsymbol{\beta}_{\xi_n}$ is subject to the Gaussian prior:

$$\Pi(\boldsymbol{\beta}_{\xi_n}) = \frac{1}{(2\pi\sigma^2_{\xi_n})^{|\xi_n|/2}} \exp\left\{-\frac{1}{2\sigma^2_{\xi_n}} \sum_{i=1}^{|\xi_n|} \beta^{*2}_i\right\}, \tag{5}$$

where $\sigma^2_{\xi_n}$ denotes a prespecified variance and works as a hyperparameter for this prior. Taking advantage of the flexibility of prior distributions, we choose $\sigma^2_{\xi_n}$ for each model such that

$$\log \Pi(\boldsymbol{\beta}_{\xi_n}) = O(1), \tag{6}$$

which implies that the prior information of $\boldsymbol{\beta}_{\xi_n}$ can be ignored for sufficiently large n. If we assume that the true model is sparse and satisfies the condition

(A$_1$) $\lim_{n\to\infty} \sum_{i=1}^{P_n} |\beta_j| < \infty$,

and if we set

$$\sigma^2_{\xi_n} = \frac{1}{2\pi} e^{C_0/|\xi_n|}, \quad \text{for some positive constant } C_0, \tag{7}$$

then substituting (7) into (5) results in that

$$\begin{aligned} |\log(\Pi(\boldsymbol{\beta}_{\xi_n}))| &= \left|-\frac{C_0}{2} - \pi e^{-C_0/|\xi_n|} \sum_{i=1}^{|\xi_n|} \beta^{*2}_i\right| \\ &\le \frac{C_0}{2} + \pi \sum_{i=1}^{P_n} \beta^2_i < \infty, \end{aligned}$$

for any model ξ_n, which implies (6) holds. Since a larger value of C_0 leads to a flatter prior distribution of $\boldsymbol{\beta}_{\xi_n}$, a large value is generally preferred for the choice of C_0. However, as shown below, this article suggests to sample from the approximate posterior distribution (14) instead of the exact posterior distribution, which gets rid of the issue of choosing the value of C_0.

Further, we let the model ξ_n be subject to the prior

$$\Pi(\xi_n) = \nu_n^{|\xi_n|}(1-\nu_n)^{P_n - |\xi_n|}, \tag{8}$$

that is, each variable has a prior probability ν_n, independent of the other variables, to be selected for the subset model. Following the derivation of BIC (e.g., Ando 2010), we have

$$\begin{aligned} &\log f(\boldsymbol{y}_n|\boldsymbol{\beta}_{\xi_n}, \xi_n, \boldsymbol{X}_n) \\ &\quad\approx \log f(\boldsymbol{y}_n|\widehat{\boldsymbol{\beta}}_{\xi_n}, \boldsymbol{X}_n) - \frac{n}{2}(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})' J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n}), \end{aligned}$$

where $\boldsymbol{y}_n = (y_1, \ldots, y_n)$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{P_n})$, $\widehat{\boldsymbol{\beta}}_{\xi_n}$ denotes the MLE of $\boldsymbol{\beta}_{\xi_n}$, and

$$J_n(\widehat{\boldsymbol{\beta}}_{\xi_n}) = -\frac{1}{n} \frac{\partial^2 \log f(\boldsymbol{y}_n|\boldsymbol{\beta}_{\xi_n}, \xi_n, \boldsymbol{X}_n)}{\partial\boldsymbol{\beta}_{\xi_n}\partial\boldsymbol{\beta}'_{\xi_n}}\bigg|_{\boldsymbol{\beta}_{\xi_n} = \widehat{\boldsymbol{\beta}}_{\xi_n}}.$$

Similarly, for the prior we have

$$\Pi(\boldsymbol{\beta}_{\xi_n}) \approx \Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) + (\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})' \frac{\partial \Pi(\boldsymbol{\beta}_{\xi_n})}{\partial \boldsymbol{\beta}_{\xi_n}}\bigg|_{\boldsymbol{\beta}_{\xi_n} = \widehat{\boldsymbol{\beta}}_{\xi_n}}.$$

Therefore,

$$\begin{aligned} &f(\boldsymbol{y}_n|\xi_n, \boldsymbol{X}_n) \\ &\quad= \int f(\boldsymbol{y}_n|\boldsymbol{\beta}_{\xi_n}, \xi_n, \boldsymbol{X}_n)\Pi(\boldsymbol{\beta}_{\xi_n})d\boldsymbol{\beta}_{\xi_n} \\ &\quad\approx \int \exp\left\{\log f(\boldsymbol{y}_n|\widehat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n)\right. \\ &\qquad\left. - \frac{n}{2}(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})^T J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})\right\} \\ &\qquad\times \left\{\Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) + (\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})' \frac{\partial \Pi(\boldsymbol{\beta}_{\xi_n})}{\partial \boldsymbol{\beta}_{\xi_n}}\bigg|_{\boldsymbol{\beta}_{\xi_n} = \widehat{\boldsymbol{\beta}}_{\xi_n}}\right\} d\boldsymbol{\beta}_{\xi_n} \\ &\quad= f(\boldsymbol{y}_n|\widehat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n)\Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) \\ &\qquad\times \int \exp\left\{-\frac{n}{2}(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})^T J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})(\boldsymbol{\beta}_{\xi_n} - \widehat{\boldsymbol{\beta}}_{\xi_n})\right\} d\boldsymbol{\beta}_{\xi_n} \\ &\quad= f(\boldsymbol{y}_n|\widehat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n)\Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) \frac{(2\pi)^{|\xi_n|/2}}{n^{|\xi_n|/2}|J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})|^{1/2}}, \end{aligned}$$

which implies that

$$\begin{aligned} \log \Pi(\xi_n|D^n) &= C + \log\{\Pi(\xi_n) f(\boldsymbol{y}_n|\xi_n, \boldsymbol{X}_n)\} \\ &\approx C + \log \Pi(\xi_n) + \log f(\boldsymbol{y}_n|\widehat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n) \\ &\quad + \log \Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) - \frac{|\xi_n|}{2}\log(n) + \frac{|\xi_n|}{2}\log(2\pi) \\ &\quad - \frac{1}{2}\log|J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})|, \end{aligned} \tag{9}$$

where C denotes a constant, and $J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})$ converges to the Fisher information matrix for a single observation and thus $\log|J_n(\widehat{\boldsymbol{\beta}}_{\xi_n})|$ is of order $O(1)$. Note that the approximation of $f(\boldsymbol{y}_n|\xi_n, \boldsymbol{X}_n)$ is identical to the Laplace approximation except where $\tilde{\boldsymbol{\beta}}_{\xi_n}$, the mode of the posterior, is used, while $\tilde{\boldsymbol{\beta}}_{\xi_n}$ and $\widehat{\boldsymbol{\beta}}_{\xi_n}$ are approximately equivalent under the prior distributions (5) and (7). The same approximation of $f(\boldsymbol{y}_n|\xi_n, \boldsymbol{X}_n)$ can also be obtained by appealing to the basic marginal likelihood identity of Chib (1995). If the following condition holds:

(A$_1$) $\|\boldsymbol{x}_i\|/\|\boldsymbol{x}_j\| = O(1)$ for all $i, j = 1, 2, \ldots, P_n$,

where $\|\cdot\|$ denotes the L_2-norm of a vector, then $\log \Pi(\widehat{\boldsymbol{\beta}}_{\xi_n}) = O(1)$ by the condition (A$_2$) and the property of MLE. Note, under condition (A$_3$), the latter implies $\|\widehat{\boldsymbol{\beta}}_{\xi_n}\| = O(\|\boldsymbol{\beta}\|)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{P_n})'$. If we ignore the term of order $O(1)$ and higher order terms in (9), and further if we let the prior probability ν_n in (8) take a value of the form

$$\nu_n = \frac{1}{1 + P_n^{\gamma}\sqrt{2\pi}}, \tag{10}$$

for some parameter γ, then we have the approximated log-posterior,

$$\log \Pi(\xi_n|D^n) \approx C + \log f\big(\boldsymbol{y}_n\big|\hat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n\big) - \frac{|\xi_n|}{2}\log(n) - |\xi_n|\gamma\log(P_n). \tag{11}$$

From the above derivation, it is clear that the accuracy of the overall approximation of the posterior is of order $O(1/n)$; that is,

$$\Pi(\xi_n|D^n) = C' f\big(\boldsymbol{y}_n\big|\hat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n\big)\big(\sqrt{n}P_n^{\gamma}\big)^{-|\xi_n|} \times \left\{1 + O\left(\frac{1}{n}\right)\right\}, \tag{12}$$

for some constant C'.

Maximizing this posterior probability is approximately equivalent to minimizing the EBIC given by

$$\text{EBIC} = -2\log f\big(\boldsymbol{y}_n\big|\hat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n\big) + |\xi_n|\log(n) + 2|\xi_n|\gamma\log(P_n),$$

which has been shown to be a consistent criterion for model selection under appropriate conditions, such as $P_n = O(n^{\kappa})$ for some $\kappa > 0, \gamma > 1 - \frac{1}{2\kappa}$ and some regularity conditions for the design matrix. The consistency means that for any $\gamma > 1 - \frac{1}{2\kappa}$, the minimum EBIC model will converge to the true model as the sample size $n \to \infty$.

To facilitate calculation of the MLE $\hat{\boldsymbol{\beta}}_{\xi_n}$, one often needs to put an upper bound on the model size $|\xi_n|$, for example, $|\xi_n| < K_n$, where K_n may depend on the sample size n. It is known that if $|\xi_n| > n$, the matrix $\tilde{\boldsymbol{X}}'_{|\xi_n|}\tilde{\boldsymbol{X}}_{|\xi_n|}$ would become singular, where $\tilde{\boldsymbol{X}}_{|\xi_n|} = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_{|\xi_n|})$ is the design matrix of model ξ_n. With this bound, the prior for the model ξ_n becomes

$$\Pi(\xi_n) \propto \nu_n^{|\xi_n|}(1 - \nu_n)^{P_n - |\xi_n|} I[|\xi_n| < K_n], \tag{13}$$

and the log-posterior distribution becomes

$$\log \Pi(\xi_n|D^n) \approx \begin{cases} C + \log f\big(\boldsymbol{y}_n\big|\hat{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n\big) & \text{if } |\xi_n| < K_n, \\ -\dfrac{|\xi_n|}{2}\log(n) - |\xi_n|\gamma\log(P_n), & \\ -\infty, & \text{otherwise.} \end{cases} \tag{14}$$

Since the negative of this log-posterior distribution is approximately reduced to the EBIC (up to a multiplicative constant), it is called a BSR.

To justify the BSR, we refer to the Bayesian theory developed by Jiang (2007). Rewrite the dataset as $D^n = (y^{(i)}, x^{(i)})_{i=1}^n$, where $x^{(i)} = (x_1^{(i)}, \ldots, x_{P_n}^{(i)})$. Let $f = f(y|x)$ denote the true density of y conditioned on x, and let $\hat{f} = \hat{f}(y|x, \xi_n)$ denote the conditional density proposed by the posterior $\Pi(\xi_n|D^n)$. Let $v_x(dx)$ denote the probability measure for x, and $v_y(dy)$ the dominating measure for the conditional densities f and $\hat{f}$. Assume that the data for n subjects are independent and identically distributed based on $f(y, x)v_x(dx)v_y(dy)$. Let

$$d(\hat{f}, f) = \sqrt{\int\int\left(\sqrt{\hat{f}} - \sqrt{f}\right)^2 v_y(dy)v_x(dx)}$$

denote the Hellinger distance between $\hat{f}$ and f. Following Theorem 2 of Jiang (2007), we have the following lemma:

Lemma 2.1. (Posterior consistency) Consider the GLM specified by (1) and (2), which satisfies the conditions (A$_1$)–(A$_3$) and $|x_j| \le 1$ for all $j = 1, \ldots, P_n$. Let ϵ_n be a sequence such that $\epsilon_n \in (0, 1]$ for each n and $n\epsilon_n^2 \succ \log(P_n)$. Suppose that the priors for the GLM are specified by (5) and (13) with the hyperparameter ν_n chosen in (10), and that γ and K_n are chosen such that the following conditions hold:

(B$_1$) $P_n \le e^{C_1 n^{\alpha}}$ for some constants $C_1 > 0$ and $\alpha \in (0, 1)$ for all large enough n;

(B$_2$) $\Delta(r_n) = \inf_{\xi_n:|\xi_n|=r_n}\sum_{j:j\notin\xi_n}|\beta_j| \le e^{-C_2 r_n}$ for some constant $C_2 > 0$, where $r_n = \lceil P_n\nu_n\rceil$ denotes the upper-rounded expectation of the prior (8);

(B$_3$) $C_2^{-1}\log(n) \le r_n \le K_n \prec n^{\beta}$ for some $\beta \in (0, q)$, where $q = \min(1 - \alpha, \delta)$ for logistic, probit, and normal linear regressions, and $q = \min\{1 - \alpha, \delta, \alpha/4\}$ for Poisson regression and exponential regressions with log-linear link functions.

Then for some $c > 0$ and for all sufficiently large n,

$$P\big\{\Pi[d(\hat{f}, f) > \epsilon_n|D^n] \ge e^{-0.5cn\epsilon_n^2}\big\} \le e^{-0.5cn\epsilon_n^2}, \tag{15}$$

where $P\{\cdot\}$ denotes the probability measure for the data D^n.

This lemma can be viewed as a corollary of Theorem 2 of Jiang (2007) for our particular choice of priors. The proof is straightforward as the consequence of the following fact: as shown in (7), the eigenvalues of the covariance matrix of the prior (5) tend to a constant as $n \to \infty$, and thus $\max\{\sigma_{\xi_n}^2, \sigma_{\xi_n}^{-2}\}$ can be bounded by BK_n for some constant $B > 0$.

If we further assume that $P_n = O(n^{\kappa})$ for some $\kappa > 0$, then $r_n = P_n/[1 + P_n^{\gamma}\sqrt{2\pi}] = O(n^{\kappa(1-\gamma)})$. Let $K_n = C_3 r_n$ for some constant $C_3 > 1$. Then α can be set to a number close to zero. If $\kappa > 1$, we can set $\delta = 1$ and thus any $\gamma \in (1 - 1/\kappa, 1)$ ensures the conditions (B$_1$)–(B$_3$) to be satisfied for logistic, probit, and normal linear regressions. If $\kappa \le 1$, we can set $\delta = \eta\kappa$ for any $0 < \eta < 1$ such that $\eta\kappa < 1 - \alpha$, then the choice of any $\gamma \in (1 - \eta, 1)$ ensures the conditions (B$_1$)–(B$_3$) to be satisfied for logistic, probit, and normal linear regressions. Pushing α to its limit 0 and η to its limit 1, we have the range $\gamma \in (0, 1)$ for $\kappa \le 1$. Given the choice of γ, the convergence rate can be

taken as

$$\epsilon_n \sim n^{-(1-\alpha-\zeta)/2},$$

for some $\zeta \in (\kappa(1-\gamma), q)$ with q as defined in (B_3). For Poisson and exponential regressions, the range of γ and the sequence of $\{\epsilon_n\}$ can be derived similarly.

On the relationship between the BSR posterior and other posterior or model selection criteria used in the literature, we have the following remarks:

- The posterior (9) is rather general. For different choices of ν_n, maximizing (9) can lead to different model selection criteria. For example, the choice $\nu_n = 1/(1+\sqrt{2\pi})$ leads to the BIC criterion.
- Hans, Dobra, and West (2007) placed a constant-variance-independent Gaussian prior on $\boldsymbol{\beta}_{\xi_n}$, that is, $\boldsymbol{\beta}_{\xi_n} \sim N(0, \tau I_{|\xi_n|})$, with τ being a prior hyperparameter to be specified by the user. Under this setting, $\log(\Pi(\boldsymbol{\beta}_{\xi_n})) = O(n)$, so the prior information cannot be ignored even when $n \to \infty$. Using the Laplace approximation, Hans, Dobra, and West (2007) obtained the approximate posterior

$$\begin{aligned}\log \Pi(\xi_n|D^n) \approx{}& \log \Pi(\xi_n) + \log f\big(\boldsymbol{y}_n|\widetilde{\boldsymbol{\beta}}_{\xi_n}, \xi_n, \boldsymbol{X}_n\big) \\ &+ \log \Pi\big(\widetilde{\boldsymbol{\beta}}_{\xi_n}\big) + \frac{|\xi_n|}{2}\log(2\pi) \\ &- \frac{1}{2}\log|\widehat{\Sigma}_n|, \end{aligned} \tag{16}$$

  where $\widetilde{\boldsymbol{\beta}}_{\xi_n}$ denotes the mode of the posterior $\Pi(\boldsymbol{\beta}_{\xi_n}|\xi_n, D^n)$, and $\widehat{\Sigma}_n$ is the negative Hessian matrix of $\log[f(\boldsymbol{y}_n|\boldsymbol{\beta}_{\xi_n}, \xi_n, \boldsymbol{X}_n)\Pi(\boldsymbol{\beta}_{\xi_n}|\xi_n)]$. Other priors, such as g, generalized-g, or hyper-g priors (Bové and Held 2011), can also be set for $\boldsymbol{\beta}_{\xi_n}$. Under these prior settings, the posterior consistency can still hold but with modified conditions of Lemma 2.1. An extra condition that is possibly necessary for the posterior consistency is on the growth rates of the largest and smallest eigenvalues of the prior covariance matrix with the model size, which, however, are usually difficult to obtain for g, generalized-g, and hyper-g priors.
- For a fully Bayesian variable selection approach, it is common to place a prior on ν_n. Scott and Berger (2010) showed that placing a prior on ν_n results in an increasing penalty for adding an extra predictor as P_n increases; that is, the fully Bayesian approach provides an automatic correction for the multiplicity involved in variable selection. It is easy to see that our choice of ν_n, given in (10), controls for the multiplicity in a straightforward way: if P_n grows large, then $\nu_n \to 0$. As discussed later, the multiplicity control is critical for the consistency of variable selection under our method. This also explains why the BIC is overly liberal for variable selection under the situation of small-n-large-P, for which $\nu_n = 1/(1+\sqrt{2\pi})$ fails to control the multiplicity.

2.2 Determination of the Prior Hyperparameter γ

The posterior (14) contains two prior hyperparameters γ and K_n, and γ plays an important role in BSR. Given the value of γ, K_n can be set to a large number such that the posterior probability $\Pi(|\xi_n| \geq K_n|D^n)$ is zero or nearly zero. Lemma 2.1 suggests a range for γ. But the range depends on an unknown parameter κ, which is not estimable for any given dataset. To overcome this obstacle, we suggest the following rule for determining the value of γ:

$$\gamma = \inf\left\{\tilde{\gamma} : \arg\max_{|\xi_n|} \Pi(|\xi_n||D^n, \tilde{\gamma}) = |\xi_{\text{map},\tilde{\gamma}}|\right\}, \tag{17}$$

which is to find the minimum value of γ for which the mode of $\Pi(|\xi_n||D^n, \gamma)$ is equal to the size of the MAP model. Here, to indicate the dependence of the posterior distribution of ξ_n on γ, we rewrite the posterior of $|\xi_n|$ as $\Pi(|\xi_n||D^n, \gamma)$ and denote the MAP model by $\xi_{\text{map},\gamma}$. If the resulting value of γ is ≥ 1, we may truncate it to a number that is slightly smaller than 1, say, 0.99.

The rationale for the rule (17) follows from Theorem 3.1 (given later), which implies that as $n \to \infty$, the posterior probability of the true model will converge to 1 and the MAP model will converge to the true model. Hence, the posterior mode of $|\xi_n|$ should be approximately equal to the size of the MAP model when n is large. In practice, if γ is overly large, then the equality $\arg\max_{|\xi_n|} \Pi(|\xi_n||D^n, \tilde{\gamma}) = |\xi_{\text{map},\tilde{\gamma}}|$ always hold, since, in this case, there will be only a single-size model sampled from the posterior and the resulting MAP model will probably be a subset of the true model. On the other hand, if γ is too small, then it will always be true that $\arg\max_{|\xi_n|} \Pi(|\xi_n||D^n, \tilde{\gamma}) > |\xi_{\text{map},\tilde{\gamma}}|$. To balance between these two extremes, we suggest the rule (17). For a given dataset, this rule makes the resulting posterior comply with its limiting theory, while including as many predictors as possible in the selected model.

In practice, the value of γ can be determined using the following procedure: (i) specify a sequence of different γ values; (ii) for each value of γ, run BSR for a short number of iterations; and (iii) choose the minimum value of γ for which the mode of $\Pi(|\xi_n||D^n, \gamma)$ coincides with the size of the MAP model. This procedure works very well for all of our examples, including both the simulated and real data examples. The resulting value of γ from this procedure is often consistent with the range suggested by Lemma 2.1. Also, we find that when the sample size is large, the choice of γ may not have much effect on the posterior. For instance, as shown in Table 2, the resulting mean, mode, and MAP models are all about the same for any $\gamma \in (0.75, 1)$ for the simulated example.

2.3 Simulation of the BSR Posterior

For simulation of the BSR posterior, we adopt the stochastic approximation Monte Carlo (SAMC) algorithm (Liang, Liu, and Carroll 2007). SAMC is an adaptive MCMC algorithm. But, unlike the adaptive Metropolis algorithm (Haario, Saksman, and Tamminen 2001), which adapts the proposal distribution, SAMC adapts the invariant distribution. In this sense, SAMC can also be viewed as a dynamic importance sampling algorithm. Indeed, as shown in Liang (2009), the samples produced by SAMC are correctly weighted with respect to the target distribution.

In the supplementary materials, we describe the details of the SAMC algorithm, particularly, how the proposal can be designed for high-dimensional GLMs. Compared with conventional MCMC algorithms, such as the Metropolis–Hastings algorithm, SAMC has a significant advantage in sample space exploration due to its self-adjusting mechanism: roughly

speaking, SAMC penalizes the visits to the oversivisited subregions and rewards the visits to the undervisited subregions, where "over" and "under" are both relative to the so-called desired sampling distribution. This mechanism makes SAMC essentially immune to the local trap problem and particularly suitable for sampling from high-dimensional space. Refer to the supplementary materials for more discussion of this issue.

3. BAYESIAN VARIABLE SELECTION

In this section, we propose a Bayesian variable selection procedure based on the marginal inclusion probability and prove its sure screening property and consistency. Identification of important variables based on the marginal inclusion probability is a widely used method for Bayesian variable selection; see, for example, Barbieri and Berger (2004) for the case of large-n-small-P normal linear models. However, two issues for this method remain unresolved. The first concerns its consistency, that is, whether this method will lead to a consistent selection of causal features for small-n-large-P GLMs. The second concerns how to determine the cutoff value for the marginal inclusion probability for a given dataset. These two issues will be addressed in Sections 3.1 and 3.2, respectively.

3.1 Consistency

With a slight abuse of notation, we let ξ_* denote the set of causal features. Let $\boldsymbol{e}_n = (e_1, e_2, \ldots, e_{P_n})$ denote the indicator vector of P_n features; that is, $e_j = 1$ if $\boldsymbol{x}_j \in \xi_*$ and 0 otherwise. Let q_j denote the marginal inclusion probability of $\boldsymbol{x}_j$, that is,

$$q_j = \sum_{\xi} e_{j|\xi} \Pi(\xi|D^n), \tag{18}$$

where $e_{j|\xi}$ indicates whether $\boldsymbol{x}_j$ is included in the model ξ. A conventional rule is to choose the features for which the marginal inclusion probability is greater than a threshold value $\hat{q}$; that is, setting $\widehat{\xi}_{\hat{q}} = \{\boldsymbol{x}_j : q_j > \hat{q}, j = 1, \ldots, P_n\}$ as an estimator of ξ_*. See, for example, Chapman et al. (2009), where the authors suggested $\hat{q} = 0.4$ for one example.

To establish the consistency of this rule, the following identifiability condition of ξ_* is needed. Let $A_{\epsilon_n} = \{\xi : d(\hat{f}(y|x, \xi), f(y|x)) \le \epsilon_n\}$, where $d(\cdot, \cdot)$ denotes the Hellinger distance between $\hat{f}$ and f. Define

$$\rho_j(\epsilon_n) = \sum_{\xi \in A_{\epsilon_n}} |e_j - e_{j|\xi}| \Pi(\xi|D^n),$$

which measures the distance between the true model and the sampled models for feature j in the ϵ_n-neighborhood A_{ϵ_n}.

(C_1) (Identifiability of ξ_*) $\max_{j \in \{1,2,\ldots,P_n\}} \rho_j(\epsilon_n) \to 0$ as $n \to \infty$ and $\epsilon_n \to 0$.

This condition states that as $n \to \infty$ and $\epsilon_n \to 0$, the true model is identifiable. In other words, when n is sufficiently large, if a model has the same density as the true density, then the model must coincide with the true model. This is a rather natural condition; otherwise, the likelihood function may lose its statistical meaning for variable selection.

Viewing $(1 - e_j, e_j)$ and $(1 - q_j, q_j)$ as two distributions defined on the space $\{0, 1\}$, we define $d(q_j, e_j)$ as the Hellinger distance between the two distributions. Lemma 3.1 concerns the upper bound of $d(q_j, e_j)$s. The proof of this lemma can be found in the Appendix.

Lemma 3.1. Assume the conditions of Lemma 2.1 and the condition (C_1) hold. Then, for any $\delta_n > 0$ and all sufficiently large n,

$$P\left[d^2(q_j, e_j) \le 2\delta_n + 2e^{-0.5cn\epsilon_n^2}\right] \ge 1 - e^{-0.5cn\epsilon_n^2},$$

for all $j = 1, 2, \ldots, P_n$.

Theorem 3.1 concerns the sure screening property and consistency of the variable selection rule $\widehat{\xi}_{\hat{q}}$, whose proof can be found in the Appendix.

Theorem 3.1. Assume that the conditions of Lemma 2.1 and the condition (C_1) hold.

(i) For any $\delta_n > 0$ and all sufficiently large n,

$$P\left(\max_{1 \le j \le P_n} |q_j - e_j| \ge 2\sqrt{\delta_n + e^{-0.5cn\epsilon_n^2}}\right) \le P_n e^{-0.5cn\epsilon_n^2}.$$

(ii) (Sure screening) For all sufficiently large n,

$$P(\xi_* \subset \widehat{\xi}_{\hat{q}}) \ge 1 - s_n e^{-0.5cn\epsilon_n^2},$$

where s_n denotes the size of ξ_*, for some choice of $\hat{q} \in (0, 1)$, preferably one not close to 0 or 1.

(iii) (Consistency) For all sufficiently large n,

$$P(\xi_* = \widehat{\xi}_{0.5}) \ge 1 - P_n e^{-0.5cn\epsilon_n^2}.$$

Regarding this theorem, we have a few remarks:

- The condition $\log(P_n) \prec n\epsilon^2$ in Lemma 2.1 implies $P_n e^{-0.5cn\epsilon_n^2} \to 0$ as $n \to \infty$.
- Theorem 3.1(i) implies that the MAP model is consistent, that is, it will converge to the true model in probability, and that the posterior probability of the true model will converge to 1, that is,

$$P(\xi_*|D^n) \to 1, \quad \text{as } n \to \infty, \tag{19}$$

which is the so-called global model consistency in Bayesian variable selection (Johnson and Rossell 2012). Regarding Bayesian variable selection, Johnson and Rossell (2012) distinguished two different types of consistency, global model consistency and pairwise model consistency. The pairwise model consistency pertains to the Bayesian factor between the true model and any other single model becoming large as the sample size increases, see, for example, Moreno, Girón, and Casella (2010). Note that the consistency of the MAP model stated above also belongs to the pairwise model consistency. For pairwise model consistency, one might have

$$P(\xi_*|D^n) \to 0, \quad \text{as } n \to \infty. \tag{20}$$

If (20) holds, then Bayesian prediction or inferential procedures that require model averaging might lose their effectiveness, as the number of models that must be averaged for valid Bayesian inference increases exponentially as P_n increases. For normal linear regression, Johnson and Rossell (2012) showed that (20) can be caused by a local prior of regression coefficients, which assigns a positive density

value to the regression coefficient vectors that contain components equal to 0, together with a prior of subset models that satisfies the condition

$$\inf_{\xi \in \mathcal{J}_+} \Pi(\xi)/\Pi(\xi_*) > \delta > 0, \tag{21}$$

for some constant δ, where $\mathcal{J}_+$ denotes the set of models containing all true predictors plus one additional predictor. It is interesting to point out that although our prior (5) for regression coefficients is local, our prior for subset models, given in (8) and (10), does not satisfy (21). For our priors, $\inf_{\xi \in \mathcal{J}_+} \Pi(\xi)/\Pi(\xi_*) = \nu_n/(1-\nu_n) \to 0$ as $n \to \infty$. That is, the automatic multiplicity correction property embedded in our priors plays a critical role for the global model consistency. For normal linear regression, Johnson and Rossell (2012) provided some prior settings that lead to the global model convergence. As a contribution of this article, we have provided a general prior setting for GLMs that leads to the same global model convergence.

- In Theorem 3.1, the word "screening" has a slightly different meaning from that in SIS (Fan and Song 2010). SIS first selects a subset of variables based on their marginal utility (i.e., screening variables based on marginal utility), and then refines the selection using Lasso or SCAD. In contrast, the rule $\widehat{\xi}_{\hat{q}}$ is based on marginal inclusion probability, and a full exploration of the model space is required to make the screening. Here, the word "screening" is used based on the following consideration: although the MAP model is consistent, it is true only in the sense of asymptotics. For any given dataset, the MAP model might not capture all true predictors. The rule $\widehat{\xi}_{\hat{q}}$ allows one to select somewhat more predictors that are worthy of further investigation, though possibly through a different and expensive experiment.
- Theorem 3.1 establishes that the rule $\widehat{\xi}_{\hat{q}}$ has the same sure screening property as SIS while under much simpler conditions. SIS requires very complicated conditions about predictors, while the conditions required by $\widehat{\xi}_{\hat{q}}$ are rather simple and can be satisfied by almost all GLMs with a sparse true model. More importantly, $\widehat{\xi}_{\hat{q}}$ uses the joint information from all predictors and thus can generally outperform SIS. As previously noted, SIS uses only the marginal information of predictors in variable screening.

Theorem 3.1 establishes the consistency of the rule $\widehat{\xi}_{\hat{q}}$ from the theoretical perspective. To implement this rule, one needs a consistent estimator for the marginal inclusion probability and a consistent estimator for the MAP model. As discussed in the supplementary materials, the estimator given in Equation (15) of the supplementary materials is consistent for the marginal inclusion probability, and the sampled highest posterior probability model provides a consistent estimator for the MAP model. Both the estimators are based on the SAMC samples drawn from the posterior $\Pi(\xi_n|D^n)$.

3.2 A Multiple Test-Based Sure Variable Screening Procedure

Although Theorem 3.1 gives us nice asymptotic properties for the variable selection rule $\widehat{\xi}_{\hat{q}}$, it is still unclear how to choose $\hat{q}$ for a given dataset. In the following, we propose a multiple-hypothesis test-based sure variable screening procedure for determining the value of $\hat{q}$. Henceforth, this procedure will be abbreviated by SVS. SVS is motivated by Theorem 3.1(i), which implies that causal features tends to have larger marginal inclusion probabilities than noncausal features.

Let $q_1, q_2, \ldots, q_{P_n}$ denote the marginal inclusion probabilities of P_n features, and let

$$z_i = \Phi^{-1}(q_i), \quad i = 1, 2, \ldots, P_n$$

denote the corresponding marginal inclusion scores (MIS), where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard Gaussian distribution. To identify the features that have significantly large MIS, we model the MIS by a two-component mixture exponential power distribution, for which the left component corresponds to noise features and the right causal features. For an m-component mixture exponential power distribution, the density function is given by

$$g(z|\vartheta_m) = \sum_{i=1}^{m} \varpi_i \psi(z|\nu_i, \sigma_i, \alpha_i), \tag{22}$$

where $\vartheta_m = (\varpi_1, \nu_1, \sigma_1, \alpha_1, \ldots, \varpi_m, \nu_m, \sigma_m, \alpha_m)$ contains all parameters of the distribution, ϖ_i is the weight of the ith component with $0 < \varpi_i < 1$ and $\sum_{i=1}^{m} \varpi_i = 1$, and

$$\psi(z|\nu_i, \sigma_i, \alpha_i) = \frac{\alpha_i}{2\sigma_i \Gamma(1/\alpha_i)} \exp\left\{-(|z-\nu_i|/\sigma_i)^{\alpha_i}\right\}, \quad -\infty < \nu_i < \infty,\ \sigma_i > 0,\ \alpha_i > 1, \tag{23}$$

where the parameters ν_i, σ_i, and α_i represent the center, dispersion, and decay rate of the distribution, respectively. For $\alpha_i = 2$, the distribution (23) is reduced to $N(\nu_i, \sigma_i^2/2)$; for $1 < \alpha_i < 2$, the distribution is heavy-tailed; and for $\alpha_i > 2$, the distribution is light-tailed.

The identifiability of (22) was established in Holzmann, Munk, and Gneiting (2006).

The parameters ϑ_m can be estimated as in Liang and Zhang (2008) by minimizing the Kullback–Leibler (KL) divergence

$$\mathrm{KL}(g_{\vartheta_m}, g) = -\int \log\left\{\frac{g(z|\vartheta_m)}{g(z)}\right\} g(z)dz,$$

where $g(z)$ denotes the unknown true density of z_i's. For a given value of m, the minimization can be done using the stochastic approximation algorithm, refer to Liang and Zhang (2008) for the details. One significant advantage of this algorithm is that it permits the general dependence between z_i's. A proof of convergence for this algorithm can be found in Zhang and Liang (2008). The cutoff value z_r, which corresponds to the setting $\hat{q} = \Phi(z_r)$, can be chosen by controlling the false discovery rate (FDR) of causal features at a prespecified test level. Figure 1 depicts such a rule. For a given rule $\Lambda_r = \{Z_i \geq z_r\}$, the FDR can be estimated by

$$\mathrm{FDR}(\Lambda_r) = \frac{P_n \hat{\varpi}_1 [1 - F(z_r|\hat{\nu}_1, \hat{\sigma}_1, \hat{\alpha}_1)]}{\#\{z_i : z_i \geq z_r\}}, \tag{24}$$

where $\#\{z_i : z_i \geq z_r\}$ denotes the number of features with the MIS greater than z_r, and $F(\cdot)$ denotes the CDF of the exponential power distribution (23). Define the q-value (Storey 2002) as

$$q_r^s(z) = \inf_{\{\Lambda_r : z \in \Lambda_r\}} \mathrm{FDR}(\Lambda_r), \tag{25}$$

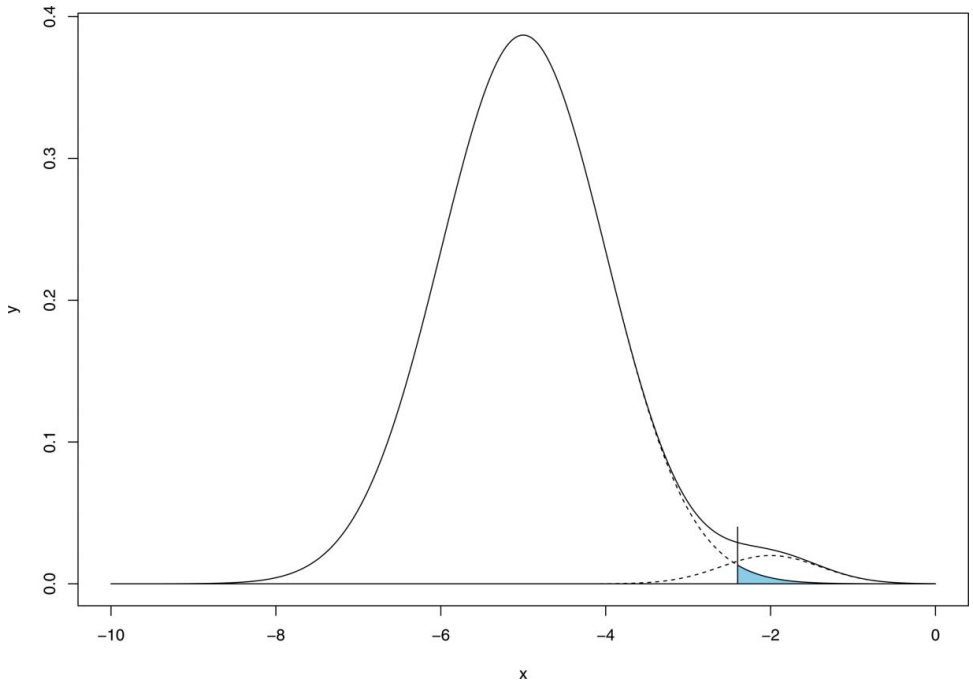

Figure 1. Illustrative plot for SVS: marginal inclusion scores are fitted by a two-component mixture exponential power distribution. The vertical line shows the cutoff z_r, and the shaded area shows the portion of falsely selected causal features.

which can be used as the reference quantity for the decision of multiple-hypothesis tests. For example, we can set the test level to be 0.01, that is, choosing z_r such that $q_r^s(z) \le 0.01$ for all $z \ge z_r$. Clearly, this rule possesses the sure screening property when n is sufficiently large. Finally, we note that other FDR methods, which account for the dependence between testing p-values, for example, Benjamini, Krieger, and Yekutieli (2006), can also be used in SVS.

4. WHY IS BAYESIAN SUBSET REGRESSION PREFERRED?

In this section, we evaluate the influence of the dimension P_n on BSR and penalized likelihood methods through a subset-full data comparison study. The full data we considered is the lymph data; see Section 7 for the description of this data. The subset data consist of only 34 genes, which correspond to the set of significant genes identified by SVS at an FDR level of 0.01 based on the marginal inclusion probabilities produced by SAMC in a run on the full data with $\gamma = 0.85$. The names of these genes are given in the supplementary materials.

The penalized likelihood methods, including Lasso, elastic net, SIS, and ISIS, were first applied to the subset data. The package we used for Lasso and elastic net is *glmnet*, which was recently developed by Friedman, Hastie, and Tibshirani (2010). This package employs the coordinate-wise optimization algorithm and appears to be the fastest for computing regularization paths. For Lasso and elastic net, the penalty term is chosen in the form

$$P_{\alpha,\lambda}(\boldsymbol{\beta}) = \lambda \left[(1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}_{[-0]}\|_{l_2}^2 + \alpha\|\boldsymbol{\beta}_{[-0]}\|_{l_1} \right],$$

where $0 < \alpha \le 1$, $\boldsymbol{\beta}_{[-0]}$ denotes the vector of regression coefficients excluding the intercept, and $\|\cdot\|_{l_2}$ and $\|\cdot\|_{l_1}$ denote, respectively, the L_2 and L_1 norms of a vector. The parameter λ is chosen by minimizing the leave-one-out cross-validation misclassification rate. The penalized likelihood methods assume that the intercept term is always included in the selected model, although this is unnecessary. Henceforth, "leave-one-out cross-validation misclassification rate" is abbreviated by "1-CVMR." Lasso corresponds to the case $\alpha = 1$ and elastic net corresponds to the case $0 < \alpha < 1$. In this article, we set $\alpha = 0.5$ for elastic net in all examples. The numerical results are summarized in Table 1, where we report the minimum 1-CVMR (r_{cv}) and the corresponding model size ($|\xi_n|$) that Lasso and elastic net achieved.

Table 1. Comparison of BSR, Lasso, elastic net, SIS, and ISIS for the subset and full data of lymph: $r_{\text{cv}}(\%)$ refers to the minimum 1-CVMR, and MAP_{min} refers to the minimum size MAP_i model with zero 1-CVMR, where MAP_i denotes the MAP model consisting of i features

	Lasso		Elastic net		SIS		ISIS		BSR			
									MAP		MAP_{min}	
Data	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$	$\lvert\xi_n\rvert$	$r_{\text{cv}}(\%)$
Sub	28	2.03	31	2.70	7	16.9	7	7.43	8	2.70	10	0
Full	23	15.5	51	16.9	7	17.6	7	10.8	8	0.68	10	0

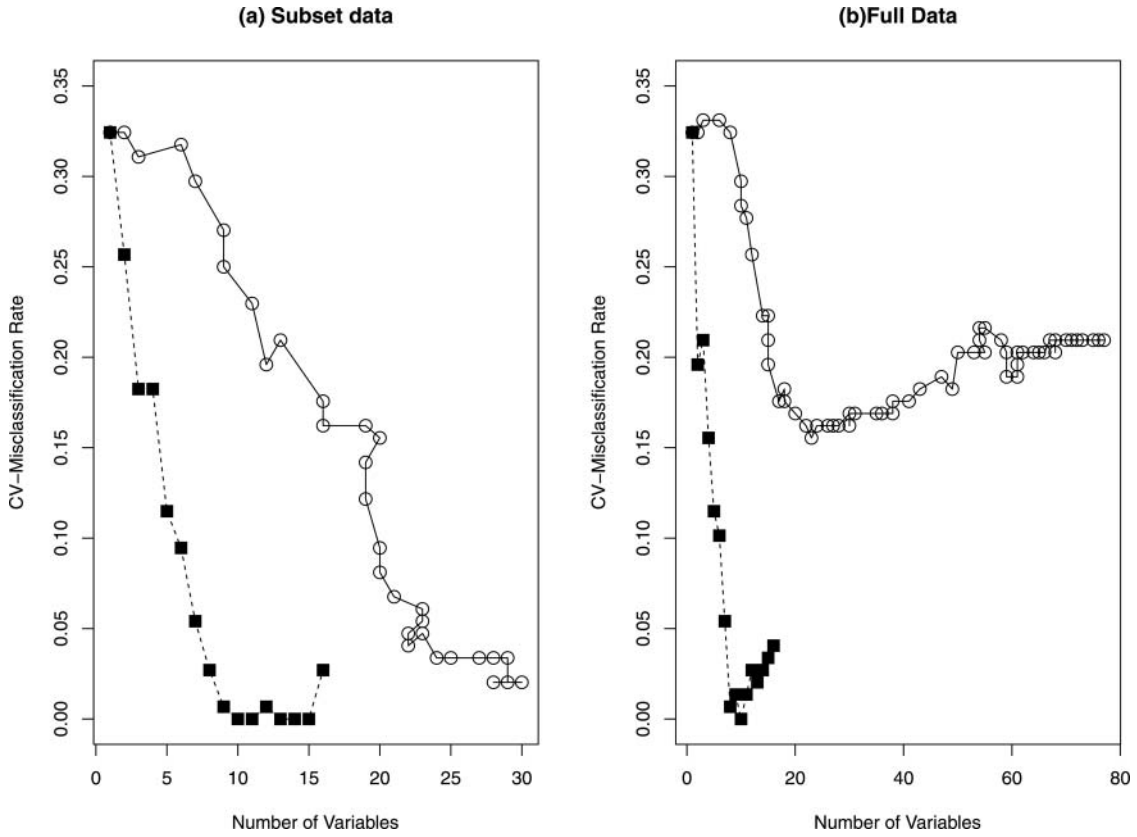

Figure 2. Comparison of Lasso and BSR for (a) the subset data and (b) the full data of lymph: 1-CVMRs of the models selected by Lasso (white circles connected by solid line) and the MAP_1–MAP_{16} models selected by BSR (black squares connected by dotted line).

The R package we used for SIS and ISIS is *SIS* recently developed by Fan et al. (2010), in which the regression coefficients of the selected model are calculated under the SCAD penalty. The results are also summarized in Table 1, where we report the 1-CVMR (r_{cv}) and the corresponding model size ($|\xi_n|$). The results were produced under the default setting with marginal likelihood ratios used as the marginal utility. The results under the setting of marginal regression coefficients were similar.

We then applied BSR to this subset data with the same prior setting as for the full data. We found two different models, both of which, considering the approximation nature of the posterior (14), should be considered as the MAP model. Each of the two models consists of eight genes and produces a perfect classification of the subjects, and the difference between their log-posterior probabilities is only 2.66×10^{-15}! The smaller 1-CVMR of the two models is 2.70% as reported in Table 1, and the larger one is 8.78%. In addition, we found many different models with 9–16 genes, each providing a perfect classification of the subjects. Among these models, some with 10, 11, 13, 14, and 15 genes have zero 1-CVMR. Therefore, we reported in Table 1 that the minimum size of the models with zero 1-CVMR is 10. This is a remarkable result. Compared with BSR, the models found by Lasso, elastic net, SIS, and ISIS are far from "optimal." Although the models found by Lasso and elastic net have an acceptable 1-CVMR, their sizes are too large. In contrast, the models found by SIS and ISIS have an acceptable size, but their 1-CVMRs are too high.

To study the influence of the dimension P_n on the performance of these methods, we then applied them to the full lymph data, which consist of 4514 predictors, as explained in Section 7. Figure 2 compares 1-CVMRs of the models produced by Lasso along its regularization path with the minimum 1-CVMRs of the MAP_i models, $i = 1, \ldots, 16$, produced by BSR with $\gamma = 0.85$ (see Section 7 for the details of BSR setting). Here, we use MAP_i to denote the MAP model consisting of i predictors. For each $i \in \{8, \ldots, 16\}$, we found multiple models from both the subset and full data, each providing a perfect classification of the subjects. Hence, only the minimum 1-CVMR of these models is reported in Figure 2. For Lasso, the minimum 1-CVMR is 2.03% for the subset data, but 15.5% for the full data. From this example, we may conclude that Lasso scales badly, its performance deteriorating as P_n increases. As shown in Table 1, elastic net, SIS, and ISIS all work similarly to Lasso, with their performance adversely affected by the dimension P_n. However, BSR does not possess this shortcoming. In the runs described above, BSR found perfect classification models with 8–16 features for the subset data, and perfect classification models with 8–23 features for the full data. The 1-CVMRs of the MAP_i models with $i = 17, \ldots, 23$ features are not shown in Figure 2(b). The explanation for this phenomenon is as follows: due to their forward variable selection nature (see, e.g., Park and Hastie 2007), Lasso and elastic net are sensitive to noise features and tend to be adversely affected by the dimension P_n. Since SIS and ISIS use only the marginal information or limited joint information of features, they are also sensitive to noise features and tend to be adversely affected by the dimension P_n. However, BSR conducts a global search over the entire model space and makes use of the joint information of all features,

hence, BSR is insensitive to noise features and less affected by increasing P_n. This comparison study implies that BSR is more suitable for high-dimensional variable selection than these penalized likelihood methods, although the latter methods are computationally more attractive.

We have two additional remarks regarding this experiment:

- BSR scales well. As previously noted, it found perfect classification models with 8–16 features for the subset model, and perfect classification models with 8–23 features for the full data. It is worth pointing out that, as shown in Figure 2, the perfect classification models from the full data tend to have higher 1-CVMRs than those from the subset data. This results because the models selected from the full data include some features other than the 34 genes, and those additional features tend to have a lower prediction ability. This suggests a possible two-stage approach for variable selection: screen candidate features via SVS in the first stage and then refine the model based on the features selected by SVS in the second stage.
- For the subset data, Lasso works reasonably well, although still inferior to BSR. The reason why Lasso tends to select a large model is understandable: due to the shrinkage of regression coefficients, the prediction ability of each feature is lowered and, hence, Lasso needs to select more features for compensation.

5. SIMULATION STUDIES

This example mimics a case–control genetic association study. The response variable y, which represents the disease status of a subject, takes value 1 for the case and 0 for the control. The explanatory variables are generated as SNPs in the human genome. Each variable x_{ij}, the genotype of SNP j of subject i, takes value 0, 1, or 2, where 0 stands for a homozygous site with the major allele, 1 stands for a heterozygous site with minor allele, and 2 stands for a homozygous site with two copies of the minor allele. Following Chen and Chen (2012), the data were generated using the following procedure:

Let n_1 and n_2 denote the numbers of cases and controls, respectively. Let $\boldsymbol{s} = \{1, 2, \ldots, k\}$ denote the set of causal SNPs for the disease. Thus, there are a total of 3^k possible genotype profiles for the k SNPs. For the SNPs belonging to $\boldsymbol{s}$, the disease risk model is given by

$$\operatorname{logit} P(y_i = 1 | \boldsymbol{x}_i(\boldsymbol{s})) = \sum_{i=1}^{k} \beta_j x_{ij},$$

for prespecified values of $\beta_1, \ldots, \beta_k$. Therefore, the true model for this example contains the intercept term and k SNPs, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. For the noncausal SNPs $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_{P_n}$, the genotype profile of each subject is generated under the assumption of Hardy–Weinberg equilibrium. That is, for each $\boldsymbol{x}_j$, for $j = k+1, \ldots, P_n$, x_{ij} is independently 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). Note that due to the logit selection, the genotypes of the causal SNPs are not uniform on $\{0, 1, 2\}$.

This example consists of 10 simulated datasets. Each was generated with $n_1 = n_2 = 500$, $P_n = 10,001$ (including the intercept term), $k = 8$, and $(\beta_1, \ldots, \beta_8) = (0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3)$. For each dataset, SAMC was run six times with $K_n \equiv 25$ and $\gamma = 0.5, 0.75, 0.85, 1.0, 1.5$, and 2.5. To evaluate the effect of γ on the performance of BSR, we tried different values of γ, even for some outside the range suggested by Lemma 2.1. Each run of SAMC consisted of 2.05×10^6 iterations, where the first 50,000 iterations were discarded for the burn-in process and the remaining iterations were used for inference. In the runs of SAMC, the sample space was partitioned according to the energy function into 401 subregions: $E_1 = \{\xi_n : U(\xi_n) \le 351\}$, $E_2 = \{\xi_n : 351 < U(\xi_n) \le 352\}, \ldots,$ $E_{400} = \{\xi_n : 749 < U(\xi_n) \le 750\}$, $E_{401} = \{\xi_n : U(\xi_n) > 750\}$, where $U(\xi_n) = -\log \Pi(\xi_n | D^n)$ is the negative log-posterior probability of the model ξ_n and also the so-called energy function. The gain factor sequence $\{a_t\}$ was chosen with $t_0 = 1000$ in the form:

$$a_t = \frac{t_0}{\max\{t, t_0\}}, \quad t = 1, 2, \ldots. \tag{26}$$

The numerical results are summarized in Table 2. They indicate that (i) as γ increases, the mean of $\Pi(|\xi_n||D^n, \gamma)$, that is, the expected size of posterior models, tends to decrease; (ii) the mode of $\Pi(|\xi_n||D^n, \gamma)$ is quite consistent with the mean of $\Pi(|\xi_n||D^n, \gamma)$; and (iii) the size of the MAP model decreases as γ increases. For this example, it follows from the rule (17) that $\gamma = 0.85$ is appropriate. The paired two-sample t-test indicates that the MAP model size and the mode of $\Pi(|\xi_n||D^n, \gamma)$ achieved at $\gamma = 0.85$ are not significantly different from each other. Note that the choice of $\gamma = 0.85$ falls into the interval $(1 - 1/\kappa, 1)$ suggested by Theorem 2.1. If we estimate κ by $\log(P_n)/\log(n) = 1.33$, then $(1 - 1/\kappa, 1) \approx (0.25, 1)$.

SVS was then applied to the 10 datasets. Figure 3 illustrates this procedure for one dataset. At a nominal FDR level of 0.01, the 13 features identified for this dataset covered all 9 true causal features. Note that for the simulated data, the distribution of MIS typically has a light right tail because of the independence of the causal and noncausal features. For the real data, the distribution of MIS may have a heavy right tail due to the general correlation among different features; see, for example, Figure 5 for the gene

Table 2. Numerical results for the simulated data: [a]p-value for the paired two-sample t-test for $H_0 : \arg\max_{|\xi_n|} \Pi(|\xi_n||D^n, \gamma) = |\xi_{\text{map},\gamma}|$ versus $H_0 : \arg\max_{|\xi_n|} \Pi(|\xi_n||D^n, \gamma) \neq |\xi_{\text{map},\gamma}|$. All estimates in the table were calculated as the average over 10 datasets with their standard deviations given in parentheses

γ	0.5	0.75	0.85	1.0	1.5	2.5
Mean of $\Pi(\lvert\xi_n\rvert\rvert D^n, \gamma)$	20.53(0.28)	9.88(0.28)	8.94(0.20)	8.21(0.19)	7.40(0.17)	6.13(0.28)
Mode of $\Pi(\lvert\xi_n\rvert\rvert D^n, \gamma)$	21.6(0.60)	9.5(0.34)	8.6(0.22)	8.1(0.23)	7.5(0.17)	6.1(0.31)
Size of MAP model	9.4(0.40)	8.8(0.42)	8.2(0.25)	8.2(0.25)	7.4(0.16)	6.2(0.33)
p-value[a]	1.93×10^{-9}	1.32×10^{-3}	0.104	0.343	0.343	0.343

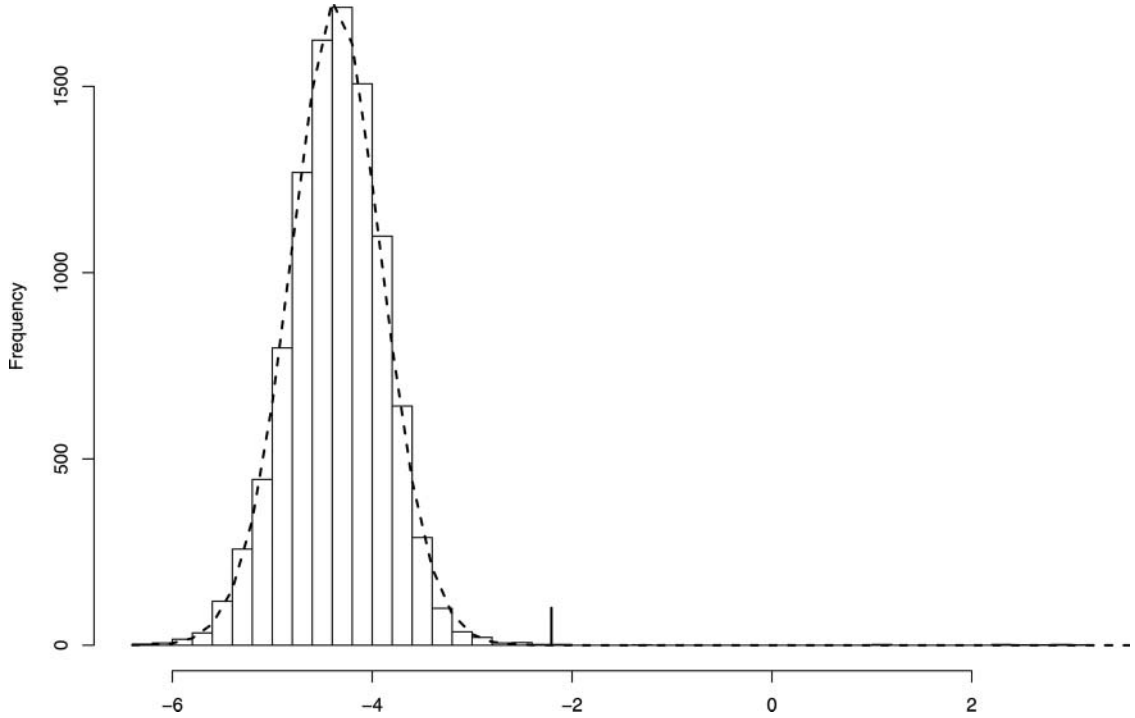

Figure 3. Illustrative plot of SVS for one simulated dataset. The dashed curve shows the estimated density function for the MIS (produced by SAMC at $\gamma = 0.85$) of nonrelevant features, and the vertical bar shows the classification rule at an FDR level of 0.01.

expression data. In this case, a considerable number of features can be identified.

To measure the performance of SVS, we calculated the false selection and negative selection rates. Let $\boldsymbol{s}_i^*$ denote the set of selected features for dataset i. Define

$$\text{fsr} = \frac{\sum_{i=1}^{10} |\boldsymbol{s}_i^* \setminus \boldsymbol{s}|}{\sum_{i=1}^{10} |\boldsymbol{s}_i^*|}, \quad \text{nsr} = \frac{\sum_{i=1}^{10} |\boldsymbol{s} \setminus \boldsymbol{s}_i^*|}{\sum_{i=1}^{10} |\boldsymbol{s}|},$$

where $|\cdot|$ denotes the cardinality of a set. The smaller the values of fsr and nsr, the better the performance of the method. The results are summarized in Table 3. SVS worked remarkably well for this example: it was able to select all causal features at $\gamma = 0.75$ and 0.85 for all datasets at a very low false selection rate as compared with Lasso, elastic net, SIS, and ISIS. When γ is far from the suggested range, for example, $\gamma = 1.5$ and 2.5, SVS tended to select more features but with some causal features missed. Note that in Table 3, the true and nominal levels of fsr are quite different. This is due to the presence of spurious features in the data. Spurious and causal features are not distinguishable based on the MIS alone.

For the simulated data, we also calculated fsr and nsr for the MAP models produced with $\gamma = 0.85$. The results are summarized in Table 4. For comparison, the penalized likelihood methods, including Lasso, elastic net, SIS, and ISIS, were also applied to this example. For each dataset, both SIS and ISIS selected 36 features. When reading Table 4, the reader needs to be noted that the size for Lasso, elastic net, SIS, and ISIS does not include the intercept term, while the size for MAP and SVS includes the intercept term. The intercept term is automatically included in all models by Lasso, elastic net, SIS, and ISIS, but it is treated as a selectable feature in BSR. The comparison shows that Lasso, elastic net, SIS, and ISIS are all too liberal. They tend to select too many irrelevant features, as indicated by their high fsr values. The MAP models are parsimonious, but they tend to miss some causal features and thus have a high nsr. In terms of both fsr and nsr, SVS appears to be the best variable screening procedure for this example.

6. GENE EXPRESSION DATA EXAMPLES

In this section, we test BSR on four gene expression datasets: lymph (Hans, Dobra, and West 2007), colon (Alon et al. 1999), leukemia (Golub et al. 1999), and prostate (Singh et al. 2002). Table 5 summarizes information of these datasets. Each dataset was preprocessed; all gene expression profiles were log-transformed and, to prevent single arrays from dominating the analysis, standardized to zero mean and unit variance. Refer to Irizarry et al. (2003) for details of the normalizing procedure.

Table 3. Results of SVS for the simulated data. [a]Average model size (over 10 datasets) with the standard deviation presented in the parentheses; [b]results at the FDR level 0.01; [c]results at the FDR level 0.001

γ	0.5[b]	0.75[b]	0.85[b]	0.85[c]	1.0[b]	1.5[b]	2.5[b]
Size[a]	48.1(3.01)	18.7(1.11)	14.2(0.70)	12.1(0.43)	13.6(0.58)	17.0(1.07)	49.9(7.43)
fsr(%)	81.70	51.87	36.62	25.62	35.29	48.24	83.57
nsr(%)	2.22	0	0	0	2.22	2.22	8.89

Table 4. Comparison of BSR with Lasso, elastic net, SIS, and ISIS for the simulated data. [a]Results of SVS at an FDR level of 0.001; [b]average model size (over 10 datasets) produced by different methods with the standard deviation presented in the parentheses

	BSR($\gamma = 0.85$)					
Methods	MAP	SVS[a]	Lasso	Elastic net	SIS	ISIS
Size[b]	8.2(0.25)	12.1(0.43)	44.9(20.48)	30.1(9.04)	36(0)	36(0)
fsr(%)	3.66	25.6	81.8	73.4	77.8	77.8
nsr(%)	12.22	0	0	0	0	0

6.1 Lymph Data

This dataset consists of $n = 148$ samples with 100 node-negative cases (low risk for breast cancer) and 48 node-positive cases (high risk for breast cancer). After prescreening, Hans, Dobra, and West (2007) selected 4512 genes that showed evidence of variation above the noise levels for further study. In addition, they included two clinical factors, estimated tumor size (in centimeters) and protein assay-based estrogen receptor status that was coded as a binary variable, as candidate predictors. This brought the total number of predictors to $P_n = 4514$.

For this dataset, SAMC was run five times with $K_n \equiv 25$ and $\gamma = 0.75$, 0.8, 0.85, 0.9, and 0.99. Each run consisted of 1.005×10^7 iterations, where the first 50,000 iterations were discarded for burn-in and the remaining iterations were used for inference. The sample space was partitioned according to the energy function into 201 subregions: $E_1 = \{\xi_n : U(\xi_n) \le 1\}$, $E_2 = \{\xi_n : 1 < U(\xi_n) \le 2\}, \ldots, E_{200} = \{\xi_n : 199 < U(\xi_n) \le 200\}$, $E_{201} = \{\xi_n : U(\xi_n) > 200\}$, where $U(\xi_n) = -\log \Pi(\xi_n|D^n)$ is the energy function. The gain factor sequence was chosen according to (26) with $t_0 = 2000$. According to the rule (17), we chose $\gamma = 0.85$ for this example. The resulting posterior distribution $\Pi(|\xi_n||D^n, \gamma = 0.85)$ is shown in Table 6, which has a mode at $|\xi_n| = 8$. The MAP model also consists of eight features.

Hans, Dobra, and West (2007) applied their SSS algorithm to this data, but with different prior distributions. They assumed that the regression coefficients follow independent prior normal distributions with a constant precision τ, that is, $\boldsymbol{\beta}_{\xi_n} \sim N(0, \tau I_d)$, where $d = |\xi_n|$ is the model size, and that ξ_n follows a binomial prior distribution

$$P(\xi_n) = \nu^d (1-\nu)^{P_n - d},$$

where ν denotes the prior probability of each predictor being selected for subset models. For the prior hyperparameters τ and ν, they typically set $\tau = 1$ and $\nu = d'/P_n$ for a small value of d' to encourage sparsity of the selected models. For this dataset, they set $\tau = 1$ and $d' = 10$ and the resulting posterior $\Pi(|\xi_n||D^n)$ had a mode at 5. Hans, Dobra, and West (2007) identified the top eight genes ranked in marginal inclusion probabilities: rgs3, dxy155e, atp6v1f, mgc8721, vdac1, gem, wsb1, and prrg1. Under the BSR priors, the top eight genes we identified are rgs3, mgc8721, dxys155e, lmx1b, tmcc1, ctnna1, hs.458986, and gng12. The number of overlap genes is three. We compared the quality of the two sets of genes by calculating their 1-CVMR: the set of genes found by Hans, Dobra, and West (2007) had a 1-CVMR of 8.11%, while that by BSR had a 1-CVMR of 4.73%.

Table 5. Summary of four gene expression datasets

Dataset	Publication	n	P	Response
Lymph	Hans, Dobra, and West (2007)	148	4514	Positive/negative node
Colon	Alon et al. (1999)	62	2000	Tumor/normal tissue
Leukemia	Golub et al. (1999)	72	3571	Subtype of leukemia
Prostate	Singh et al. (2002)	102	6033	Tumor/normal tissue

For comparison, we also applied Lasso, elastic net, SIS, and ISIS to these data. By minimizing 1-CVMR, Lasso selected 23 genes with a minimum 1-CVMR of 15.5%, and elastic net selected 51 features (including the estimated tumor size and 50 genes) with a minimum 1-CVMR of 16.9%. Both SIS and ISIS selected seven genes, and the resulting 1-CVMRs were 17.6% and 10.8%, respectively. As reported in Table 1, the MAP model of BSR consisted of eight genes, and its 1-CVMR was only 0.68%. In addition, BSR found a model of 10 genes that produced zero 1-CVMR for these data! This example indicates that BSR can provide a dramatic improvement over the penalized likelihood methods for variable selection.

Figure 5(a) shows that genes relevant to the positivity of lymph nodes can be identified using SVS. At an FDR level of 0.01, SVS identified 34 genes; and at the level 0.001, it identified 21 genes. Figure 4 compares the q-values of the genes selected by BSR, SIS, ISIS, Lasso, and elastic net. As indicated by the plots, the genes selected by the four penalized likelihood methods are quite inconsistent with those selected by BSR. The penalized likelihood methods sometimes selected low-MIS genes. To justify our conclusion that BSR had selected important genes for identifying cases with positive lymph nodes, we applied all five methods to the subset data formed by the 34 genes identified by SVS at the FDR level 0.01. The results are presented in Section 4. As shown in Table 1, the performance of all the four methods, Lasso, elastic net, SIS, and ISIS, were improved on this subset data. This experiment shows that increasing the dimension P_n causes the performance of the penalized likelihood methods to deteriorate, while this does not happen for BSR.

6.2 Colon, Leukemia, and Prostate Data

For each of these datasets, SAMC was run five times with $K_n \equiv 25$ and $\gamma = 0.75$, 0.8, 0.85, 0.9, and 0.99. Each run consisted of 5.05×10^6 iterations, where the first 50,000 iterations were discarded for burn-in and the remaining iterations were used for inference. The sample space was partitioned

Table 6. Posterior distribution $\Pi(|\xi_n||D^n, \gamma = 0.85)$ obtained by BSR for the lymph data

	Model size ($|\xi_n|$)							
Data	≤ 5	6	7	8	9	10	11	≥ 12
Lymph	1.1×10^{-4}	1.54×10^{-4}	2.17×10^{-4}	0.646	0.237	0.086	0.024	6.47×10^{-3}

according to the energy function $U(\xi_n) = -\log \Pi(\xi_n|D^n)$ into 101 subregions: $E_1 = \{\xi_n : U(\xi_n) \leq 1\}$, $E_2 = \{\xi_n : 1 < U(\xi_n) \leq 2\}, \ldots, E_{100} = \{\xi_n : 99 < U(\xi_n) \leq 100\}$, $E_{101} = \{\xi_n : U(\xi_n) > 100\}$. Since these datasets contain smaller numbers of samples than lymph, a narrower energy range was considered and thus a shorter number of iterations was run. The gain factor sequence was chosen according to (26) with $t_0 = 1000$. According to the rule (17), we chose $\gamma = 0.99$ for all these data. The resulting posterior distributions of $|\xi_n|$ are shown in Table 7. For the colon data, the MAP model consists of three features, but $\Pi(|\xi_n||D^n)$ contains two modes at 2 and 4, respectively. For the multimodal case, we calculated the posterior mean of $|\xi_n|$ and found that the mean was 4.43 at $\gamma = 0.9$ and 3.21 at $\gamma = 0.99$. Therefore, we chose $\gamma = 0.99$ for this dataset.

As suggested by Table 7, BSR tends to select parsimonious models for these data. The MAP model for the colon data consists of three genes, g.451, g.576, and g.2000, where the number corresponds to the line number of the colon dataset posted on M. Dettling's homepage *http://stat.ethz.ch/~dettling/bagboost.html* and this is the same for the leukemia and prostate datasets. For the colon data, we also found many perfect classification models containing four or more genes. For the leukemia data, the MAP model contains two genes, but it is not unique. For example, (g.435,g.956), (g.626,g.1393), (g.888,g.956), (g.956, g.1205), (g.956, g.2196), and (g.3292,g.3441) all form MAP models with perfect classifications of the subjects. For this dataset, we also found many perfect classification models containing three or more genes. For the prostate data, the MAP

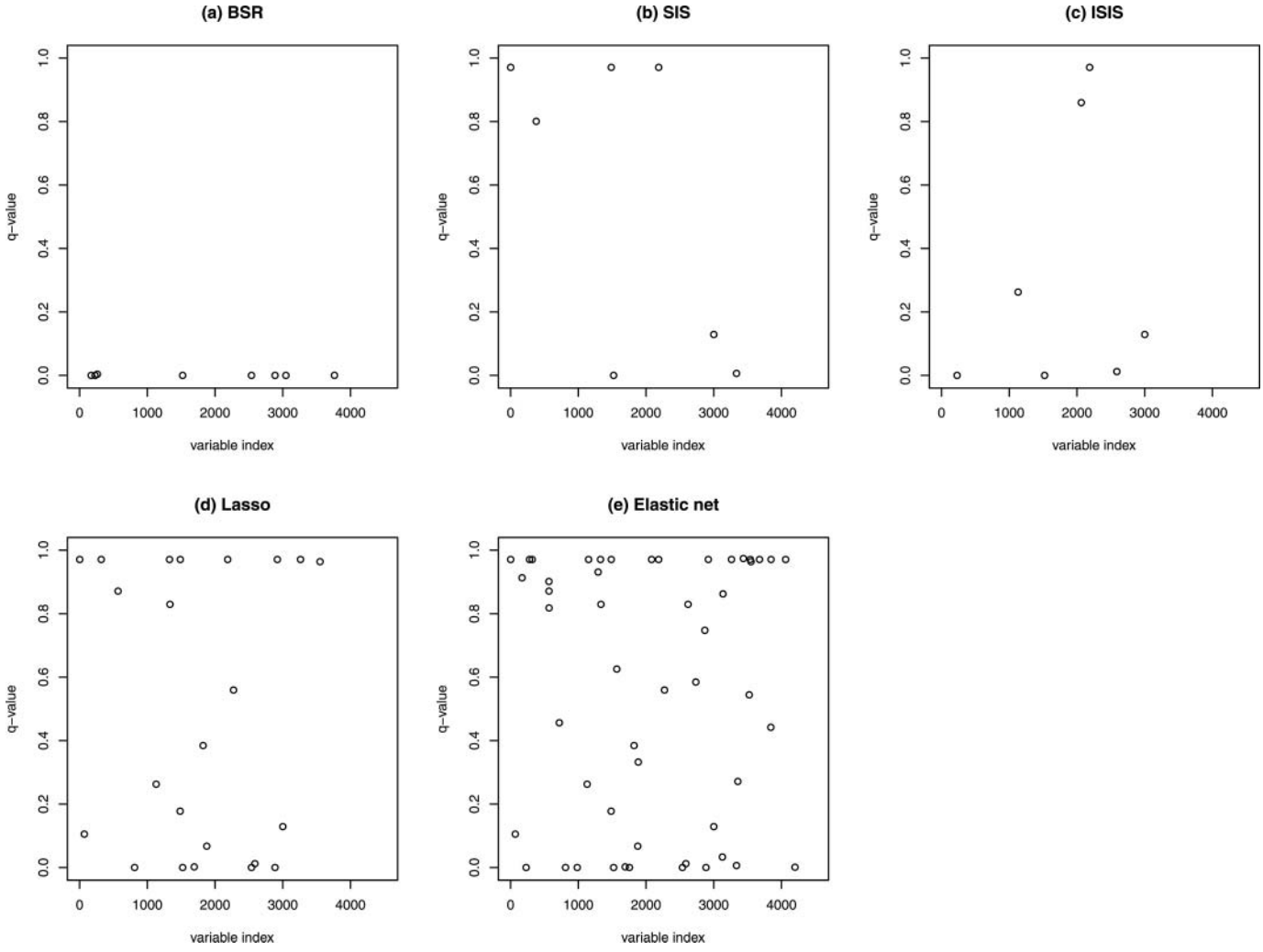

Figure 4. Plots of q-values for the genes selected by BSR, SIS, ISIS, Lasso, and elastic net. The model shown in (a) is one of the MAP models found by BSR with $\gamma = 0.85$ and it has a 1-CVMR of 2.7%. The models shown in (b), (c), (d), and (e) were selected by SIS, ISIS, Lasso, and elastic net, and have 1-CVMRs 17.6%, 10.8%, 15.5%, and 16.9%, respectively.

Table 7. Posterior distribution $\Pi(|\xi_n||D^n, \gamma = 0.99)$ obtained by BSR for colon, leukemia, and prostate data

	Number of variables ($|\xi_n|$)									
Dataset	1	2	3	4	5	6	7	8	9	≥ 10
Colon	0.001	0.387	0.201	0.260	0.118	0.029	0.004	< 0.001	< 0.001	< 0.001
Leukemia	0.023	0.561	0.337	0.070	0.007	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Prostate	0	0.005	0.098	0.681	0.191	0.024	0.002	< 0.001	< 0.001	< 0.001

model contains three genes g.2619, g.2614, and g.4898. In addition, we found many perfect classification models containing four or more genes. These results are summarized in Table 8. Note that in BSR, the perfect classification models are also MAP_i models due to the use of a L_0-penalty-like prior. The results are surprising. For each dataset, BSR could find many models of different sizes, each producing a perfect classification for the subjects and having zero 1-CVMR.

Figure 5(b)–5(d) shows that the genes relevant to the disease can be identified based on the MIS through the SVS procedure. Due to the general correlation between gene expression profiles, a considerable number of relevant genes were identified for each dataset. At an FDR level of 0.01, we identified 84, 47, and 101 genes for the colon, leukemia, and prostate data, respectively. At the level 0.001, we identified 59, 26, and 31 genes for the colon, leukemia, and prostate data, respectively. At the level 10^{-9}, we identified 12, 11, and 3 genes for colon, leukemia, and prostate data, respectively. Surprisingly, for the three datasets, all the genes contained in the MAP models can be found at an FDR level of 10^{-9}.

For comparison, the penalized likelihood methods, Lasso, elastic net, SIS, and ISIS, were also applied to these data with the results summarized in Table 8. The numerical results indicate that BSR performed much better than the penalized likelihood methods for these datasets. The models selected by BSR tend to be sparser and, more importantly, of lower cross-validation error. Figure 6 compares the 1-CVMRs of the models selected by Lasso along its regularization paths and the minimum 1-CVMRs of the MAP_i models sampled by BSR. The comparison indicates that BSR consistently worked better than Lasso for all three examples.

7. COMPUTATION TIME

Table 9 summarizes the CPU time of BSR for the four gene expression data examples at the selected values of γ. The settings of BSR have been described previously and the CPU times were measured on a Dell desktop of 3.0 GHz. It is easy to see that the CPU times are in an acceptable range for all the examples, given the cost of data collection.

After adjusting for the number of iterations, a simple linear regression for the CPU time versus $[E_{D^n}(|\xi_n|)]^3$ has an R^2 of 97.9%. A cubic transformation for $E_{D^n}(|\xi_n|)$ is used here, because the CPU time for inverting a matrix is known in the cubic order of the matrix size. We have also tried other examples (not reported in the article). Including their CPU time with those in Table 9, a linear regression for the CPU time versus n, P_n, and $[E_{D^n}(|\xi_n|)]^3$ indicates that the CPU time depends mainly on $[E_{D^n}(|\xi_n|)]^3$, followed by n and P_n in order of significance. This analysis implies that BSR can be applied to datasets with an extremely large P_n, as the true model size usually does not grow much with P_n.

For problems where $E_{D^n}(|\xi_n|)$ is large, the computation can be accelerated by table-listing the sampled models, taking advantage of BSR for which the posterior distribution of regression coefficients is degenerated. This can avoid repeated parameter estimation for the models that have been previously sampled. Since parameter estimation has been a dominating factor of computation for these problems, the CPU time saved with the table-listing approach can be substantial.

8. DISCUSSION

In this article, we have proposed a new prior setting for GLMs, which leads to a BSR with the MAP model approximately equivalent to the minimum EBIC model. Under mild conditions, we have established consistency of the posterior based on the result of Jiang (2007). Further, we have proposed a variable screening procedure based on the marginal inclusion probability and showed that the proposed procedure shares the same theoretical properties of sure screening and consistency as the SIS procedure proposed by Fan and Song (2010). Since the proposed

Table 8. Comparison of BSR, Lasso, elastic net, SIS, and ISIS for three gene expression data. [a]Minimum 1-CVMR (r_{cv}%) and the corresponding model size ($|\xi_n|$); [b]1-CVMR (r_{cv}%) and the corresponding model size ($|\xi_n|$); [c]zero 1-CVMR models among MAP_i's sampled by SAMC

									BSR			
	Lasso[a]		Elastic net[a]		SIS[b]		ISIS[b]		MAP		MAP^c_*	
Dataset	$\lvert\xi_n\rvert$	r_{cv}%	$\lvert\xi_n\rvert$	r_{cv}%	$\lvert\xi_n\rvert$	r_{cv}%	$\lvert\xi_n\rvert$	r_{cv}%	$\lvert\xi_n\rvert$	r_{cv}%	$\lvert\xi_n\rvert$	r_{cv}%
Colon	15	12.9	26	11.3	3	12.9	3	12.9	3	4.84	4–7,9	0
Leukemia	16	4.17	37	2.78	4	4.17	2	0	2	0	2–10	0
Prostate	3	6.86	42	5.88	5	6.86	4	4.90	3	3.92	4–9	0

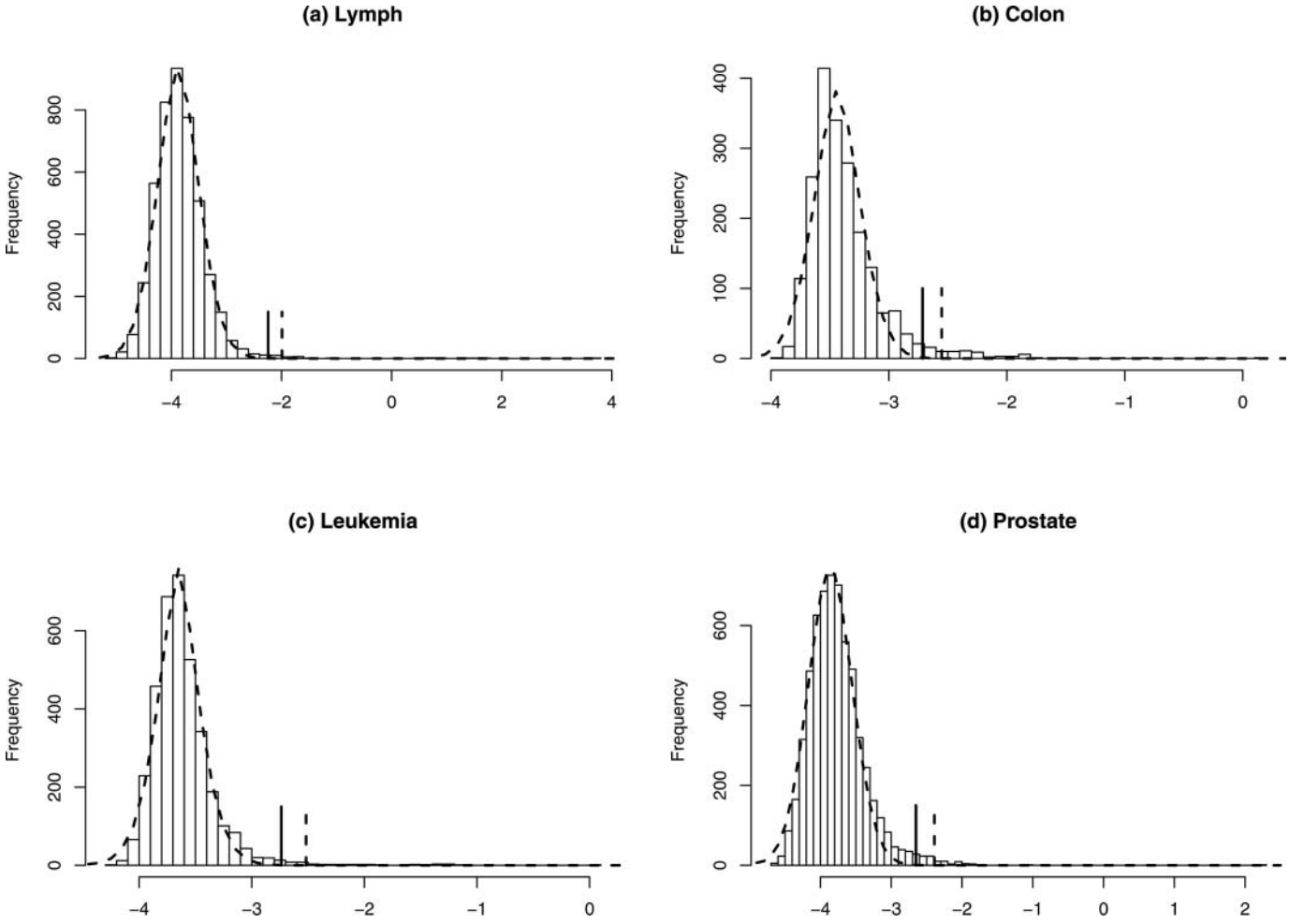

Figure 5. Illustrative plots of SVS for (a) lymph data, (b) colon data, (c) leukemia data, and (d) prostate data. The dashed curve shows the fitted density for the MIS of nonrelevant genes, and the vertical bars show the classification rules at the FDR level 0.01 (solid line) and the FDR level 0.001 (dashed line).

procedure makes use of the joint information of all predictors, it generally outperforms SIS and its iterative extension ISIS.

For Bayesian variable selection, this article has involved three different types of consistency, posterior consistency, global model consistency, and pairwise model consistency. As suggested by Theorem 3.1, the global model consistency is the strongest one. The posterior consistency together with the identifiability of the true model implies the global model consistency, and the global model consistency leads directly to the pairwise consistency of the MAP model. As previously noted, from the point of view of Bayesian model averaging, choosing priors that lead to the global model consistency is of great interest. Otherwise, the number of models that need to be averaged for valid Bayesian inference may increase exponentially with P_n. In this article, we have provided a simple but general prior setting for a wide class of GLMs that leads to the global model consistency.

Table 9. Summary of CPU times of BSR: the CPU times are measured (in hours) on a Dell desktop of 3.0 GHz and 8 GB RAM, and $E_{D^n}(|\xi_n|)$ denotes the posterior expectation of $|\xi_n|$

Dataset	n	P_n	γ	Iterations ($\times 10^6$)	$E_{D^n}(\lvert\xi_n\rvert)$	CPU (hr)
Lymph	148	4514	0.85	10.05	8.5	12.7
Colon	62	2000	0.99	5.05	3.2	1.0
Leukemia	72	3571	0.99	5.05	2.5	1.7
Prostate	102	6033	0.99	5.05	4.1	1.5

In this article, we have also made extensive comparisons of BSR with the popular penalized likelihood methods, including Lasso, elastic net, SIS, and ISIS. Through a comparison study conducted on the subset and full data, we have found that the performance of the penalized likelihood methods tend to deteriorate with the dimension P_n. Given the stable performance of BSR on the subset and full data, we conclude that BSR is more suitable than these penalized likelihood methods for high-dimensional variable selection, although the latter are computationally more attractive. Our further results on the simulated data and four gene expression datasets make this conclusion more convincing, see, for example, Tables 4 and 8.

Although our numerical examples are all for logistic regression, Lemma 2.1, and Theorem 3.1 hold for other GLMs with constant dispersion, which include probit, Poisson, exponential, and normal linear regressions (with constant variance). A further

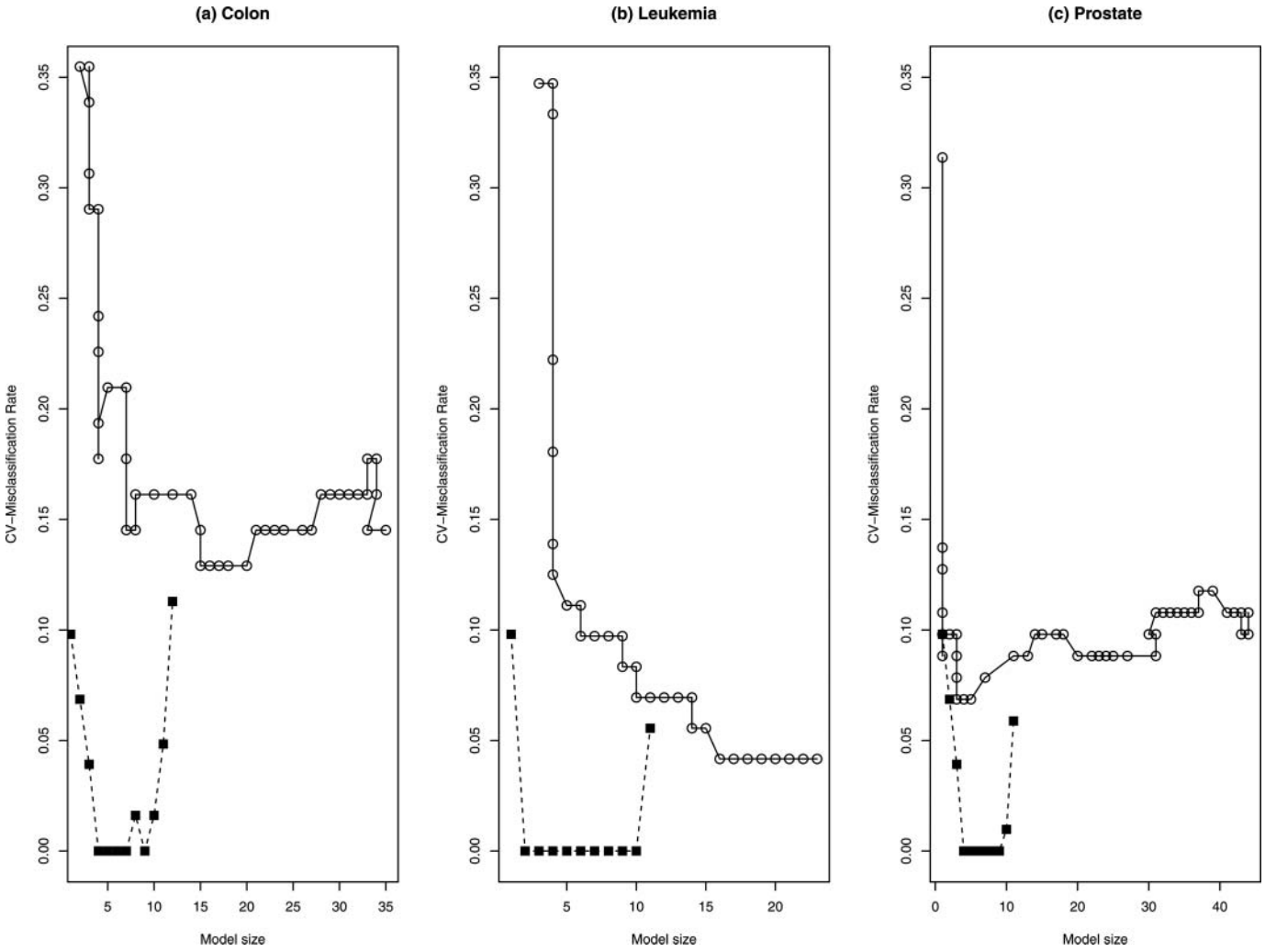

Figure 6. Comparison of 1-CVMRs produced by Lasso (while circles connected by solid line) and BSR (black squares connected by dotted line) for (a) colon data, (b) leukemia data, and (c) prostate data.

extension of this study to the GLMs that include a dispersion parameter would be of great interest.

Finally, we note that the theoretical properties and computational advantages of the proposed method are all established under the assumption that the sample size is large. When the sample size is small, one may need to sample from the exact posterior distribution instead of the approximated one as derived in Section 2.1. In this scenario, one may also consider to switch the prior distribution (5) to another one that accounts for the possible correlation between covariates and/or facilitates the simulation from the posterior, for example, the hyper-g prior (Bové and Held 2011) or the conjugate prior (Chen and Ibrahim 2003; Chen et al. 2008). A further study of this problem, that is, how to make variable selection for high-dimensional regression with a small sample size, would be of great interest.

APPENDIX: THEORETICAL PROOFS

A.1 Proof of Lemma 3.1

We note that the Hellinger distance is bounded above by $\sqrt{2}$, that the squared Hellinger distance is convex in both its arguments, and that $d^2(e_j, e_{j|\xi}) = 2|e_j - e_{j|\xi}|$, as both e_j and $e_{j|\xi}$ are binary variables. Here, $(1 - e_j, e_j)$ and $(1 - e_{j|\xi}, e_{j|\xi})$ are viewed as two distributions defined on the space $\{0, 1\}$. Then

$$\begin{aligned} d^2(q_j, e_j) &= d^2\left(\sum_{\xi} e_{j|\xi}\Pi(\xi|D^n), e_j\right) \le \sum_{\xi} d^2(e_{j|\xi}, e_j)\Pi(\xi|D^n) \\ &= \sum_{\xi\in A_{\epsilon_n}} d^2(e_{j|\xi}, e_j)\Pi(\xi|D^n) + \sum_{\xi\in A^c_{\epsilon_n}} d^2(e_{j|\xi}, e_j)\Pi(\xi|D^n) \\ &\le 2\rho_j(\epsilon_n) + 2\Pi\left(A^c_{\epsilon_n}\middle|D^n\right). \end{aligned}$$

Let $\delta'_n = \Pi(A^c_{\epsilon_n}|D^n)$. For any $\delta_n > 0$ and for sufficiently large n,

$$P[d^2(q_j, e_j) \le 2\delta_n + 2\delta'_n] \ge P[\Pi[d(\hat{f}, f) > \epsilon_n|D^n] \le \delta'_n].$$

Taking $\delta'_n = e^{-0.5cn\epsilon_n^2}$, by Lemma 2.1, we have

$$P\left[d^2(q_j, e_j) \le 2\delta_n + 2e^{-0.5cn\epsilon_n^2}\right] \ge 1 - e^{-0.5cn\epsilon_n^2},$$

which completes the proof.

A.2 Proof of Theorem 3.1

Let $\delta(q_j, e_j)$ denote the total variation distance between the two distributions $(1 - q_j, q_j)$ and $(1 - e_j, e_j)$. It is easy to see that $\delta(q_j, e_j) = |q_j - e_j|$. By the inequality of Hellinger distance versus total variation distance,

$$|q_j - e_j| \le \sqrt{2}d(q_j, e_j).$$

By Lemma 3.1, we have

$$P\left[|q_j - e_j| > 2\sqrt{\delta_n + e^{-0.5cn\epsilon_n^2}}\right] \le e^{-0.5cn\epsilon_n^2}, \quad j = 1, 2, \ldots, P_n.$$

Then part-(i) follows from Boole's inequality of probability.

Since for all $\boldsymbol{x}_j \in \xi_*$, we have $e_j = 1$ and 0 otherwise, then by part-(i), we have

$$q_j \ge \hat{q} \quad \text{for all } \boldsymbol{x}_j \in \xi_* \text{ and } \quad q_j < \hat{q} \quad \text{for all } \boldsymbol{x}_j \notin \xi_*,$$

in probability 1 when n is sufficiently large.

Define $B_n = \{\max_{\boldsymbol{x}_j \in \xi_*} q_j \ge \hat{q}\}$ and $E_n = \{\max_{\boldsymbol{x}_j \notin \xi_*} q_j < \hat{q}\}$. By Boole's inequality,

$$P\left(B_n^c\right) \le s_n e^{-0.5cn\epsilon_n^2} \quad \text{and} \quad P\left(E_n^c\right) \le (P_n - s_n)e^{-0.5cn\epsilon_n^2},$$

for sufficiently large n. The former implies the sure screening property, that is, part-(ii),

$$P(\xi_* \subset \widehat{\xi}_{\hat{q}}) \ge 1 - s_n e^{-0.5cn\epsilon_n^2},$$

and the latter implies the vanishing of false selection rates,

$$P\left(\xi_*^c \subset \widehat{\xi}_{\hat{q}}^c\right) = P(\widehat{\xi}_{\hat{q}} \subset \xi_*) \ge 1 - (P_n - s_n)e^{-0.5cn\epsilon_n^2}.$$

By Bonferroni's inequality, we have

$$P(\xi_* = \widehat{\xi}_{\hat{q}}) \ge 1 - P_n e^{-0.5cn\epsilon_n^2}, \tag{27}$$

which implies part-(iii) when $\hat{q}$ is set to 0.5.

SUPPLEMENTARY MATERIALS

Part I: The SAMC algorithm.

Part II: Selected features of the subset data studied in Section 4.

[Received January 2012. Revised November 2012.]

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Mack, S., and Levine, J. (1999), "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 USA*, 96, 6745–6750. [599]

Ando, T. (2010), *Bayesian Model Selection and Statistical Modeling*, New York: Chapman & Hall. [591]

Bae, K., and Mallick, B. K. (2004), "Gene Selection Using a Two-Level Hierarchical Bayesian Model," *Bioinformatics*, 20, 3423–3430. [590]

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [594]

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006), "Adaptive Linear Step-up Procedures That Control the False Discovery Rate," *Biometrika*, 93, 491–507. [596]

Bottolo, L., and Richardson, S. (2010), "Evolutionary Stochastic Search for Bayesian Model Exploration," *Bayesian Analysis*, 5, 583–618. [590]

Bové, D. S., and Held, L. (2011), "Hyper-*g* Priors for Generalized Linear Models," *Bayesian Analysis*, 6, 387–410. [593,604]

Broman, K. W., and Speed, T. P. (2002), "A Model Selection Approach for the Identification of Quantitative Trait Loci in Experimental Crosses," *Journal of The Royal Statistical Society,* Series B, 64, 641–656. [590]

Chapman, J. M., Onnie, C. M., Prescott, N. J., Fisher, S. A., Mansfield, J. C., Mathew, C. G., Lewis, C. M., Verzilli, C. J., and Whittaker, J. C. (2009), "Searching for Genotype-Phenotype Structure: Using Hierarchical Log-linear Models in Crohn Disease," *American Journal of Human Genetics*, 84, 178–187. [594]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection With Large Model Space," *Biometrika*, 94, 759–771. [590]

——— (2012), "Extended BIC for Small-*n*-Large-*P* Sparse GLM," *Statistica Sinica*, 22, 555–574. [590,598]

Chen, M.-H., and Ibrahim, J. G. (2003), "Conjugate Priors for Generalized Linear Models," *Statistica Sinica*, 13, 461–476. [604]

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. [604]

Chib, S. (1995), "Marginal Likelihood From Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321. [592]

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010), "R Package: Sure Independence Screening," available at *http://cran.r-project.org/web/packages/SIS*. [597]

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. [589]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of The Royal Statistical Society,* Series B, 70, 849–911. [590]

——— (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [589]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultra-Dimensional Variable Selection via Independent Learning: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [590]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Model With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [590,595,602]

Figueiredo, M. A. T. (2003), "Adaptive Sparseness for Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159. [590]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [596]

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Collerm, H., Loh, M. L., Downing, J. P., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. [599]

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242. [593]

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for Regression With Many Candidate Predictors," *Journal of the American Statistical Association*, 102, 507–516. [590,593,599,600]

Holzmann, H., Munk, A., and Gneiting, T. (2006), "Identifiability of Finite Mixtures of Elliptical Distributions," *Scandinavian Journal of Statistics*, 33, 753–763. [595]

Hsu, S. J. (1995), "Generalized Laplace Approximation in Bayesian Inference," *Canadian Journal of Statistics*, 23, 399–410. [590]

Jiang, W. (2006), "On the Consistency of Bayesian Variable Selection for High Dimensional Binary Regression and Classification," *Neural Computation*, 18, 2762–2776. [590]

——— (2007), "Bayesian Variable Selection for High Dimensional Generalized Linear Models: Convergence Rates of the Fitted Densities," *The Annals of Statistics*, 35, 1487–1511. [590,592,602]

Johnson, V. E., and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660. [594,595]

Liang, F. (2009), "On the Use of Stochastic Approximation Monte Carlo for Monte Carlo Integration," *Statistics & Probability Letters*, 79, 581–587. [593]

Liang, F., Liu, C., and Carroll, R. J. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320. [593]

Liang, F., Truong, Y. K., and Wong, W. H. (2001), "Automatic Bayesian Model Averaging for Linear Regression and Applications in Bayesian Curve Fitting," *Statistica Sinica*, 11, 1005–1029. [590]

Liang, F., and Zhang, J. (2008), "Estimating FDR Under General Dependence Using Stochastic Approximation," *Biometrika*, 95, 961–977. [595]

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–676. [590]

Moreno, E., Girón, F. J., and Casella, G. (2010), "Consistency of Objective Bayes Factors as the Model Dimension Grows," *The Annals of Statistics*, 38, 1937–1952. [594]

Park, M. Y., and Hastie, T. (2007), "L_1-Regularization Path Algorithm for Generalized Linear Models," *Journal of The Royal Statistical Society,* Series B, 69, 659–677. [597]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [590]

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010), "Bayesian Models for Sparse Regression Analysis of High Dimensional Data," *Bayesian Statistics*, 9, 539–568. [590]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [593]

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [599]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of The Royal Statistical Society,* Series B, 64, 479–498. [595]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of The Royal Statistical Society,* Series B, 58, 267–288. [589,590]

Zhang, C.-H. (2009), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [589]

Zhang, J., and Liang, F. (2008), "Convergence of Stochastic Approximation Under Irregular Conditions," *Statistica Neerlandica*, 62, 393–403. [595]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of The Royal Statistical Society,* Series B, 67, 301–320. [589]