Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 420412, 11 pages
http://dx.doi.org/10.1155/2013/420412

Research Article

Applications of Bayesian Gene Selection and Classification with Mixtures of Generalized Singular g-Priors

Wen-Kuei Chien[1] **and Chuhsing Kate Hsiao**[2,3]

[1] *Biostatistics Center, Taipei Medical University, Taipei 11031, Taiwan*
[2] *Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei 10055, Taiwan*
[3] *Bioinformatics and Biostatistics Core, Division of Genomic Medicine, Research Center for Medical Excellence, National Taiwan University, Taipei 10055, Taiwan*

Correspondence should be addressed to Chuhsing Kate Hsiao; ckhsiao@ntu.edu.tw

Received 4 September 2013; Revised 10 November 2013; Accepted 10 November 2013

Academic Editor: Ao Yuan

Recent advancement in microarray technologies has led to a collection of an enormous number of genetic markers in disease association studies, and yet scientists are interested in selecting a smaller set of genes to explore the relation between genes and disease. Current approaches either adopt a single marker test which ignores the possible interaction among genes or consider a multistage procedure that reduces the large size of genes before evaluation of the association. Among the latter, Bayesian analysis can further accommodate the correlation between genes through the specification of a multivariate prior distribution and estimate the probabilities of association through latent variables. The covariance matrix, however, depends on an unknown parameter. In this research, we suggested a reference hyperprior distribution for such uncertainty, outlined the implementation of its computation, and illustrated this fully Bayesian approach with a colon and leukemia cancer study. Comparison with other existing methods was also conducted. The classification accuracy of our proposed model is higher with a smaller set of selected genes. The results not only replicated findings in several earlier studies, but also provided the strength of association with posterior probabilities.

1. Introduction

Recent advancement in oligonucleotide microarray technologies has resulted in production of thousands of gene expression levels in a single experiment. With such vast amount of data, one major task for researchers is to develop classification rules for prediction of cancers or cancer subtypes based on gene expression levels of tissue samples. The accuracy of such classification rules may be crucial for diagnosis and treatment, since different cancer subtypes may require different target-specific therapies. However, the development of good and efficient classification rules has not been straightforward, either because of the huge number of genes collected from a relatively small number of tissue samples or because of the model complexity associated with the biological mechanism. The identification of a smaller set of relevant genes to characterize different disease classes, therefore, has been a challenging task. Procedures which are efficient in gene selection as well as in classification do play an important role in cancer research.

Many approaches have been proposed for classes classification. For example, several analyses identified a subset of classifying genes with t-statistics, regression model approach, mixture model, Wilcoxon score test, or the between-within classes sum of squares (BSS/WSS) [1–7]. These methods are univariate in the sense that each gene is tested individually. Others started with an initial step of dimension reduction before classification procedures, such as the principle components analysis (PCA) [8–10] and the partial least squares algorithm (PLS algorithm) [11–15]. These methods may reduce dimension (the number of genes) effectively but may not be biologically interpretable. To capture the gene-gene correlations, researchers proposed the pair-based method [16], correlation-based feature selection [17], and the Markov random field prior [18]. Although these methods can model the gene-gene interaction, they can be computationally time-consuming.

Bayesian approach can accommodate naturally the interplay between genes via prior distributions, under the setting of regression models. Examples included the Bayesian hierarchical mixture model [19–21] and a logistic or probit link with latent variables and stochastic search variable selection (SSVS) procedure for binary and multicategorical phenotypes [22–25]. To consider all genes simultaneously, most Bayesian approaches adopt a multivariate analysis with a natural conjugate prior $N(\mathbf{0}, c(\mathbf{X}^T\mathbf{X})^{-1})$, called g-prior, for the regression parameters $\boldsymbol{\beta}$ [26]. This *a priori* distribution utilizes the design matrix as the prior covariance matrix of $\boldsymbol{\beta}$ and can lead to a relatively simple posterior distribution. However, if the number of genes is much larger than the number of samples available, the dimension of $\mathbf{X}$ becomes large and a high degree of multicollinearity may occur. In that case, the covariance matrix of Zellner's g-prior becomes nearly singular. Modifications included the gsg-prior distribution with the Moore-Penrose generalized inverse matrix [27] and use of a ridge parameter [28, 29]. Alternatively, other researchers focused on the scalar c in $c(\mathbf{X}^T\mathbf{X})^{-1}$ which controls the expected size of the nonzero regression coefficients. For instance, it was reported that the final results are insensitive to the values of c between 10 and 100, and the value $c = 100$ has been suggested after extensive examinations [30]. Instead of fixing c at a constant, George and Foster [31] proposed an empirical Bayes estimate for c, while Liang and colleagues [32] suggested a hyper-g prior, a special case of the incomplete inverse-gamma prior in Cui and George [33].

The main purpose of this research is the application of fully Bayesian approaches with a hyperprior on c. Specifically we adopted an inverse-gamma prior $\mathrm{IG}(1/2, n/2)$ which was commented earlier that it could lead to computational difficulty. Therefore, we outlined a MCMC algorithm and demonstrated its implementation. In this paper, we considered a probit regression model for classification with SSVS to identify the influential genes, augmented the response variables $Y_1, Y_2, \ldots, Y_n$ with latent variables $Z_1, Z_2, \ldots, Z_n$, and converted the probit model to a Gaussian regression problem with the generalized singular g-prior (gsg-prior). For the choice of c, we assigned a hyperprior for the uncertainty in c. This hyperprior is intuitive and differs from those in [32, 33]. Finally, we defined an indicator variable γ_j for the jth gene and perform MCMC methods to generate posterior samples for gene selection and class classification. The rest of the paper is arranged as follows. In Section 2, we briefly described the model specification including the data augmentation approach and SSVS methods. Under this hyperprior on c, we also demonstrated the implementation of the Bayesian inference. Applications of three cancer studies, acute leukemia, colon cancer, and large B-cell lymphoma (DLBCL), were presented in Section 3. Conclusion and discussion were given in Section 4.

2. Model and Notation

Let $(\mathbf{X}, \mathbf{Y})$ indicate the observed data,

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

where x_{ij} denotes the expression level of the jth gene from the ith sample and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ denotes the response vector, where $Y_i = 1$ indicates that sample i is a cancer tissue and $Y_i = 0$ for normal tissue. Assume that $Y_1, Y_2, \ldots, Y_n$ are n independent random variables with $p_i = \Pr(Y_i = 1)$.

2.1. Probit Model with Latent Variable. The gene expression measurements can be linked to the response outcome with a probit regression model:

$$p_i = \Pr(Y_i = 1) = \Phi(\alpha + \mathbf{X}_i\boldsymbol{\beta}), \qquad (2)$$

where α represents the intercept, $\mathbf{X}_i$ is the ith row in the $n \times p$ design matrix $\mathbf{X}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, and Φ is the standard normal cumulative distribution function.

To perform statistical inference under this probit regression model, we first adopt n independent latent variables $Z_1, Z_2, \ldots, Z_n$, where

$$Z_i = \alpha + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \; i = 1, \ldots, n, \qquad (3)$$

and the Z_i corresponds to the disease status as

$$Y_i = \begin{cases} 1, & \text{if } Z_i > 0, \\ 0, & \text{if } Z_i \leq 0. \end{cases} \qquad (4)$$

The use of such latent variables helps to determine which category the ith sample is to be classified. Note that multiplying a constant on both sides in (3) does not change the model; thus a unit variance is considered for ε_i.

If a noninformative prior is assumed for $\boldsymbol{\beta}$, then the posterior covariance matrix of $\boldsymbol{\beta}$ given $\mathbf{Z} \equiv (Z_1, Z_2, \ldots, Z_n)$ becomes $(\mathbf{X}^T\mathbf{X})^{-1}$. However, due to the enormous size of microarray data, $(\mathbf{X}^T\mathbf{X})^{-1}$ may be nearly singular, and variable selection for dimension reduction is needed. We define for variable selection the vector $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, \ldots, \gamma_p)$ whose elements are all binary, where

$$\gamma_i = \begin{cases} 1, & \text{if } \beta_i \neq 0 \text{ (the } i\text{th gene selected)}, \\ 0, & \text{if } \beta_i = 0 \text{ (the } i\text{th gene not selected)}. \end{cases} \qquad (5)$$

Given $\boldsymbol{\gamma}$, we denote p^γ as the number of 1's in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}^\gamma$ a $p^\gamma \times 1$ reduced vector containing the regression coefficients β_j if its corresponding γ_j is 1. Accordingly, for all $\gamma_j = 1$, the corresponding columns in $\mathbf{X}$ are collected to build $\mathbf{X}^\gamma$, an $n \times p^\gamma$ reduced gene expression matrix. Given $\boldsymbol{\gamma}$, the probit regression model in (3) can be written as

$$Z_i = \alpha + \mathbf{X}_i^\gamma\boldsymbol{\beta}^\gamma + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \; i = 1, \ldots, n, \qquad (6)$$

where $\mathbf{X}_i^\gamma$ is the ith row in $\mathbf{X}^\gamma$.

2.2. Choice of Prior Distributions. To complete the model specification, we assign a normal $N(0,h)$ prior for the intercept α with a large h indicating no *a priori* information. For the regression parameters, the commonly applied g-prior $\boldsymbol{\beta}^{\gamma} \mid \boldsymbol{\gamma}, c \sim N(\mathbf{0}, c(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{-1})$ may not work if the sample size n is less than the number p^{γ}, leading to the results that $\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma}$ is not of full rank and $(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{-1}$ does not exist. Therefore, we consider the *gsg*-prior distribution with $(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{+}$ as the pseudoinverse of $\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma}$ for $\boldsymbol{\beta}^{\gamma}$ conditioning on $(\boldsymbol{\gamma}, c)$, $\boldsymbol{\beta}^{\gamma} \mid \boldsymbol{\gamma}, c \sim N(\mathbf{0}, c(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{+})$. This would solve the singularity problem. Next, we assign for $\boldsymbol{\gamma}$ and c the priors

$$\begin{gathered}\pi(c) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} c^{-3/2} e^{-n/(2c)}, \\ \gamma_i \sim \operatorname{Ber}(\pi_i), \quad 0 \le \pi_i \le 1,\ i = 1, \dots, p,\end{gathered} \tag{7}$$

and assume that γ_i are independent for $i = 1, \dots, p$. Note that here the π_i's are of small values, implying a small set of influential genes.

We now complete the model specification:

$$\begin{gathered}\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T, \quad \text{where} \\ p_i = \Pr(Y_i = 1) = \Phi(\alpha + \mathbf{X}_i\boldsymbol{\beta}), \\ Z_i = \alpha + \mathbf{X}_i^{\gamma}\boldsymbol{\beta}^{\gamma} + \varepsilon_i, \quad \text{where} \\ Y_i = 1 \text{ if } Z_i > 0, \text{ and } 0 \text{ otherwise} \\ \boldsymbol{\beta}^{\gamma} \mid \boldsymbol{\gamma}, c \sim N\left(\mathbf{0}, c\left(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma}\right)^{+}\right), \\ \pi(c) \sim \operatorname{IG}\left(\frac{1}{2}, \frac{n}{2}\right), \\ \gamma_i \sim \operatorname{Ber}(\pi_i).\end{gathered} \tag{8}$$

Note that $Y_i = 1$ if the ith sample is a cancer tissue, α is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients, Φ is the standard normal cumulative distribution function, and $\mathbf{X}$ is the design matrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}. \tag{9}$$

And $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, \dots, \gamma_p)$ contains the binary γ_i, where $\gamma_i = 1$ if the ith gene is selected ($\beta_i \ne 0$), $\boldsymbol{\beta}^{\gamma}$ is a $p^{\gamma} \times 1$ reduced vector containing the regression coefficients β_j if its corresponding γ_j is 1, p^{γ} is the number of 1's in $\boldsymbol{\gamma}$, and $\mathbf{X}_i^{\gamma}$ is the ith row in $\mathbf{X}^{\gamma}$.

2.3. Computation and Posterior Inference. Based on the prior distributions specified in previous sections, the joint posterior distribution can be derived as

$$\begin{aligned} & P(\mathbf{Z}, \alpha, \boldsymbol{\beta}^{\gamma}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ & \quad \propto \left[\exp\left\{-\frac{\sum_{i=1}^{n}(Z_i - \alpha - \mathbf{X}_i^{\gamma}\boldsymbol{\beta}^{\gamma})^2}{2}\right\} \prod_{i=1}^{n} I(A_i)\right] \\ & \qquad \cdot \exp\left(-\frac{\alpha^2}{2h}\right) \\ & \qquad \cdot \left[\exp\left(-\frac{\boldsymbol{\beta}^{\gamma T}\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma}\boldsymbol{\beta}^{\gamma}}{2c}\right) \prod_{i=1}^{m_{\gamma}} \lambda_i^{-1/2}\right] \\ & \qquad \cdot \left[\prod_{i=1}^{p} \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}\right] \\ & \qquad \cdot \left[c^{-3/2}\exp\left(-\frac{n}{2c}\right)\right], \end{aligned} \tag{10}$$

where

$$A_i = \begin{cases} \{Z_i : Z_i > 0\} & \text{if } Y_i = 1, \\ \{Z_i : Z_i \le 0\} & \text{if } Y_i = 0, \end{cases} \tag{11}$$

and $\lambda_1, \lambda_2, \dots, \lambda_{m_{\gamma}}$ ($m_{\gamma} \le p_{\gamma}$) are the nonzero eigenvalues of $(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{+}$. From (10), $\boldsymbol{\beta}^{\gamma}$ given $(\mathbf{Z}, \alpha, \boldsymbol{\gamma}, c, \mathbf{Y}, \mathbf{X})$ is a multivariate normal distribution with a covariance matrix $c(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{+}/(c+1)$. In the case where $\mathbf{X}^{\gamma}$ is not of full column rank, the problem of convergence may occur in the MCMC algorithm because the covariance matrix is not positive definite and the multivariate normal distribution becomes degenerated. To avoid this problem and speed up the computations, we integrate out α and $\boldsymbol{\beta}^{\gamma}$ in (10) following Yang and Song's [27] suggestion and derive

$$\begin{aligned} & p(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ & \quad \propto \frac{1}{|\Sigma_{\gamma}|^{1/2}} \exp\left(-\frac{\mathbf{Z}^T\Sigma_{\gamma}^{-1}\mathbf{Z}}{2}\right) \prod_{i=1}^{n} I(A_i) \\ & \qquad \cdot \prod_{i=1}^{p} \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i} c^{-3/2} e^{-n/2c}, \end{aligned} \tag{12}$$

where $\Sigma_{\gamma} = \mathbf{I}_n + h\mathbf{1}\mathbf{1}^T + c\mathbf{X}^{\gamma}(\mathbf{X}^{\gamma T}\mathbf{X}^{\gamma})^{+}\mathbf{X}^{\gamma T}$. As the posterior distribution is not available in an explicit form, we use the MCMC technique to obtain posterior sample observations. The computational sampling scheme is as follows.

(1) Draw $\mathbf{Z}$ from $p(\mathbf{Z} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$, where

$$p(\mathbf{Z} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c) \propto N(\mathbf{0}, \Sigma_{\gamma}) \prod_{i=1}^{n} I(A_i). \tag{13}$$

The conditional distribution of $\mathbf{Z}$ given $(\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$ is a multivariate truncated normal. Since it is difficult to directly sample $\mathbf{Z}$ from this distribution, we draw samples Z_i, $i = 1, \dots, n$, from $p(Z_i \mid \mathbf{Z}_{(-i)}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$, where $\mathbf{Z}_{(-i)}$ is the vector of $\mathbf{Z}$ without the ith element [34].

(2) Draw $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c)$, where

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c) \propto \frac{1}{\left|\Sigma_{\boldsymbol{\gamma}}\right|^{1/2}} \exp\left(-\frac{\mathbf{Z}^T \Sigma_{\boldsymbol{\gamma}}^{-1} \mathbf{Z}}{2}\right) \prod_{i=1}^{p} \pi_i^{\gamma_i} (1 - \pi_i)^{1-\gamma_i}. \tag{14}$$

Similar to the above procedure, we draw samples γ_i, $i = 1, \dots, n$, from $p(\gamma_i \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c)$. It can be shown that

$$\begin{aligned} p\left(\gamma_i \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right) &= \frac{p\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right)}{p\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right) + p\left(\gamma_i = 0 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right)} \\ &= \left(1 + \frac{1 - \pi_i}{\pi_i} \rho\right)^{-1}, \end{aligned} \tag{15}$$

where

$$\begin{aligned} \rho &= \left|\Sigma_{\boldsymbol{\gamma}^1} \Sigma_{\boldsymbol{\gamma}^0}^{-1}\right|^{1/2} \exp\left\{\frac{\mathbf{Z}^T \left(\Sigma_{\boldsymbol{\gamma}^1}^{-1} - \Sigma_{\boldsymbol{\gamma}^0}^{-1}\right) \mathbf{Z}}{2}\right\}, \\ \boldsymbol{\gamma}^1 &= \left(\gamma_1, \dots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \dots, \gamma_p\right), \\ \boldsymbol{\gamma}^0 &= \left(\gamma_1, \dots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \dots, \gamma_p\right), \end{aligned} \tag{16}$$

$\Sigma_{\boldsymbol{\gamma}}^1$ and $\Sigma_{\boldsymbol{\gamma}}^0$ are similar to $\Sigma_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma}$ replaced by $\boldsymbol{\gamma}^1$ and $\boldsymbol{\gamma}^0$, respectively.

(3) Draw c from $p(c \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma})$, where

$$\begin{aligned} p(c \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}) &\propto p(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ &\propto \frac{1}{\left|\Sigma_{\boldsymbol{\gamma}}\right|^{1/2}} \exp\left(-\frac{\mathbf{Z} \Sigma_{\boldsymbol{\gamma}}^{-1} \mathbf{Z}}{2}\right) \cdot c^{-3/2} e^{-n/2c}. \end{aligned} \tag{17}$$

The above distribution does not belong to any standard distribution, so we will use Metropolis-Hastings algorithm to sample c.

The iteration therefore starts with initial values of $\mathbf{Z}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, and $c^{(0)}$, and our MCMC procedures at the tth iteration are as follows.

Step 1. Draw $Z_i^{(t)}$ from $p(Z_i \mid \mathbf{Z}_{(-i)}^{(t-1)}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}^{(t-1)}, c^{(t-1)})$, $i = 1, \dots, n$.

Step 2. For $i = 1, \dots, p$, calculate $p_i^{(t)} \equiv p(\gamma_i^{(t)} = 1 \mid \boldsymbol{\gamma}_{(-i)}^{(t-1)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}^{(t)}, c^{(t-1)})$, generate a random number u_i from $U(0,1)$, and let

$$\gamma_i^{(t)} = \begin{cases} 1, & u_i < p_i^{(t)}, \\ 0, & \text{otherwise}. \end{cases} \tag{18}$$

Step 3. Draw c from (17) by the following steps:

(i) maximize (17) to obtain c_{opt};

(ii) generate the proposal value

$$c^{(t)} = c_{\text{opt}} + \varepsilon^{(t)}, \tag{19}$$

where $\varepsilon^{(t)}$ follows a normal $N(\mu, \sigma^2)$ truncated in a positive region (a,b) with a density q;

(iii) accept $c^{(t)}$ with the acceptance probability:

$$R = \min\left\{1, \frac{p\left(c^{(t)} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}\right)}{p\left(c^{(t-1)} \mid |\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}\right)} \cdot \frac{q\left(c^{(t-1)} - c_{\text{opt}}\right)}{q\left(c^{(t)} - c_{\text{opt}}\right)}\right\}. \tag{20}$$

After the initial burn-in period, we obtain the MCMC samples $\{(\mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}),\ t = 1, \dots, M\}$ which are next used to estimate the posterior gene inclusion probability by

$$\widehat{p}\left(\gamma_i = 1 \mid \mathbf{Y}, \mathbf{X}\right) = \frac{1}{M} \sum_{t=1}^{M} I\left(\gamma_i^{(t)} = 1\right), \tag{21}$$

and genes with higher posterior inclusion probabilities are considered more relevant to classification.

2.4. Classification. To assess the performance of our procedures, testing data sets are considered. For example, a testing set $(X_{\text{new}}, Y_{\text{new}})$ is available, and the predictive probability of Y_{new} given X_{new} is

$$\begin{aligned} p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}\right) = \int p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}, \mathbf{Z}, \boldsymbol{\gamma}, c\right) p\left(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}\right) d\left(\mathbf{Z}, \boldsymbol{\gamma}, c\right). \end{aligned} \tag{22}$$

Based on the MCMC samples, we estimate the probability with

$$\widehat{p}\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}\right) = \frac{1}{M} \sum_{t=1}^{M} p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}, \mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}\right). \tag{23}$$

When there are no testing sets available, we adopt the leave-one-out cross-validation (LOOCV) method to evaluate

Table 1: The posterior inclusion probability and description of the leading 20 genes for the colon cancer study. Genes identified in other studies were also noted.

Gene	Probability	Description
Z50753	0.1519	*H. sapiens* mRNA for GCAP-II/uroguanylin precursor[abc]
D14812	0.1303	Human mRNA for ORF, complete cds[bc]
H06524	0.1163	Gelsolin precursor, plasma (*Homo sapiens*)[ac]
R87126	0.1081	Myosin heavy chain, nonmuscle (*Gallus gallus*)[abc]
H08393	0.1012	Collagen alpha-2(XI) chain (*Homo sapiens*)[abc]
T62947	0.0987	60S ribosomal protein L24 (*Arabidopsis thaliana*)[abc]
T57882	0.0881	Myosin heavy chain, nonmuscle type A (*Homo sapiens*)[b]
R88740	0.0594	Atp synthase coupling factor 6, mitochondrial precursor (*Homo sapiens*)[bc]
J02854	0.0527	Myosin regulatory light chain 2, smooth muscle isoform (*Homo sapiens*); contains TAR1 repetitive element[ab]
T94579	0.0494	Human chitotriosidase precursor mRNA, complete cds[b]
H64807	0.0490	Placental folate transporter (*Homo sapiens*)[bc]
M59040	0.0439	Human cell adhesion molecule (CD44) mRNA, complete cds[c]
R55310	0.0437	S36390 mitochondrial processing peptidase[c]
M82919	0.0333	Human gamma aminobutyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds[bc]
H20709	0.0330	Myosin light chain alkali, smooth-muscle isoform (*Homo sapiens*)[bc]
T92451	0.0319	Tropomyosin, fibroblast, and epithelial muscle-type (*Homo sapiens*)[a]
R33481	0.0312	Transcription factors ATF-A and ATF-A-DELTA (*Homo sapiens*)[b]
L06175	0.0309	*Homo sapiens* P5-1 mRNA, complete cds
T64012	0.0309	Acetylcholine receptor protein, delta chain precursor (*xenopus laevis*)
H09719	0.0300	Tubulin alpha-6 chain (*Mus musculus*)

[a] Gene also identified in Ben-Dor et al. [38].
[b] Gene also identified in Furlanello et al. [39].
[c] Gene also identified in Chu et al. [40].

the performance with the training data. Because the predictive probability for Y_i is

$$p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}\right) = \left(\iiint p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}, c\right)^{-1} \times p\left(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}\right) d\mathbf{Z} d\boldsymbol{\gamma} dc \right)^{-1}, \quad (24)$$

where $\mathbf{Y}_{(-i)}$ denotes the vector of $\mathbf{Y}$ without the ith element. We estimate this probability based on the generated MCMC samples,

$$\widehat{p}\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}\right) = \frac{M}{\sum_{t=1}^{M} p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}, \mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}\right)^{-1}}. \quad (25)$$

3. Applications

In this section, we applied the fully Bayesian approach and the reference prior to three cancer studies: colon cancer, leukemia, and a large B-cell lymphoma (DLBCL) study [35–37]. We also compared the performance of this approach with other existing gene selection and classification methods. These data have been extensively studied with various methods but we only included a limited set of them. Others can be found in the reference lists of the work cited here.

3.1. Colon Cancer Study. The data of the colon cancer study contained 2000 expression levels from 40 tumor and 22 normal colon tissues. These expression levels were first transformed with a base 10 logarithmic function and then standardized to zero mean and unit variance for each gene. We then performed the MCMC sampler fixing the h in Σ_γ at 100 and $\pi_i = \Pr(\gamma_i = 1) = 0.005$ for all $i = 1, \ldots, p$. We burned in the first 12000 iterations, collected every 30th sample, and obtained 6700 posterior points in total for further analysis. The leading 20 genes with the largest posterior inclusion probabilities were presented in Table 1. This list was compared with the findings in three other studies [38–40] and similar findings were denoted in Table 1. The first 19 genes were identified in at least one of the three studies. For reference, Figure 1 displays the 100 largest posterior probabilities of the 100 corresponding genes.

For classification, we adopted the external leave-one-out cross-validation (LOOCV) procedure to evaluate the performance of classification with the selected genes. The procedures were the following: (i) removing one sample from the training set; (ii) ranking the genes in terms of t-statistics using the remaining samples and retaining the top 50 genes as the starting set to reduce computational burden; (iii) selecting the p^* most influential genes from the 50 genes based on our Bayesian method; and (iv) using these p^* genes to classify the previously removed sample. The procedures were repeated for each sample in the dataset. With different choices of p^* like $p^* = 6$, $p^* = 10$, and $p^* = 14$, the error rates were 0.1452, 0.1452, and 0.1129, respectively. The performance of other

Table 2: Performance comparison of different procedures with LOOCV for the colon cancer study.

Methods	No. of genes	LOOCV error rate	LOOCV accuracy
Bayesian g-prior	6	0.1452 (9/62)	0.8548 (53/62)
Bayesian g-prior	10	0.1452 (9/62)	0.8548 (53/62)
Bayesian g-prior	14	**0.1129 (7/62)**	**0.8871 (55/62)**
SVM[a]	1000	**0.0968 (6/62)**	**0.9032 (56/62)**
Classification tree[b]	200	0.1452 (9/62)	0.8548 (53/62)
1-Nearest-neighbor[b]	25	0.1452 (9/62)	0.8548 (53/62)
LogitBoost, estimated[b]	25	0.1935 (12/62)	0.8065 (50/62)
LogitBoost, 100 iterations[b]	10	0.1452 (9/62)	0.8548 (53/62)
AdaBoost, 100 iterations[b]	10	0.1613 (10/62)	0.8387 (52/62)
MAVE-LD[c]	50	0.1613 (10/62)	0.8387 (52/62)
IRWPLS[d]	20	**0.1129 (7/62)**	**0.8871 (55/62)**
SGLasso[e]	19	0.1290 (8/62)	0.8710 (54/62)
MRMS + SVM + D1[f]	5	0.1290 (8/62)	0.8710 (54/62)
MRMS + SVM + D2[f]	33	0.1452 (9/62)	0.8548 (53/62)
t-test + probit regression	6	0.1452 (9/62)	0.8548 (53/62)
t-test + probit regression	10	0.1774 (11/62)	0.8226 (51/62)
t-test + probit regression	14	0.2258 (14/62)	0.7742 (48/62)

[a] Proposed by Furey et al. [41].
[b] Proposed by Dettling and Bühlmann [42].
[c] Proposed by Antoniadis et al. [43].
[d] Proposed by Ding and Gentleman [44].
[e] Proposed by Ma et al. [45].
[f] Proposed by Maji and Paul [46].

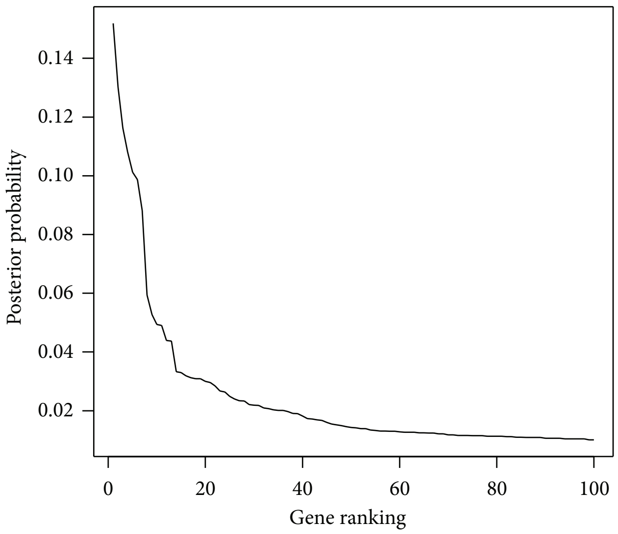


Figure 1: The largest 100 posterior probabilities of the genes for colon cancer study.

methods, including SVM [41]; classification tree followed by 1-Nearest-neighbor and LogitBoost with 100 iterations [42]; MAVE-LD [43]; IRWPLS [44]; supervised group Lasso (SGLasso, [45]) and MRMS [46]; and t-test for single markers in probit regression was summarized in Table 2. SVM had the smallest error rate, but it apparently included too many genes (1000 in this set). One other method MRMS+SVM+D1 performed better, with one more correct classification, than our proposed procedure when 6 or 10 genes were selected.

3.2. Leukemia Study. Next we considered the leukemia study with gene expression levels from 72 tissues including 47 acute lymphoblastic leukemia (ALL) patients and 25 acute myeloid leukemia (AML) subjects. These data contained 38 training and 34 testing samples. The training data contained 27 ALL cases and 11 AML cases, whereas the testing data were with 20 ALL cases and 14 AML cases. As described in other studies [2], the preprocessing steps such as thresholding and filtering were applied first and then followed by a base 10 logarithmic transformation. A total of 3571 genes were left for analysis. Next, we standardized the data across samples, and we ranked these genes by the same MCMC procedures described earlier. The top 20 genes with the largest posterior inclusion probabilities were presented in Table 3, and genes identified by other studies [36, 41, 47, 48] were also noted. For reference, Figure 2 displays the 100 largest posterior probabilities of the 100 corresponding genes.

For the classification procedure, similar to the procedures for colon cancer study, we selected p^* most influential genes from a starting set of 50 genes and next used them to examine the testing data. With $p^* = 6$, 10, or 14 genes, only the 61st and 66th observations were misclassified by our procedure. We also compared the results with weighted voting machine [36], MAVE-LD [43], two-step EBM [47], KIGP + PK [48], and t-test for single markers with probit regression, as summarized in Table 4. Note that although MAVE-LD and two-step EBM methods performed better than our proposed

Table 3: The posterior inclusion probability and description of the leading 20 genes for the leukemia study. Genes identified in other studies were also noted.

Gene	Probability	Description
X95735	0.0691	Zyxin[abc]
M27891	0.0519	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)[abc]
M23197	0.0302	CD33 cD33 antigen (differentiation antigen)[abc]
Y12670	0.0251	LEPR leptin receptor[a]
X85116	0.0226	Epb72 gene exon 1[ab]
D88422	0.0196	CYSTATIN A[bc]
X62654	0.0196	ME491 gene extracted from *H. sapiens* gene for Me491/CD63 antigen[b]
X04085	0.0195	Catalase (EC 1.11.1.6) 5′flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)[a]
L09209	0.0195	APLP2 amyloid beta (A4) precursor-like protein 2[bc]
HG1612-HT1612	0.0186	Macmarcks[bc]
M16038	0.0186	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog[abc]
U50136	0.0181	Leukotriene C4 synthase (LTC4S) gene[ab]
M55150	0.0172	FAH fumarylacetoacetate[ab]
M92287	0.0172	CCND3 cyclin D3[bc]
M22960	0.0168	PPGB protective protein for beta-galactosidase (galactosialidosis)[bc]
X70297	0.0168	CHRNA7 cholinergic receptor, nicotinic, and alpha polypeptide 7[b]
X51521	0.0163	VIL2 Villin 2 (ezrin)[b]
M63138	0.0154	CTSD cathepsin D (lysosomal aspartyl protease)[ab]
M27783	0.0154	ELA2 elastase 2, neutrophil[c]
U81554	0.0137	CaM kinase II isoform mRNA

[a] Gene also identified in Golub et al. [36].

[b] Gene also identified in Ben-Dor et al. [38].

[c] Gene also identified in in Lee et al. [22].

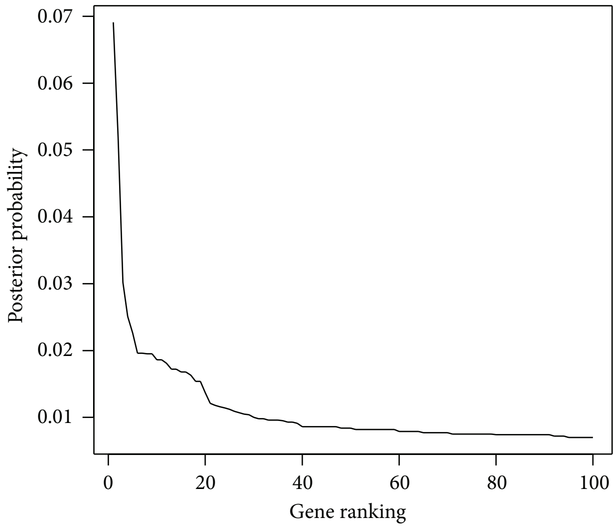


Figure 2: The largest 100 posterior probabilities of the genes for leukemia study.

procedure, both methods used more genes (50 and 512) and yet achieved only one less misclassification. Among this list, our procedure apparently considered a smaller set of genes with a satisfactory performance.

3.3. Diffuse Large B-Cell Lymphoma (DLBCL) Study. This study collected 58 samples from DLBCL patients and 19 samples from follicular lymphoma [37]. The original dataset contained 7129 genes. After the preprocessing steps such as thresholding and filtering were applied and a base 10 logarithmic transformation was conducted, a total of 6285 genes were left for analysis. Next, we standardized the data across samples and ranked these genes by the same MCMC procedures described in earlier sections. The error rates for $p^* = 6$, 10, or 14 under LOOCV were 0.0519, 0.0649, and 0.0779, and the accuracy was between 0.92 and 0.95, as listed in Table 5. To achieve a smaller error rate, we considered $p^* = 5$ and obtained a smaller rate 0.0390, the same rate achieved by the hyperbox enclosure (HBE) method [49]. Similar to the discussion in the previous two applications, our proposed model can achieve the same or smaller error rate with a smaller set of genes.

4. Conclusion and Discussion

In this Bayesian framework, we considered a mixture of g-prior to complete a fully Bayesian analysis for gene selection and cancer classification. Different from other existing methods that treated the c as a fixed value, we incorporated its uncertainty by assuming a reference inverse-gamma prior distribution. Earlier studies mentioned this prior, but considered it difficult to derive posterior inference. We therefore

Table 4: Performance comparison of different procedures for the leukemia study.

Methods	No. of genes	Testing error rate	Testing accuracy
Bayesian g-prior	6	**0.0588 (2/34)**	**0.9412 (32/34)**
Bayesian g-prior	10	0.0588 (2/34)	0.9412 (32/34)
Bayesian g-prior	14	0.0588 (2/34)	0.9412 (32/34)
Weighted voting machine[a]	50	0.1471 (5/34)	0.8529 (29/34)
MAVE-LD[b]	50	**0.0294 (1/34)**	**0.9706 (33/34)**
Two-step EBM[c]	32	0.1471 (5/34)	0.8529 (29/34)
Two-step EBM[c]	256	0.0588 (2/34)	0.9412 (32/34)
Two-step EBM[c]	512	**0.0294 (1/34)**	**0.9706 (33/34)**
KIGP + PK[d]	20	0.0588 (2/34)	0.9412 (32/34)
t-test + probit regression	6	0.1765 (6/34)	0.8235 (28/34)
t-test + probit regression	10	0.0882 (3/34)	0.9118 (31/34)
t-test + probit regression	14	0.1176 (4/34)	0.8824 (30/34)

[a] Proposed by Gloub et al. [36].
[b] Proposed by Antoniadis et al. [43].
[c] Proposed by Ji et al. [47].
[d] Proposed by Zhao and Cheung [48].

Table 5: Performance comparison of different procedures with LOOCV for the colon cancer study.

Methods	No. of genes	LOOCV error rate	LOOCV accuracy
Bayesian g-prior	5	0.0390 (3/77)	0.9610 (74/77)
Bayesian g-prior	6	0.0519 (4/77)	0.9481 (73/77)
Bayesian g-prior	10	0.0649 (5/77)	0.9351 (72/77)
Bayesian g-prior	14	0.0779 (6/77)	0.9221 (71/77)
Bayesian g-prior	20	0.0779 (6/77)	0.9221 (71/77)
HBE	6	0.0390 (3/77)	0.9610 (74/77)
t-test + probit regression	6	0.1169 (9/77)	0.8831 (68/77)
t-test + probit regression	10	0.1558 (12/77)	0.8442 (65/77)
t-test + probit regression	14	0.2208 (17/77)	0.7792 (60/77)

outlined the implementation for computation under this model setting for future applications. This approach is more flexible in the process of model building. This model is able to evaluate how influential a gene can be with posterior probabilities that can be used next for variable selection. Such an approach is useful in biomedical interpretations for the selection of relevant genes for disease of interest. When compared with other existing methods, our proposed procedure achieves a better or comparable accurate rate in classification with fewer genes. In the analyses of colon cancer and leukemia studies, we replicate several relevant genes identified by other research groups. The findings have accumulated evidence for further laboratory research.

In the application section, we listed only the results from $p^* = 6$, 10, and 14 selected genes. Other values for p^* have been tried and the performance remains good. For instance, the pink line in Figures 3 and 4 displays the accuracy of the proposed procedure when the number of selected genes p^* varies between 5 and 20 for the colon cancer and leukemia study, respectively. For the colon cancer study, the largest accuracy 0.8871 occurs at $p^* = 14$, while other values of p^* lead to the accuracy between 0.8387 and 0.8871. These

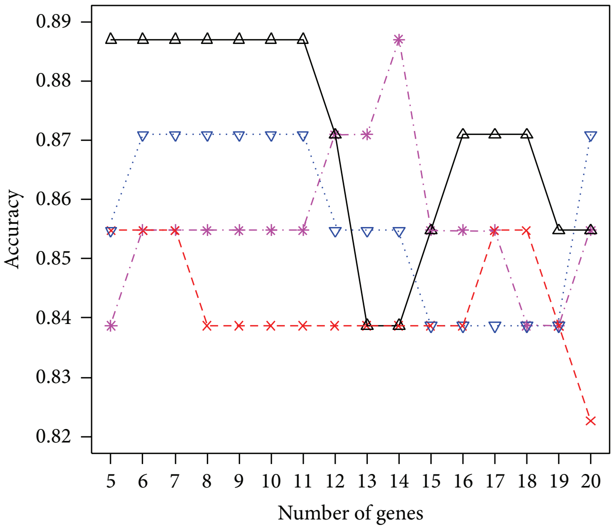


Figure 3: The accuracy of the proposed procedure at different numbers ($p^* = 5, \ldots, 20$) of selected genes with c following the generalized g-prior (pink line) or fixed at constant 5 (red line), 10 (blue), or 20 (black) for the colon cancer study.

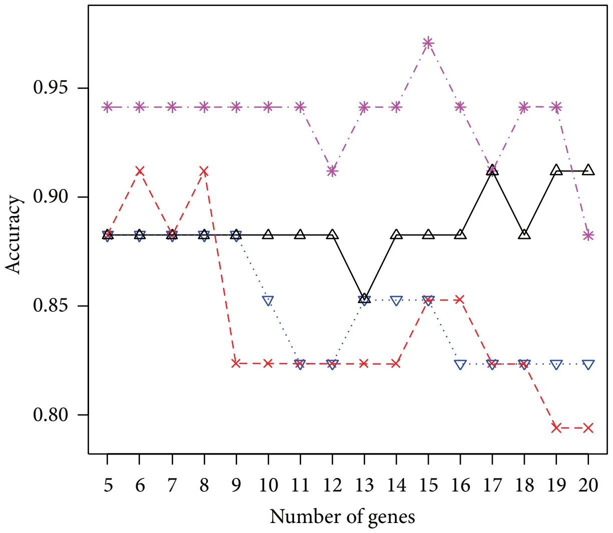


FIGURE 4: The accuracy of the proposed procedure at different numbers ($p^* = 5, \ldots, 20$) of selected genes with c following the generalized g-prior (pink line) or fixed at constant 5 (red line), 10 (blue), or 20 (black) for the leukemia study.

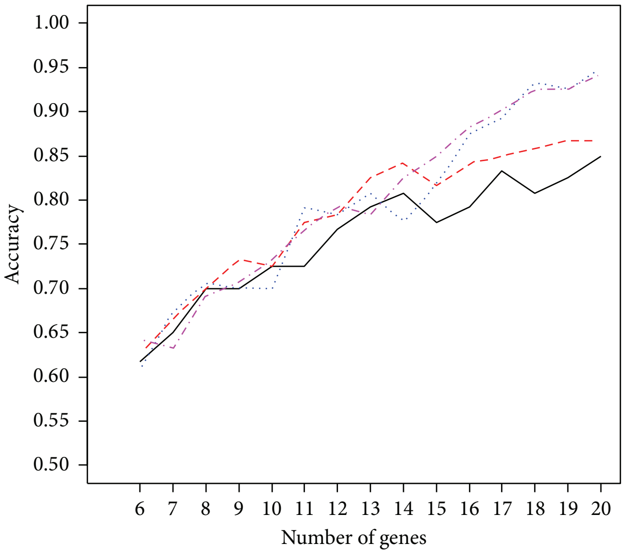


FIGURE 5: Average accuracy when the number of genes ranges from 1 to 15 under the mixtures of g-priors on c (pink line), c fixed at 5 (black), c at 50 (red), and c at 500 (blue).

correspond to at least 52 correctly identified subjects out of 62. For the leukemia study, the largest accuracy 0.9706 occurs at $p^* = 15$. Other values of p^* all lead to an accuracy larger than 90% except when $p^* = 20$ (accuracy is $0.8824 = 30/34$). In addition, we compared the results under the proposed generalized g-prior with c fixed at a constant. The colored lines in Figures 3 and 4 are for c fixed at 5 (red line), 10 (blue), or 20 (black), respectively. Again, results under the prior distribution assumption lead to a higher accuracy with a less number of selected genes. Another issue is related to the choice of the number of genes in the starting set. We have considered 50 in all three applications. This value can certainly be changed. However, the computational complexity increased as the value becomes larger. This cost in computation remains a research topic for future research.

To compare the performance of a stochastic c and a constant c, we also conducted a small simulation study to investigate the effect of assigning a prior on c versus fixing c at different constant values. We used the R package penalizedSVM [50, 51] to simulate three data sets; each contains 500 genes with 15 genes associated with the disease. The numbers of training and testing sample were 200 and 40, respectively. We then conducted the gene selection procedures with a prior on c, $c = 5$, $c = 50$, and $c = 500$ at $p^* = 1, 2, \ldots, 15$ and recorded the accuracy under each setting. Figure 5 plots the average accuracy with the pink line standing for the accuracy under the mixtures of g-priors on c, the black line for $c = 5$, the red line for $c = 50$, and the blue line for $c = 500$. It can be observed that only when c is assigned with a very large number like 500, the corresponding accuracy can be slightly better than that under a prior for the uncertainty in c. This again supports the use of the mixtures of g-priors for a better and robust result.

Here in this paper we have focused on the analysis of binary data. However, the probit regression model can be extended to a multinomial probit model to solve the multiclass problems, and the Bayesian inference can be carried out similarly. Such analysis will involve a larger computational load and further research in this direction is needed. Another point worth mentioning is the inclusion of interactions between genes. Further research can incorporate a power prior into the prior of $\boldsymbol{\gamma}$ [52] or include information on gene-gene network structure [18] to complete the procedure for variable selection.

Acknowledgment

Part of this research was supported by NSC 100-2314-B-002-107-MY3.

References

[1] V. T. Chu, R. Gottardo, A. E. Raftery, R. E. Bumgarner, and K. Y. Yeung, "MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis," *Genome Biology*, vol. 9, no. 7, article R118, 2008.

[2] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

[3] A. Hirakawa, Y. Sato, D. Hamada, and I. Yoshimura, "A new test statistic based on shrunken sample variance for identifying differentially expressed genes in small microarray experiments," *Bioinformatics and Biology Insights*, vol. 2, pp. 145–156, 2008.

[4] W. Pan, J. Lin, and C. T. Le, "A mixture model approach to detecting differentially expressed genes with microarray data," *Functional and Integrative Genomics*, vol. 3, no. 3, pp. 117–124, 2003.

[5] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.

[6] A. Gusnanto, A. Ploner, F. Shuweihdi, and Y. Pawitan, "Partial least squares and logistic regression random-effects estimates for gene selection in supervised classification of gene expression data," *Journal of Biomedical Informatics*, vol. 4, pp. 697–709, 2013.

[7] Y. Liang, C. Liu, X. Z. Luan et al., "Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification," *BMC Bioinformatics*, vol. 14, article 198, 2013.

[8] G.-Z. Li, H.-L. Bu, M. Q. Yang, X.-Q. Zeng, and J. Y. Yang, "Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis," *BMC Genomics*, vol. 9, no. 2, article S24, 2008.

[9] A. Wang and E. A. Gehan, "Gene selection for microarray data analysis using principal component analysis," *Statistics in Medicine*, vol. 24, no. 13, pp. 2069–2087, 2005.

[10] S. Bicciato, A. Luchini, and C. Di Bello, "PCA disjoint models for multiclass cancer analysis using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 571–578, 2003.

[11] X. Q. Zeng, G. Z. Li, M. Q. Yang, G. F. Wu, and J. Y. Yang, "Orthogonal projection weights in dimension reduction based on partial least squares," *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, vol. 1, pp. 100–115, 2009.

[12] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 32–44, 2007.

[13] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[14] J. X. Liu, Y. Xu, C. H. Zheng, Y. Wang, and J. Y. Yang, "Characteristic gene selection via weighting principal components by singular values," *PLoS ONE*, vol. 7, no. 7, Article ID e38873, 2012.

[15] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biology Direct*, vol. 7, article 33, 2012.

[16] T. Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, vol. 3, no. 4, pp. 1–17, 2002.

[17] Y. Wang, I. V. Tetko, M. A. Hall et al., "Gene selection from microarray data for cancer classification—a machine learning approach," *Computational Biology and Chemistry*, vol. 29, no. 1, pp. 37–46, 2005.

[18] F. C. Stingo and M. Vannucci, "Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data," *Bioinformatics*, vol. 27, no. 4, pp. 495–501, 2011.

[19] J. G. Ibrahim, M.-H. Chen, and R. J. Gray, "Bayesian models for gene expression with DNA microarray data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 88–99, 2002.

[20] Y.-C. Wei, S.-H. Wen, P.-C. Chen, C.-H. Wang, and C. K. Hsiao, "A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies," *European Journal of Human Genetics*, vol. 18, no. 8, pp. 942–947, 2010.

[21] B. Peng, D. Zhu, and B. P. Ander, "An Integrative Framework for Bayesian variable selection with informative priors for identifying genes and pathways," *PLoS ONE*, vol. 8, no. 7, Article ID 0067672, 2013.

[22] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: a Bayesian variable selection approach," *Bioinformatics*, vol. 19, no. 1, pp. 90–97, 2003.

[23] N. Sha, M. Vannucci, M. G. Tadesse et al., "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, vol. 60, no. 3, pp. 812–819, 2004.

[24] X. Zhou, K.-Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 249–259, 2004.

[25] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[26] A. Zellner, "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*, pp. 233–243, North-Holland, Amsterdam, The Netherlands, 1986.

[27] A.-J. Yang and X.-Y. Song, "Bayesian variable selection for disease classification using gene expression data," *Bioinformatics*, vol. 26, no. 2, pp. 215–222, 2010.

[28] M. Baragatti and D. Pommeret, "A study of variable selection using g-prior distribution with ridge parameter," *Computational Statistics and Data Analysis*, vol. 56, no. 6, pp. 1920–1934, 2012.

[29] E. Leya and M. F. J. Steel, "Mixtures of g-priors for Bayesian model averaging with economic applications," *Journal of Econometrics*, vol. 171, no. 2, pp. 251–266, 2012.

[30] M. Smith and R. Kohn, "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, vol. 75, no. 2, pp. 317–343, 1996.

[31] E. I. George and D. P. Foster, "Calibration and empirical bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.

[32] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 410–423, 2008.

[33] W. Cui and E. I. George, "Empirical Bayes versus fully Bayes variable selection," *Journal of Statistical Planning and Inference*, vol. 138, no. 4, pp. 888–900, 2008.

[34] C. P. Robert, "Convergence control methods for Markov chain Monte Carlo algorithms," *Statistical Science*, vol. 10, pp. 231–253, 1995.

[35] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[36] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[37] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[38] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559–583, 2000.

[39] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "Entropy-based gene ranking without selection bias for the predictive classification of microarray data," *BMC Bioinformatics*, vol. 4, article 54, 2003.

[40] W. Chu, Z. Ghahramani, F. Falciani, and D. L. Wild, "Biomarker discovery in microarray gene expression data with Gaussian processes," *Bioinformatics*, vol. 21, no. 16, pp. 3385–3393, 2005.

[41] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[42] M. Dettling and P. Bühlmann, "Boosting for tumor classification with gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 1061–1069, 2003.

[43] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 563–570, 2003.

[44] B. Ding and R. Gentleman, "Classification Using Generalized Partial Least Squares," Bioconductor Project Working Papers, 2004, http://www.bepress.com/bioconductor /paper5.

[45] S. Ma, X. Song, and J. Huang, "Supervised group Lasso with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, article 60, 2007.

[46] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray data," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.

[47] Y. Ji, K.-W. Tsui, and K. Kim, "A novel means of using gene clusters in a two-step empirical Bayes method for predicting classes of samples," *Bioinformatics*, vol. 21, no. 7, pp. 1055–1061, 2005.

[48] X. Zhao and L. W.-K. Cheung, "Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data," *BMC Bioinformatics*, vol. 8, article 67, 2007.

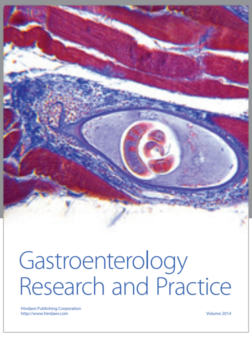
[49] O. Dagliyan, F. Uney-Yuksektepe, I. H. Kavakli, and M. Turkay, "Optimization based tumor classification from microarray gene expression data," *PLoS ONE*, vol. 6, no. 2, Article ID e14579, 2011.

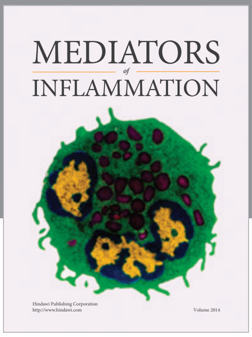
[50] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[51] G. M. Fung and O. L. Mangasarian, "A feature selection Newton method for support vector machine classification," *Computational Optimization and Applications*, vol. 28, no. 2, pp. 185–202, 2004.

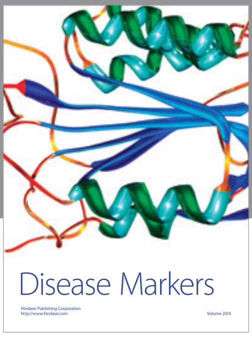
[52] A. Krishna, H. D. Bondell, and S. K. Ghosh, "Bayesian variable selection using an adaptive powered correlation prior," *Journal of Statistical Planning and Inference*, vol. 139, no. 8, pp. 2665–2674, 2009.

The Scientific World Journal

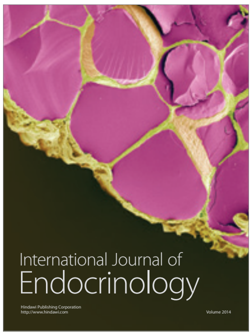

Gastroenterology Research and Practice


MEDIATORS of INFLAMMATION

Journal of Diabetes Research


Disease Markers

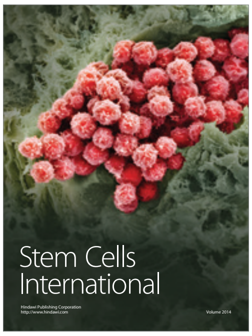
Journal of Immunology Research

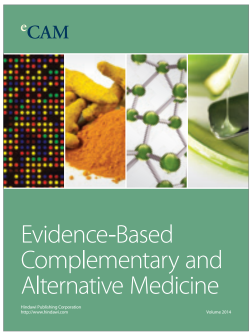

International Journal of Endocrinology

PPAR Research

BioMed Research International

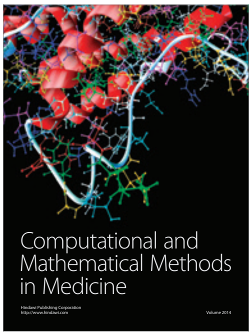
Journal of Ophthalmology

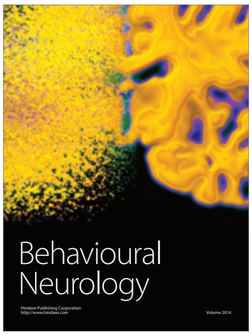

Stem Cells International

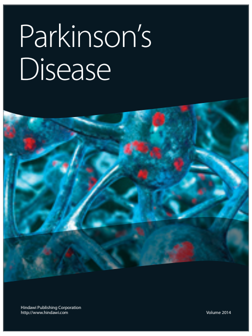

eCAM
Evidence-Based Complementary and Alternative Medicine

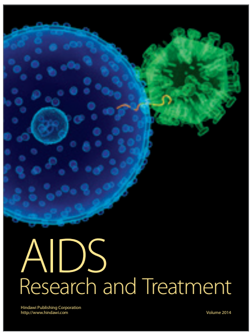
Journal of Obesity

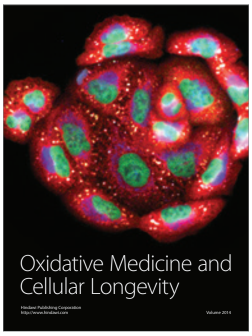
Journal of Oncology


Computational and Mathematical Methods in Medicine


Behavioural Neurology


Parkinson's Disease


AIDS Research and Treatment


Oxidative Medicine and Cellular Longevity