Statistics and Its Interface Volume 11 (2018) 385–395

Sparse Bayesian variable selection for classifying high-dimensional data

Aijun Yang, Heng Lian, Xuejun Jiang*, and Pengfei Liu

Identifying differentially expressed genes for classifying experiment classes is an important application of microarrays. Methods for selecting important genes are of much significance in accurate classification. Owing to the large number of genes and many of them are irrelevant, insignificant or redundant, standard statistical methods do not work well. The modification of existing methods is needed to achieve better analysis of microarray data. We present a stochastic variable selection approach for gene selection with different two level hierarchical prior distributions for regression coefficients. These priors can be used as a sparsity-enforcing mechanism to perform gene selection for classification. Using simulation-based MCMC methods for simulating parameters from the posterior distribution, an efficient algorithm is developed and implemented. This algorithm is robust to the choices of initial values, and produces posterior probabilities of related genes for biological interpretation. To highlight the potential applications of the proposed approach, we provide examples of the well-known colon cancer data and leukemia data in microarray literature.

Keywords and phrases: Sparse priors, Stochastic variable selection, Classification, High-dimensional data.

1. INTRODUCTION

With the development of microarray technology, researchers can rapidly measure the levels of thousands of genes expressed in a single experiment. One important application of this microarray technology is to classify the samples into different diagnostic categories using their gene expression profiles. One current difficulty is that the microarray data often consist of a large number of genes compared to the number of samples. Some genes could be related to a particular type of diagnostic category. However, many of the genes are irrelevant or redundant and affect the accuracy of classification. Therefore, robust and accurate gene selection methods are required because effective gene selection methods often lead to a compact classifier with better interpretability and accuracy.

Gene selection problem basically can be treated as a variable selection problem associated with a linear regression models problem in statistics. Most of the proposed methods in the literature are univariate methods and have following fundamental disadvantages: Firstly these methods do not take the correlations between genes into account. Consequently, some insignificant genes are selected while some useful but weakly significant genes may be omitted. Secondly these methods are not probabilistic models, thus can not produce the inclusion probabilities for the selected genes, which are helpful for achieving better biological interpretation. Thirdly, the classification procedure of these methods consists of two steps. In the first step, standard techniques such as t or F tests are used to select some significant genes. In the second step, those selected genes are used to fit a classification model for cancer classification. However, as the classification models play no part in the initial gene selection, this classification procedure often results in accumulated errors in the final class prediction (Dougherty, 2001).

Bayesian multivariate methods, which can take into account the correlations among genes, are proposed in the literature for developing probabilistic models for the purpose of gene selection and cancer classification. Recently some Bayesian formulations of neural networks and support vector machine (SVM) have been proposed (Chakraborty et al., 2004; Mallick et al., 2005) for cancer classification. However, these methods have some limitations, such as, they cannot self sufficiently select the significant genes, and their performance is often highly dependent on an efficient gene selection prior to model fitting. By using the stochastic search technique and SVM, Chakraborty et al. (2007) and Chakraborty (2009) developed a Bayesian kernal logistic model for multiclass data and a Bayesian kernal probit model for binary data, respectively. These two methods both try to conduct gene selection and class prediction simultaneously. However, nonlinear kernels (polynomial and Gaussian kernel) are required in using such Bayesian SVM methods. Therefore, it is not straightforward to interpret the direct relationship between the genes and the cancer types.

In some Bayesian literature, Bayesian probit/multinomial probit regression models (Lee et al., 2003; Zhou et al., 2004a; Zhou et al., 2004b; Sha et al., 2004) and Bayesian logistic regression model (Zhou et al., 2004c) are proposed for classification purposes. In such models, the gene selection can be automatically performed by indexing the genes of the models. As a linear model is used by these approaches to establish the relationship between the genes and the cancer types, how the genes finally explain the tumor behavior can be tracked down. However, they adopted the g-prior (Zell-

*Corresponding author.

ner, 1986) for unknown parameters of regression coefficients. For situations with high-dimensional genes or even a small set of genes, there exists a possibility of multicollinearity, then the covariance matrix involved in the g-prior is nearly singular (Gupta and Ibrahim, 2007). Taking such g-prior may lead to the collapse of the MCMC algorithm and other convergence problems (Yang and Song, 2010).

Alternatively, sparse methods have been proposed for gene selection and classification. These methods assume that only a subset (which is often considered small) of genes has significant effect on the cancer types, and the other subset of genes which have little or no effect can be eliminated so as to better estimate the significant genes. Sparse methods are preferable as they can lead to a better outcome of sample prediction using fewer genes. Many such methods have been developed to improve separation of significant genes from insignificant genes. Sparse Bayesian methods that used heavy-tailed priors for the regression coefficients encourage a large proportion of those coefficients to be shrunk to a value close to zero. The degree of sparseness of the methods can be adjusted by changing the prior distribution of the regression coefficients. Many prior have been studied including: the student t (Bae and Mallick, 2004), the double exponential (Park and Casella, 2008)(leading to Bayesian LASSO) and the elastic net (Li and Lin, 2010)(leading to Bayesian elastic net). However, these methods still have some disadvantages: (a) perform only shrinkage of the regression coefficients towards zero but do not automatically implement variable selection; (b) just select a single model for classification, but do not take into account the model uncertainty which is especially important if prediction is the main objection. More recently, Chakraborty and Guo (2011) suggested to use Bayesian hybrid Huberized SVM (BHHSVM) with elastic net prior for the regression coefficients for gene selection and classification simultaneously. But BHHSVM is implementationally more complicated and computationally slower than Bayesian probit regression (Mallick et al., 2011).

Currently some variable selection techniques, such as Gibbs Variable Selection (GVS), Reversible Jump MCMC (RJMCMC; Green, 1995) and Stochastic Search Variable Selection (SSVS; George and McCulloch, 1993), can be incorporated into Bayesian sparse methods to do variable selection. GVS has the advantage that the posterior distribution is not affected by pseudo-priors, but it needs pseudo-priors on all regression coefficients of the model. The merit of RJMCMC is that the specification for pseudo-priors is not required, and the number of variables selected at each iteration is assumed to be a random variable; whereas diffuse priors will often lead to the fewest parameter model being chosen. The advantage of SSVS is that it can be applied to a wide variety of models, and the users are allowed to indicate which models they think are more likely.

In this paper, we propose an integrated sparse Bayesian variable selection method for classification using sparse Bayesian method and SSVS technique. The gene selection is conducted by indexing the genes of the model under this method. Compared with the methods discussed above, the novelty of our method may be summarized as follows: (a) sparse Bayesian variable selection is the big novelty of our method. (b) our method can take model uncertainty into account, and the importance of genes are measured by calculating the relative frequency of each gene selected through the MCMC method. (c) the relative frequencies of genes are sparser than that in Bayesian probit/logistic regression models, then it is helpful for us to select significant genes. (d) by rewriting the heavy-tailed priors as a two level hierarchical model, an efficient MCMC algorithm is designed to visit models of any size.

For gene selection and classification of diagnostic category, we consider a multivariate Bayesian regression model with two-level hierarchical (TH) Bayesian framework and a stochastic search variable selection (SSVS) method. Moreover, unlike the method based on approximation, we perform full Bayesian analysis through the Markov chain Monte Carlo (MCMC; Gilks et al., 1996) based stochastic search algorithm. In developing our TH-SSVS algorithm, an efficient sampling scheme is implemented. In addition, the TH-SSVS approach produces the posterior probabilities for the selected genes, which is helpful for achieving better biological interpretation. We illustrate the advantage of our method on two well-known microarray data sets: Colon cancer data (Alon et al., 1999) and Acute leukemia data (Golub et al., 1999), which have been extensively used in the literature to demonstrate various classification procedures (Nguyen and Rocke, 2002; Le Cao and Chabrier, 2008; among others). Our results show that the proposed TH-SSVS approach reduced the number of genes selected and produced prediction accuracy comparable to that of the existing variable selection and classification methods.

The remainder of the paper is structured as follows. In Section 2, we briefly review the statistical model and describe hierarchical prior distributions for variable selection; we also give details of the Bayesian analysis of the posterior distribution, including a discussion of efficient sampling scheme, and discuss the classification in this section. Section 3 illustrates the performance of the method for two publicly available data sets. In section 4 we apply our method on one simulated data set. Section 5 gives some discussions and conclusions.

2. METHOD

2.1 Probit model

Suppose the data set has n observations with p predictors. Let $Y = (Y_1, \cdots, Y_n)$ denote the observed binary responses. For example, Y_i=1 indicates that sample i is normal or one type of cancer and Y_i=0 indicates that sample i is cancer or another type of cancer. For each sample i, let x_{ij} be the measurement of the expression level of the j-th gene for the i-th sample; hence we have the following data matrix $\mathbf{X}$ of covariates:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

We model the dependence of Y_i on X_i as $p_i = P(Y_i = 1) = \Phi(\alpha + X_i\beta)$, where α represents the intercept, and $\beta = (\beta_1, \cdots, \beta_p)^T$ is a p by 1 dimensional vector of regression coefficients, X_i is the i-th row of $\mathbf{X}$, and Φ is the standard normal cumulative distribution function relating p_i with $\alpha + X_i\beta$. We follow Albert and Chib (1993) and augment Y_i with one latent variables Z_i to convert the probit model to a regression model with inequality constraints on the latent variables. More specifically, we define

$$Z_i = \alpha + X_i\beta + \varepsilon_i, \tag{1}$$

where the disturbance or noise term ε_i are independently and identically distributed as $N(0, 1)$. The relationship between Y_i and Z_i is

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0, \\ 0 & \text{if } Z_i \le 0. \end{cases} \tag{2}$$

In order to index the possible subsets of genes for performing gene selection, we introduce a latent binary vector $\gamma = (\gamma_1, \cdots, \gamma_p)$, such that

$$\gamma_i = \begin{cases} 1 & \text{if the } i\text{-th gene is included in the model,} \\ 0 & \text{if the } i\text{-th gene is excluded from the model.} \end{cases} \tag{3}$$

This indicator is used to induce a mixture prior on the regression coefficients.

Given γ, let $p_\gamma = \sum_{i=1}^{p} \gamma_i$, β_γ be a p_γ by 1 vector consisting of all the nonzero elements of β, and $\mathbf{X}_\gamma$ be an n by p_γ matrix of covaraites consisting of all the columns of $\mathbf{X}$ corresponding to those elements of γ that are equal to 1. Adopting these notations, model (1) can be rewritten as

$$Z_i = \alpha + \mathbf{X}_{i,\gamma}\beta_\gamma + \varepsilon_i, \tag{4}$$

where $\mathbf{X}_{i,\gamma}$ is the i-th row of $\mathbf{X}_\gamma$.

2.2 Prior specification

The choice of the prior distributions for the unknown parameters is very important in the Bayesian SSVS approach. In this paper, prior distributions for α, β_γ, and γ with the structure $p(\alpha, \beta_\gamma, \gamma) = p(\alpha)p(\beta_\gamma|\gamma)p(\gamma)$ is considered.

The prior distribution of α is taken as

$$\alpha \sim N(0, h), \tag{5}$$

where h is a hyperparameter representing the variance of the univariate normal distribution. Since α is not our focus, a specified value is assigned to h. According to Lamnisos et al. (2009), a large value of h is taken.

For more crucial regression coefficient parameter β, we consider sparse priors in this paper. Sparse priors play an important role in Bayesian regression modeling, and has been shown to be useful in a more general problem of learning a sparse model in high-dimensional space (Wainwright et al., 2006). In contrast to a prior assumption of independently and normally distributed coeficients sharing a common variance, sparse priors are heavy tailed and peaked at zero, and can better accommodate large regression coefficients. Two particular sparse priors are student t and Laplacian distributions. In regression problems, study and use of the Laplacian prior distribution have become popular in part due to its connections to the Lasso procedure of Tibshirani (1996). However, the variable selection property is ad hoc from a Bayesian perspective. Under the absolutely continuous student t or Laplacian prior distribution, the prior probability of the event $\beta_i = 0$ is zero, and so the posterior probability of such an event must also be zero. In order for posterior inferences about events such as $\beta_i = 0$ to be coherent, prior probability mass must be allocated to these events. By the definition of γ_i, if $\gamma_i = 0$, the i-th gene is excluded from the model, it is natural to force $\beta_i = 0$, and if $\gamma_i = 1$, we assign a student t or Laplacian prior for β_i. Within the class of sparse priors for β_i, scale mixtures of normal distributions have received extensive attention. Therefore, the student t prior or Laplacian prior can be presented as a two level hierarchical model. The complete hierarchical probability distribution for β_i given γ_i are given below.

At the first level, the prior distribution of regression coefficient β_i given γ_i is assumed to be

$$p(\beta_i|\gamma_i) = (1 - \gamma_i)\delta(0) + \gamma_i N(0, \lambda_i), \tag{6}$$

where $\delta(0)$ is a point mass at 0, λ_i is the variance of β_i when γ_i is equal to one.

At the second level, we assume two different prior distributions for λ_i

Model I: $\lambda_i \sim \text{IG}(\frac{a}{2}, \frac{2}{b})$, where $\text{IG}(\frac{a}{2}, \frac{2}{b})$ denotes an inverse gamma distribution, and a and b are hyperparameters with the density function proportional to $u^{-(\frac{a}{2}+1)}\exp(-\frac{b}{2u}), u > 0$.

Model II: $\lambda_i \sim \text{Ga}(1, \frac{\tau}{2})$, where $\text{Ga}(1, \frac{\tau}{2})$ has the density function $\frac{\tau}{2}\exp(-\frac{\tau u}{2}), u > 0$, where τ is a hyperparameter.

For the prior specification on γ, a widely used prior is

$$p(\gamma) = \prod_{i=1}^{p} \theta_i^{\gamma_i}(1 - \theta_i)^{1-\gamma_i}, \quad 0 \le \theta_i \le 1, \tag{7}$$

that is $p(\gamma_i = 1) = \theta_i, i = 1, \cdots, p$. This prior assumes that the i-th gene is included in the model independently with a prior probability θ_i.

2.3 Computation

Denote $Z = (Z_1, \cdots, Z_n)^T, \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$. Under the model and prior specifications in the above sections, the joint posterior distribution under Model I or Model II is given by

$$p(Z, \alpha, \beta_\gamma, \mathbf{\Lambda}, \gamma|Y, \mathbf{X}) \tag{8}$$

$$\propto \exp\{-\frac{\sum_{i=1}^{n}(Z_i-\alpha-X_{i,\gamma}\beta_\gamma)^2}{2}\}\prod_{i=1}^{n}\mathrm{I}(A_i)$$
$$\times \exp(-\frac{\alpha^2}{2h})\times\prod_{i\in m(\gamma)}\lambda_i^{-\frac{1}{2}}\exp(-\frac{(\beta_i|\gamma_i)^2}{2\lambda_i})$$
$$\times\prod_{i=1}^{p}\theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}$$
$$\times\prod_{i=1}^{p}\lambda_i^{-(\frac{a}{2}+1)}\exp(-\frac{b}{2\lambda_i})\Big\{\text{or}\quad\exp(-\frac{\tau\lambda_i}{2})\Big\},$$

where A_i is equal to either $\{Z_i : Z_i > 0\}$ or $\{Z_i : Z_i \le 0\}$ corresponding to $Y_i = 1$ or $Y_i = 0$, respectively; $\mathrm{I}(\cdot)$ is an indicator function; and $m(\gamma)$ is the subscript set of those elements of γ that are equal to 1.

The posterior distribution in (8) cannot be expressed in an explicit form; therefore, we use an MCMC technique, namely the Gibbs sampler (Geman and Geman, 1984), to generate observations from this posterior distribution. Because α is rarely of interest, we marginalize it out for the purpose of simplicity and speed (Park and Casella, 2008). To make the sampling scheme efficiently explore the space of 2^p variables, we jointly update correlated components to improve the results.We can in turn update Z, β_γ, Λ and γ based on $p(Z, \Lambda|\mathbf{X}, Y, \beta, \gamma) \propto p(Z|\mathbf{X}, Y, \Lambda, \gamma)p(\Lambda|\beta, \gamma)$ and $p(\beta_\gamma, \gamma|\mathbf{X}, Z, \Lambda) \propto p(\beta_\gamma|\mathbf{X}, Z, \Lambda, \gamma)p(\gamma|\mathbf{X}, Z, \Lambda)$.The conditional distributions for implementing our sampling scheme are given below:

$(i)p(Z|\mathbf{X}, Y, \Lambda, \gamma)$: It can be shown that

$$p(Z|\mathbf{X}, Y, \Lambda, \gamma) \propto N(0, \Sigma_\gamma)\prod_{i=1}^{n}\mathrm{I}(A_i), \tag{9}$$

with $\Sigma_\gamma = h1_n1_n^T + \mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^T + \mathbf{I}_n$, which is a multivariate truncated normal distribution. In (9), β is marginalized out from the posterior distribution $p(Z|\mathbf{X}, Y, \beta, \Lambda, \gamma)$ to reduce autocorrelation between β and Z, thus to improve mixing in the Markov chain. Direct sampling from (9) is known to be difficult. We follow the method of Devroye (1986) to simulate samples from the univariate truncated normal distribution $p(Z_i|Z_{(-i)}, \mathbf{X}, Y, \Lambda, \gamma)$, where $Z_{(-i)}$ is the vector of Z without the i-th element.

$(ii)p(\Lambda|\beta, \gamma)$: The posterior distribution of the i-th diagonal element of Λ under Model I is

$$\lambda_i|\beta_i, \gamma_i \sim \mathrm{IG}(\frac{a+1}{2}, \frac{2}{b+\beta_i^2}). \tag{10}$$

The posterior distribution of λ_i under Model II is

$$\lambda_i^{-1}|\beta_i, \gamma_i \sim \mathrm{InvGauss}(\frac{\sqrt{\tau}}{|\beta_i|}, \tau), \tag{11}$$

where InvGauss denotes the inverse Gaussian distribution with the probability density function

$$\mathrm{InvGauss}(\iota, \kappa) = \sqrt{\frac{\kappa}{2\pi u^3}}\exp\{-\frac{\kappa(u-\iota)^2}{2\iota^2 u}\}, u > 0. \tag{12}$$

We use the algorithm given in Chhikara and Folks (1989) to generate the random observations from the inverse Gaussian distribution.

$(iii)p(\beta_\gamma|\mathbf{X}, Z, \Lambda, \gamma)$: The full conditional distribution of β_γ is

$$\beta_\gamma|\mathbf{X}, Z, \Lambda, \gamma \sim N(\Omega_\gamma\mathbf{X}_\gamma^T\Phi Z, \Omega_\gamma), \tag{13}$$

where $\Phi = (h1_n1_n^T + \mathbf{I}_n)^{-1}$, and $\Omega_\gamma = (\mathbf{X}_\gamma^T\Phi\mathbf{X}_\gamma + \Lambda_\gamma^{-1})^{-1} = \Lambda_\gamma - \Lambda_\gamma\mathbf{X}_\gamma^T\Phi(\Phi\mathbf{X}_\gamma\Lambda_\gamma\mathbf{X}_\gamma^T\Phi + \Phi)^{-1}\Phi\mathbf{X}_\gamma\Lambda_\gamma$. The matrix inversion for calculating Ω_γ is computed using the well-known Sherman-Morrison-Woodbury formula, which can make the computation much faster when data are high-dimensional with small sample size.

$(iv)p(\gamma|\mathbf{X}, Z, \Lambda)$: This conditional distribution is proportional to $|\Sigma_\gamma|^{-\frac{1}{2}}\exp(-\frac{Z^T\Sigma_\gamma^{-1}Z}{2})\prod_{i=1}^{p}\theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}$. We marginalize out β from the conditional distribution $p(\gamma|\mathbf{X}, Z, \beta, \Lambda)$ so that the Markov chain would be nonreducible (Panagiotelisa and Smith, 2008). For implementing an efficient sampling scheme, we draw a component γ_i of γ conditionally on $\gamma_{(-i)}$, where $\gamma_{(-i)}$ is the vector of γ without the i-th element, and

$$p(\gamma_i|\gamma_{(-i)}, \mathbf{X}, Z, \Lambda) \propto |\Sigma_\gamma|^{-\frac{1}{2}}\exp(-\frac{Z^T\Sigma_\gamma^{-1}Z}{2})\theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}. \tag{14}$$

Because γ_i is binary, we can get the conditional probabilities of $p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z, \Lambda)$ and $p(\gamma_i = 0|\gamma_{(-i)}, \mathbf{X}, Z, \Lambda)$. Denote $\gamma^1 = (\gamma_i, \cdots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \cdots, \gamma_p)$ and $\gamma^0 = (\gamma_i, \cdots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \cdots, \gamma_p)$, and similarly define Σ_{γ^1} and Σ_{γ^0} as Σ_γ in (9). It can be shown that:

$$p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z, \Lambda) = (1 + \frac{1-\theta_i}{\theta_i}\rho)^{-1}, \tag{15}$$

where

$$\rho = |\Sigma_{\gamma^1}\Sigma_{\gamma^0}^{-1}|^{\frac{1}{2}}\exp\{-\frac{Z^T(\Sigma_{\gamma^1}^{-1} - \Sigma_{\gamma^0}^{-1})Z}{2}\}. \tag{16}$$

As a result, an explicit form of the conditional distribution in (15) can be derived.

To implement the Gibbs sampler, we start with an initial value $(Z^{(0)}, \Lambda^{(0)}, \beta_\gamma^{(0)}, \gamma^{(0)})$, and continue as follows: at the $(k+1)$-th iteration with the k-th value $(Z^{(k)}, \Lambda^{(k)}, \beta_\gamma^{(k)}, \gamma^{(k)})$,

step (a): For $i = 1, \cdots, n$, draw $Z_i^{(k+1)}$ from the univariate truncated normal distribution $p(Z_i^{(k)}|Z_{(-i)}^{(k)}, \mathbf{X}, Y, \Lambda^{(k)}, \gamma^{(k)})$.

step (b): For $i = 1, \cdots, p$, if $\gamma_i = 1$ draw $\lambda_i^{(k+1)}$ from the conditional distribution (10) and (11) for Model I and Model II, respectively; if $\gamma_i = 0$, set $\lambda_i^{(k+1)} = \lambda_i^{(k)}$.

step (c): Draw $\beta_\gamma^{(k+1)}$ from the conditional distribution (13).

step (d): For $i = 1, \cdots, p$, generate a random number u_i from a uniform distribution $U[0,1]$, calculate the probability $p_i^{(k+1)} = p(\gamma_i^{(k+1)} = 1 | \gamma_{(-i)}^{(k)}, \mathbf{X}, Z^{(k+1)}, \Lambda^{(k+1)})$ via (15) and (16), and update γ_i as follows:

$$\gamma_i^{(k+1)} = \begin{cases} 1 & \text{if } p_i^{(k+1)} \le u_i, \\ 0 & \text{otherwise.} \end{cases}$$

Under mild regularity conditions and for sufficiently large T, $(Z^{(T)}, \Lambda^{(T)}, \beta_\gamma^{(T)}, \gamma^{(T)})$ simulated from the above Gibbs sampler can be regarded as an observation from the joint posterior distribution $p(Z, \beta_\gamma, \Lambda, \gamma | Y, \mathbf{X})$, see Geman and Geman (1984). We collect MCMC samplers $\{(Z^{(k)}, \Lambda^{(k)}, \beta_\gamma^{(k)}, \gamma^{(k)}), k = 1, \cdots, M\}$ after a suitable burn-in period. An initial value of $\gamma^{(0)}$ can be obtained by randomly selecting a small number of genes and assigning 1 to the corresponding entries of $\gamma^{(0)}$. In contrast, Bae and Mallick (2004) used two sample t statistic to identify a certain number of significant genes for getting $\gamma^{(0)}$. Our method seems more reasonable as we usually have little prior information about which genes are significant among the large number of genes. The MCMC algorithm in our method is robust to the choice of $\gamma^{(0)}$ and encounters no problem in convergence. Note also that the MCMC algorithm focuses on generating $(Z^{(k)}, \Lambda^{(k)}, \beta_\gamma^{(k)}, \gamma^{(k)})$, which is important and sufficient for gene selection and classification, while the less important α is not simulated. The relative frequency of each gene can be calculated as

$$\hat{p}(\gamma_i = 1 | \mathbf{X}, Y) = \frac{1}{M} \sum_{k=1}^{M} \gamma_i^{(k)}. \tag{17}$$

This gives an estimate of the posterior gene inclusion probability as a measure of the relative importance of the i-th gene. Genes with high posterior inclusion probabilities are relevant to classification.

2.4 Classification

The performance of a classification rule is best assessed by applying the rule created on the training set to the test set. If no test set is available, we use the sample based leave one out cross-validation (LOOCV) method (Gelfand, 1996). Let $Y_{(-i)}$ be the vector of Y without the i-th element. An LOOCV predictive probability for Y_i can be calculated as

$$\begin{aligned} p(Y_i | Y_{(-i)}, \mathbf{X}) &= \int p(Y_i | Y_{(-i)}, \mathbf{X}, Z, \beta, \mathbf{\Lambda}, \gamma) \\ &\quad \times p(Z, \beta, \mathbf{\Lambda}, \gamma | Y_{(-i)}, \mathbf{X}) dZ d\beta d\mathbf{\Lambda} d\gamma. \end{aligned} \tag{18}$$

If a test set Y_{new} is available, the predictive posterior probability of Y_{new} given the new covariate X_{new} is

$$\begin{aligned} p(Y_{\text{new}} | Y, X_{\text{new}}) &= \int p(Y_{\text{new}} | Y, X_{\text{new}}, Z, \beta, \mathbf{\Lambda}, \gamma) \\ &\quad \times p(Z, \beta, \mathbf{\Lambda}, \gamma | Y) dZ d\beta d\mathbf{\Lambda} d\gamma. \end{aligned} \tag{19}$$

This probability can be approximated by Monte Carlo integration as follows:

$$\begin{aligned} \hat{p}(Y_{\text{new}} | Y, X_{\text{new}}) &= \frac{1}{M} \sum_{k=1}^{M} p(Y_{\text{new}} | Y, X_{\text{new}}, Z^{(k)}, \beta^k, \mathbf{\Lambda}^{(k)}, \gamma^{(k)}). \end{aligned} \tag{20}$$

3. EMPIRICAL STUDIES

We now illustrate the practical utility of the proposed TH-SSVS approach via two well-known data sets: the colon cancer data analyzed initially by Alon et al. (1999), and the leukemia data analyzed by Golub et al. (1999). The performance in gene selection and prediction accuracy of the TH-SSVS approach will be compared with the existing gene selection and classification methods.

3.1 Colon cancer data

Alon et al. (1999) used Affymetrix Oligonucleotide Array to measure expression levels of 40 tumor and 22 normal colon tissues for 6,500 human genes. These samples were collected from 40 different colon cancer patients, in which 22 patients supplied both normal and tumor samples. A selection of 2,000 genes based on highest minimal intensity across the samples was conducted by Alon et al. (1999), and the data are publicly available at http://microarray.princeton.edu/oncology/affydata/. Alon et al. (1999) discussed the application of clustering methods for analyzing expression patterns of different cell types. One cluster consists of 5 tumors and 19 normal tissues, while the second contains 35 tumors and 3 normal tissues. We analyzed these data further by taking a base 10 logarithmic of each expression level, and then standardized each tissue sample to zero mean and unit variance across the genes.

In our Bayesian analysis based on the TH-SSVS approach, we set $a = 6, b = 8, \tau = 1, \theta_i = 0.005, i = 1, \cdots, p$, and $h = 100$. To check convergence, three chains with different initial values of Z and γ are run. The initial values $\gamma^{(0)}$ were obtained based on randomly selecting different 25 genes for models I and II from a total of 2,000 genes, and setting $\gamma_i^{(0)} = 1$ if the i-th gene is among the selected genes and $\gamma_i^{(0)} = 0$ otherwise. Three diagnostic plots recommended by Smith and Kohn (1996) and Brown et al. (1998) were used to check convergence. Fig.1(a) shows that the most significant genes, which are determined by the posterior gene inclusion probabilities. Fig.1(b) plots the number of selected genes versus the iteration number, and Fig.1(c) plots the log relative posterior probabilities of selected genes, $\log(p(\gamma | Y, \mathbf{X}, Z))$, versus the iteration number. Fig.1(b) and Fig.1(c) indicate that the the chain converged well enough

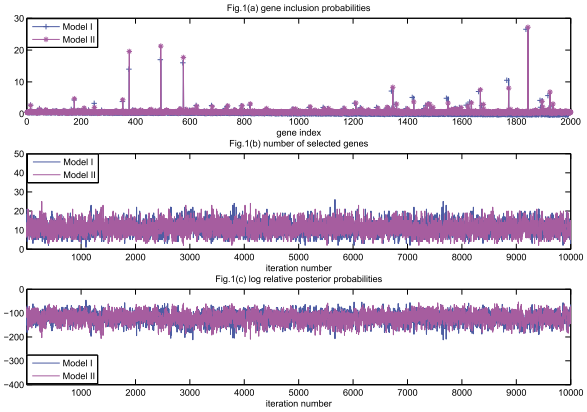

Figure 1. Fig.1(a) shows the gene inclusion probabilities (in percentages) versus the gene index, Fig.1(b) and Fig.1(c) show the number of selected genes and the log relative posterior probabilities of selected genes versus the first 10,000 iteration number, respectively.

within 10,000 iterations. We collected 50,000 observations after 10,000 burn-in iterations to get the estimates of the posterior gene inclusion probabilities (see (17)).

The 18 most significant genes ranked by the posterior gene inclusion probabilities (see Fig. 1(a)) for models I and II are presented in Table 1. At least seven of them were also selected by Ben-Dor et al. (2000). One of the top-ranked genes listed in Table 1 is uroguanylin precursor Z50753. Notterman et al. (2001) showed that a reduction of uroguanylin might be an indication of colon tumors; and Shailubhai et al. (2000) reported that treatment with uroguanylin has a positive therapeutic significance to the reduction in precancerous colon ploys. The second selected gene in Table 1 is R87126 (myosin heavy chain, nonmuscle). The isoform B of R87126 acts as a tumor supressor and is well-known as a component of the cytoskeletal network (Yam et al. 2001, among others). The discriminative power of gene J02854 also has a biological interpretation, because it is known to be an intracellular target of integrins, affecting cell motility (Keely et al., 1998).

Since there is no test set available, it is common to evaluate the performance of the classification methods for a selected subset of genes by the LOOCV procedure (Lachenbruch and Mickey, 1968; McLachlan, 1992 and Gelfand, 1996). Some existing methods in the literature calculated the LOOCV error within the gene selection process. However, as pointed out by the referees, this internal LOOCV procedure is biased and provides optimistic results. Therefore, an external LOOCV procedure proposed by Ambroise and McLachlan (2002) was used in our analysis. Similar to many other multivariate methods, this procedure is challenged by server memory requirements and large computational time. According to the traditional attempts to overcome these problems (see Antoniadis, et al. 2003; Le Cao and

Table 1. Colon cancer data: strongly significant genes for classifying normal and tumor tissues

No.	Clone ID	Gene annotation
1	H06524$^{\text{I,II}}$	Gelsolin precursor, plasma (human)$^{+}$
2	R87126$^{\text{I,II}}$	MYOSIN HEAVY CHAIN, NONMUSCLE^{+}
3	D14812$^{\text{I,II}}$	Human mRNA for ORF, complete cds
4	Z50753$^{\text{I,II}}$	H.sapiens mRNA for GCAP-II/uroguanylin precursor^{+}
5	H08393$^{\text{I,II}}$	COLLAGEN ALPHA 2(XI) CHAIN^{+}
6	T62947$^{\text{I,II}}$	60S RIBOSOMAL PROTEIN L24^{+}
7	M82919$^{\text{I,II}}$	Human gamma amino butyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds.
8	H64807$^{\text{I,II}}$	PLACENTAL FOLATE TRANSPORTER
9	J02854$^{\text{I,II}}$	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM(HUMAN);$^{+}$
10	H11084$^{\text{I,II}}$	Vascular endothelial growth factor
11	R99907$^{\text{I,II}}$	INTERFERON REGULATORY FACTOR 2
12	T94579$^{\text{I,II}}$	Human chitotriosidase precursor mRNA, complete cds.
13	M36634$^{\text{I,II}}$	Human vasoactive intestinal peptide mRNA,$^{+}$
14	T57882$^{\text{I,II}}$	Myosin heavy chain, nonmuscle type A
15	R55310$^{\text{I,II}}$	S36390 Mitochondrial processing reptidase
16	T64012$^{\text{I}}$	ACETYLCHOLINE RECEPTOR PROTEIN, DELTA CHAIN PRECURSOR
17	H09719$^{\text{I,II}}$	TUBULIN ALPHA-6 CHAIN (Mus musculus)
18	M63391$^{\text{I}}$	Human desmin gene, complete cds.
19	T92451$^{\text{II}}$	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE;$^{+}$
20	X62048$^{\text{II}}$	H.sapiens Wee1 hu gene.$^{+}$

+: Ben-Dor et al. (2000)

Chabrier, 2008), we perform the external LOOCV procedure as follows: 1) omit one observation of the training set, 2) based on the remaining observations, the p^* most significant genes were chosen by our TH-SSVS approach, 3) the chosen pfi genes were used to classify the left out sample, and 4) go back to step 1) and select another observation. This process was repeated for all observations in the training set until each observation had been held out and predicted exactly once. From the full set of 2,000 genes, our method on average selected only 10 genes at each MCMC step. The performance of our method with p^*=10 is summarized in Table 2. With 10 genes, models I and II misclassified 2 tumor tissues (T33, T36) and 1 normal tissue (N36). Alon et al. (1999), using a muscle index based on the average intensity of ESTs, misclassified 5 tumor tissues (T2, T30, T33, T36, T37) and 3 normal tissues (N8, N12, N34). Furey et al. (2000), applying the support vector machine (SVM) with 1,000 or 2,000 genes, misclassified 3 tumor tissues (T30, T33, T36) and 3 normal tissues (N8, N34, N36). It is interesting to notice that N36 and T36 were originated from the same patient, and both were consistently misclassified by SVM and TH-SSVS approaches. Our LOOCV results have been compared with the following classification methods: support vector machine (SVM; Furey et al., 2000); LogitBoost optimal, LogitBoost

Table 2. Comparison of LOOCV performance of different approaches for Colon cancer data

	Method	No. of genes	LOOCV error rate
1	SVM[a]	1000 or 2000	0.0968
2	LogitBoost, optimal[b]	2000	0.1290
3	Classification tree[b]	200	0.1452
4	MAVE-LD[c]	50	0.1613
5	1-nearest-neighbor[b]	25	0.1452
6	LogitBoost, estimated[b]	25	0.1935
7	SGLasso[c]	19	0.1290
8	LogitBoost, 100 iterations[b]	10	0.1452
9	AdaBoost, 100 iterations[b]	10	0.1613
10	BPR[e]	22	0.1129
11	L1-SVM[f]	15	0.0968
12	SVM-RFE[g]	24	0.0806
13	TH-SSVS^{I,II}	10	0.0484

a: Furey et al. (2000);
b: Dettling and Bühlmann (2003);
c: Antoniadis et al. (2003);
d: Ma et al. (2007);
e: Lee et al. (2003);
f: Bradley et al. (1998);
g: Guyon et al. (2002).

estimated, LogitBoost 100 iterations, AdaBoost 100 iterations, 1-nearest-neighbor, and Classification tree (Dettling and Buhlmann, 2003); MAVE-LD (Antoniadis et al., 2003), Supervised group Lasso (SGLasso; Ma et al., 2007), L1-SVM (Bradley et al., 1998), SVM-RFE (recursive feature elimination) (Guyon et al., 2002) and Bayesian probit regression (BPR) method (Lee et al., 2003). The summary is presented in Table 2. It is clear from the comparison that our method, which used fewer genes, is better than or comparable to the other popular classification methods.

To assess the sensitivity of the Bayesian results to the inputs of hyperparameters in the prior distributions, we reanalyzed the data set by using different values of a, b, τ, h, and θ_i. For instance, $b = 16, \tau = 0.5, h = 200$, and $\theta_i = 0.007$, the identification of the relevant genes and the performance of classification are essentially the same as before.

3.2 Leukemia data

We further illustrate the performance of our classification procedure on the leukemia dataset (Golub et al., 1999), which is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. This gene expression level was obtained from Affymetrix high-density oligonucleotide arrays containing $p = 6{,}817$ human genes. Golub et al. (1999) gathered bone marrow or peripheral blood samples from 72 patients suffering either from acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML), which were identified based on myeloid (bone marrow related) and their origins, lymphoid (lymph or lymphatic tissue related), respectively. The data comprise 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML, which have already been divided into a training set consisting of 38 samples of which 11 are AML and 27 are ALL; and a test set of 34 samples of which 20 are ALL and 14 are AML.

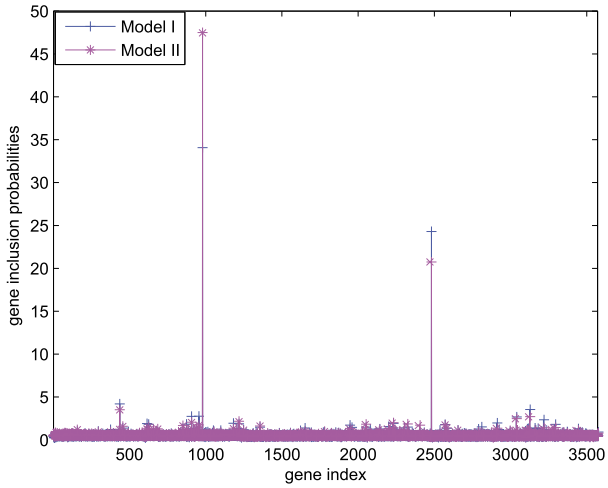

Figure 2. Fig. 2 shows the gene inclusion probabilities (in percentages) versus the gene index for leukemia data.

Based on the protocol given in Dudoit et al. (2002), the following preprocessing steps were taken for the data: (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with max/min$\leq$5 and (max-min)$\leq$ 500, where max and min refer respectively to the maximum and minimum expression levels of a particular gene across samples; and (iii) base 10 logarithmic transformation. The filtering resulted in 3,571 genes. We further transformed the gene expression data to have mean zero and standard deviation one across samples. We applied the Bayesian TH-SSVS method with the same inputs of the hyperparameters as in the first example. An initial value of γ was similarly obtained as before via 25 randomly selected genes from a total of 3,571 genes.

The posterior gene inclusion probabilities for models I and II are presented in Figure 2, and the relevant genes selected on the basis of these probabilities are reported in Table 3. Moreover, the relevant genes selected by Golub et al. (1999) and Ben-Dor et al. (2000) are also shown. One of the most significant gene is Zyxin. Macclama et al. (1996) has shown that Zyxin encodes an LIM domain protein localized at focal contacts in adherent erythroleukemian cells. It has also been recently demonstrated that Zyxin exports from the nucleus by intrinsic leucine rish nuclear export sequences, and enter the nucleus through association with other proteins.Wang and Gilmore (2003) reported that misregulation of nuclear functions of Zyxin protein seems to be associated with pathogenic effects. Therefore, it is not surprising that Zyxin plays an important role in classifying AML and ALL. Among the top-ranked genes we also found CD33 antigene with known expression specificity to AML (Sobol et al.

Table 3. Leukemia data: strongly significant genes for discriminating ALL and AML samples

No	Gene ID	Gene descriptions
1	M27891^{I,II}	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)$^{+*}$
2	X95735^{I,II}	Zyxin^{+*}
3	D88422^{I,II}	CYSTATIN A*
4	M27783^{I,II}	ELA2 Elastatse 2, neutrophil
5	M23197^{I,II}	CD33 antigen (differentiation antigen)$^{+*}$
6	M16038^{I,II}	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog^{+*}
7	L09209^{I,II}	APLP2 Amyloid beta (A4) precursor-like protein 2*
8	M83652^{I}	PFC Properdin P factor, complement*
9	U22376^{I}	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds^{+*}
10	M84526^{I}	DF D component of complement (adipsin)$^{+*}$
11	X62654^{I,II}	ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen*
12	J04615^{I}	SNRPN Small nuclear ribonucleoprotein polypeptide N*
13	M92287^{I,II}	CCND3 Cyclin D3*
14	J05243^{I,II}	SPTAN1 Spectrin, alpha, non-erythrocytic1 (alpha-fodrin)*
15	M11722^{I}	Terminal transferase mRNA*
16	Y12670^{I,II}	LEPR Leptin receptor^{+}
17	X85116^{I}	Epb72 gene exon 1^{+*}
18	U82759^{I}	GB DEF = Homeodomain protein HoxA9 mRNA
19	X74262^{II}	RETINOBLASTOMA BINDING PROTEIN P48^{+}
20	X04085^{II}	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)$^{+}$
21	X82240^{II}	TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/lymphoma 1
22	L47738^{II}	Inducible protein mRNA^{+*}
23	U05259^{II}	MB-1 gene^{+*}
24	HG1612^{II}	Macmarcks*
25	M22960^{II}	PPGB Protective protein for beta-galactosidase*

+: Golub et al. (1999);
∗: Ben-Dor et al. (2000).

Table 4. The comparison of classification methods for the leukemia data

	Method	No. of genes	Test error rate
1	SVM^{a}	25 to 1000	0.0588 to 0.1176
2	WVM^{b}	50	0.1471
3	MAVE-LD^{c}	50	0.0294
4	MAVE-NPLD^{c}	50	0.0294
5	PLS-LD^{d}	50	0.0294
6	PLS-QDA^{d}	50	0.1765
7	BPR^{e}	28	0.0294
8	L1-SVM^{f}	24	0.0000
9	SVM-RFE^{g}	32	0.0294
10	TH-SSVS^{I,II}	12	0.0000

a: Furey et al. (2000);
b: Golub et al. (1999);
c: Antoniadis et al. (2003);
d: Nguyen and Rocke (2002).
e: Lee et al. (2003);
f: Bradley et al. (1998);
g: Guyon et al. (2002).

1987), CD63 antigene known as a member of the tranmembrane 4 superfamily (Smith et al., 1995), and Macmarks involves in growth and metastasis of certain tumors (Spizz and Blackshear, 1997).

Only 12 genes are selected on average by our method, and the $p^* = 12$ genes were used to conduct prediction on the test set. In Table 4, we compare our classification results with the following popular classification methods: SVM (Furey et al., 2000); weighted voting machine (Golub et al., 1999); MAVE-LD and MAVE-NPLD (Antoniadis et al., 2003); PLS-LD and PLS-QDA (Nguyen and Rocke, 2002); L1-SVM (Bradley et al., 1998); SVM-RFE (recursive feature elimination) (Guyon et al., 2002) and Bayesian probit regression (BPR) method (Lee et al., 2003). Our results, with fewer genes, are better than or comparable to those obtained by the above existing methods in the literature. Most of the methods have also done well in accurately classifying the leukemia types in the test set. So here we do not gain much with respect to classification accuracy, but this well known data set is used as a validation of continued reliable performance of our TH-SSVS method.

4. SIMULATION STUDY

4.1 Simulation setup

This section illustrates our proposed TH-SSVS method using simulated data and demonstrates the effectiveness of our method for binary classification problems. We compared the performance of TH-SSVS method with L1-SVM (Bradley et al., 1998), SVM-RFE (recursive feature elimination) (Guyon et al., 2002) and Bayesian probit regression (BPR) method (Lee et al., 2003). L1-SVM, SVM-RFE and BPR methods can do gene selection and predict the tumor class simultaneously like our TH-SSVS method. The optimal tuning and other parameters in these four methods are obtained by LOOCV technique on the training set. To further show that our model is robust to the choice of the prior parameters, three different prior settings are tried for the simulation study. Prior set 1: $a = 6, b = 16, \tau = 1, \theta_i = 0.005, i = 1, \cdots, p, h = 100$; Prior set 2: $a = 6, b = 8, \tau = 0.5, \theta_i = 0.005, i = 1, \cdots, p, h = 100$; Prior set 3: $a = 6, b = 8, \tau = 1, \theta_i = 0.007, i = 1, \cdots, p, h = 100$.

In the simulation, we use the leukemia data set (Golub et al., 1999) and use BWSS criteria to extract top 40 genes, so $x_i, i = 1, \cdots, 40$ are those microarray measurements from the leukemia data set, and generate $x_i \sim N(0, 0.01), i =$

Table 5. Simulation Results

	Method	No. of variables	Misclassification error rate	SD	TP	TN
1	BPR	45	4.82	4.52	75.00	99.13
2	L1-SVM	49	4.10	3.53	72.50	98.98
3	SVM-RFE	41	4.69	4.34	67.50	99.28
4	TH-SSVS^{I,II}-1	37	2.95	2.54	87.50	99.89
5	TH-SSVS^{I,II}-2	37	2.95	2.48	85.00	99.85
6	TH-SSVS^{I,II}-3	38	3.02	2.79	80.00	99.69

$41, \cdots, 2000$. The covariates are designed so that only 2% of the variables are relevant for classification and the rest of the variables are redundant. We generate 72 samples in total, and randomly split it into 38 samples in the training set and 34 samples in the test set. The class labels are kept as in the original data set. Therefore, the number of covariates is larger than the number of samples. We replicate the simulation 100 times. Then we want to see how precisely our proposed method can select the 40 relevant variables from the redundant variables, and check whether the prediction performance of the proposed method is better than that of the other four methods.

4.2 Simulation results

We report the simulation results of different methods in Table 5. The simulation results of our proposed method under three different prior settings are denoted as TH-SSVS-1, TH-SSVS-2, and TH-SSVS-3. In this table, the numbers in the second row under "No. of Variables" are the median number of total variables selected by L1-SVM, SVM-RFE, BPR and TH-SSVS methods. These variables are then used by different methods for classification. Since we exactly know which variables are relevant and which are redundant in this simulation, we can check how successfully the methods are able to select the relevant variables and eliminate the redundant variables. The median true positive (TP) and true negative (TN) rates are included in Table 5. The TP rate is the percentage of truly relevant variables selected, and the TN rate is the percentage of redundant variables not selected. From Table 5, it can be seen that our proposed TH-SSVS method produces more than 80% TP and TN rates, which are respectively higher than that of the other three methods. Ideally we would like to have both TP and TN rates as high as possible. The big novelty of our method is sparse Bayesian variable selection, and from the reported TP and TN rates we see that our method is able to select the true relevant variables and discard the redundant variables and produce a more sparse model.

From the misclassification results summarized in Table 5, we can clearly see that in terms of average misclassification error, our TH-SSVS method attains lower misclassification error than L1-SVM, SVM-RFE and BPR methods. The misclassification results of our method under three different priors are similar. Thus we can draw the conclusion that our method is not sensitive to the choice of the prior. However, it should be pointed out that the prior inclusion probabilities θ_i should be set to a small value, as a large value may lead to many irrelevant variables selected in the model and might reduce the prediction accuracy.

5. DISCUSSION

In this paper, we propose a sparse Bayesian probit regression model together with stochastic search variable selection approach for gene selection and cancer classification. Our proposed model employs two different sparsity-enforcing priors for the regression coefficients. These sparsity-enforcing priors can be rewritten in two level hierarchical manner for simplicity. Simulation-based MCMC method is introduced to estimate the unknown parameters. Moreover, by integrating some parameters out from some full conditional distributions and jointly updating the parameters, we can apply an efficient sampling scheme to simulate samples from the posterior distributions. Other nice features of our approach also include the flexibility in choosing the initial value of γ, and the ability in providing posterior gene inclusion probabilities to achieve biological interpretation. We demonstrate the performance of our TH-SSVS approach on the colon cancer and leukemia data sets. With a small subset of relevant gene, our approach compared favorably with other popular approaches in performing disease classification.

While a, b, τ and θ are treated as known hyperparameters, they can be treated as unknown parameters with hierarchical prior distributions to them. The project of extension to more than two categories using multinomial probit model is ongoing. We assume that genes are independent. In our future research, we will extend the model to account for a interaction structure between genes.

ACKNOWLEDGEMENTS

We thank the editor and reviewers for insightful comments, which have led to a significant improvement of this article. Supported by the grant of Natural Science Foundation of China (11501294,11501261), China Postdoctoral Science Foundation (2015M580374,2016T90398), Natural Science Foundation of Guangdong (2016A030313856), Jiangsu Qinglan Project(2017), Open Project Program of the Key Laboratory of Statistical Information Technology and Data Mining (SDL201704) and Project of Natural Science Research in Jiangsu Province (15KJB110007).

Received 1 March 2014

REFERENCES

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679. MR1224394

Alon, U., et al. (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.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* **99** 6562–6566.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics* **19** 1–8.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20** 3423–3430.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *J Comput. Biol.* **7** 559–583.

Bradley, P. and Mangasarian, O. (1998). Feature selection via concave minimization and support vector machines. *In: Proceedings of the 15th International Conference on Machine Learning*, 82–90.

Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statist. Soc. B* **60** 627–641. MR1626005

Chakraborty, S., Ghosh, M., Maiti, T. and Tewari, A. (2004). Bayesian neural networks for bivariate binary data: An application to prostate cancer study. *Statistics in Medicine* **24** 3645–3662. MR2212305

Chakraborty, S., Ghosh, M., Mallick, B. K., Ghosh, D. and Dougherty, E. (2007). Gene Expression-Based Glioma Classification Using Hierarchical Bayesian Vector Machines. *Sankhya* **69** 514–547. MR2460007

Chakraborty, S. (2009). Bayesian Binary Kernel Probit Model for Microarray Based Cancer Classification and Gene Selection. *Computational Statistics and Data Analysis* **53** 4198–4209. MR2744317

Chakraborty, S. and Guo, R. (2011). Bayesian Hybrid Huberized SVM and its Applications in High Dimensional Medical Data. *Computational Statistics and Data Analysis* **55(3)** 1342–1356. MR2741419

Chhikara, R. and Folks, L. (1989). *The inverse gaussian distribution: theory, methodology, and applications*. Marcel Dekker, New York.

Cristianini, N. and Shawe-Taylor, J. (1999). *An Introduction to SVM*. Cambridge University Press, Cambridge.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** 1061–1069.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York. MR0836973

Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics* **2** 28–34.

Dudoit, Y., Yang, H., Callow, M. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97** 77–87. MR1963389

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16** 906–914.

Gelfand, A. (1996). Model determination using sampling-based methods. *in Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (ed.), Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, pp. 145–158. MR1397969

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **6** 721–741.

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

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in practise*. Chapman and Hall, London. MR1397966

Golub, T. R., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* **82(4)** 711–732. MR1380810

Gupta, M. and Ibrahim, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102** 867–880. MR2411650

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., et al. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46** 389–422.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Element of Statistical Learning*. Springer-Verlag, New York. MR1851606

Keely, P., Parise, L. and Juliano, R. (1998). Integrins and GTPases in tumour cell growth, motility and invasion. *Trends In Cell Biology* **8** 101–107.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10** 1–11. MR0223016

Lamnisos, D., Griffin, J. E. and Steel, F. J. Mark (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics* **18** 592–612. MR2751643

Le Cao K.-A. and Chabrier, P. (2008). ofw: an R package to selection continuous variables for multi-class classification with a stochastic wrapper method. *Journal of Statistical Software* **28** 1–16.

Lee, K. E. et al. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19** 90–97.

Li, Y., Campbell, C. and Tipping, M. (2002). Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics* **18** 1332–1339.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis* **5(1)** 847–866. MR2596439

Liu, X., Krishnan, A. and Mondry, A. (2005). An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* **6** 76.

Ma, S., Song, S. and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 1471–2105.

Maccalma, T. et al. (1996). Molecular characterization of human zyxin. *J. Biol. Chem.* **271** 31470–31478.

Mallick, B. K., Ghosh, D., Ghosh, M. (2005). Bayesian classification of tumors using gene expression data. *Journal of the Royal Statistical Society, B* **67** 219–232. MR2137322

Mallick, B. K., Chakraborty, S., Ghosh, M. (2011). Comment on Article by Polson and Scot. *Bayesian Analysis* **6(1)** 25–30. MR2781804

Mamitsuka, H. (2006). Selecting features in microarray classification using ROC curves. *Pattern Recognition* **39** 2393–2404.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley, New York. MR1190469

Nguyen, D. V. and Rocke, D. M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **18** 1216–1226.

Notterman, D. et al. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research* **61** 3124–3130.

Panagiotelisa, A. and Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high dimensional additive models. *Journal of Econometrics* **143** 291–316. MR2389611

Park, K. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60** 812–819. MR2089459

Shailubhai, K. et al. (2000). Uroguanylin treatment suppresses polyp formation in the Apc(Min/+) mouse and induces apoptosis in hu-

man colon adenocarcinoma cells via cyclic GMP. *Cancer Research* **60** 5151–5157.

Smith, D. et al. (1995). Antibodies against human CD63 activate transfected rat basophilic leukemia (RBL-2H3) cells. *Molecular Immunology* **32** 1339–1344.

Smith, M. and Kohn, R. (1996). Nonparametric regression via Bayesian variable selection. *Journal of Econometrics* **75** 317–343.

Sobol, R. et al. (1987). Clinical importance of myeloid antigen expression in adult acute lymphoblastic leukemia. *N. Eng. J. Med.* **316** 1111–1117.

Spizz, G. and Blackshear, P. (1997). Identification and characterization of cathepsin B as the cellular MARCKS cleaving enzyme. *J. Biol. Chem.* **272** 23833–23842.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99** 6567–6572.

Wainwright, M., Ravikumar, P. and Lafferty, J. (2006). High-dimensional graphical model selection using L1- regularized logistic regression. *Advances in Neural Information Processing Systems* **19**.

Wang, Y. and Gilmore, T. (2003). Zyxin and paxillin proteins: focal adhesion plaque lim domain proteins go nuclear. *Biochimica et Biophysica Acta* **1593** 115–120.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. (2001). Feature selection for SVMs. *Adv. Neural Inform. Process. Syst.* **13** 668–674.

Yam, J., Chan, K. and Hsiao, W. (2001). Suppression of the tumorigenicity of mutant p53- transformed rat embryo fibrob- 97 lasts through expression of a newly cloned rat nonmuscle myosin heavy chain-B. *Oncogene* **20** 58–68.

Yang, A. and Song, X. (2010). Bayesian variable selection for disease classiffication using gene expression data. *Bioinformatics* **26(2)** 215–222.

Yang, K., Cai, Z., Li, J. and Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC Bioinformatics* **7** 228.

Yeung, K. Y., Bumgarner, R. E. and Raftery, A. E. (2005). Bayesian Model Averaging: Development of an improved multiclass, gene selection and classification tool for microarray data. *Bioinformatics* **21** 2394–2402.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* **5** 1205–1224. MR2248015

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *In: Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. North Holland, Amsterdam, pp. 233–243. MR0881437

Zhou, X., Wang, X., and Dougherty, E. (2004a). A Bayesian approach to non linear probit gene selection and classification. *Journal of the Franklin Institute* **341** 137–156. MR2060946

Zhou, X., Wang, X., and Dougherty, E. (2004b). Gene prediction using multinomial probit regression with Bayesian gene selection. *EURASIP Journal on Applied Signal Processing* **1** 115–124.

Zhou, X., Liu, K., and Wong, S. (2004c). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics* **37** 249–259.

Aijun Yang
College of Economics and Management
Nanjing Forestry University
Jiangsu
China
Key Laboratory of Statistical Information Technology and Data Mining
State Statistics Bureau
Chengdu
China
E-mail address: ajyang81@163.com

Heng Lian
Department of Mathematics
City University of Hong Kong
Kowloon Tong
Hong Kong
China
E-mail address: henglian@cityu.edu.hk

Xuejun Jiang
Department of Mathematics
South University of Science and Technology of China
Shenzhen
China
E-mail address: jiangxj@sustc.edu.cn

Pengfei Liu
School of Mathematics and Statistics
Jiangsu Normal University
Xuzhou
China
E-mail address: jiangxj@sustc.edu.cn