Bayesian Variable Selection for High Dimensional Data Analysis

YANG, Aijun

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

The Chinese University of Hong Kong
June 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Thesis/Assessment Committee

Professor LEE Sik-Yum (Chair)

Professor SONG Xin-Yuan (Thesis Supervisor)

Professor CHEUNG Siu-Hung(Committee Member)

Professor TANG Man-Lai (External Examiner)

Abstract of thesis entitled:

Bayesian Variable Selection for High Dimensional Data Analysis

Submitted by YANG, Aijun

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in May 2010

In the practice of statistical modeling, it is often desirable to have an accurate predictive model. Modern data sets usually have a large number of predictors. For example, DNA microarray gene expression data usually have the characteristics of fewer observations and larger number of variables. Hence parsimony is especially an important issue. Best-subset selection is a conventional method of variable selection. Due to the large number of variables with relatively small sample size and severe collinearity among the variables, standard statistical methods for selecting relevant variables often face difficulties.

The second part of the thesis proposes a Bayesian stochastic variable selection approach for gene selection based on a probit regression model with a generalized singular g-prior distribution for regression coefficients. Using simulation-based MCMC methods for simulating parameters from the posterior distribution, an efficient and dependable algorithm is implemented. It is also shown that this algorithm is robust to the choice of initial values, and produces posterior probabilities of related genes for biological interpretation. The performance of the proposed approach is compared with other popular methods in gene selection and classification via the well known colon cancer and

leukemia data sets in microarray literature.

In the third part of the thesis, we propose a Bayesian stochastic search variable selection approach for multi-class classification, which can identify relevant genes by assessing sets of genes jointly. We consider a multinomial probit model with a generalized g-prior for the regression coefficients. An efficient algorithm using simulation-based MCMC methods are developed for simulating parameters from the posterior distribution. This algorithm is robust to the choice of initial value, and produces posterior probabilities of relevant genes for biological interpretation. We demonstrate the performance of the approach with two well-known gene expression profiling data: leukemia data and lymphoma data. Compared with other classification approaches, our approach selects smaller numbers of relevant genes and obtains competitive classification accuracy based on obtained results.

The last part of the thesis is about the further research, which presents a stochastic variable selection approach with different two-level hierarchical prior distributions. 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 can be developed and implemented.

摘要

在統計建模的實際應用中，通常希望有准確地預測模型。現代數據集通常具有很多的變量。例如，DNA 微陣列基因數據就是有少量的觀測值和大量的變量。因此，模型的簡約性是一個很重要的問題。最佳變量選擇是一種傳統的變量選擇方法。但是由於觀測值少，變量多以及變量之間強相關性的原因，傳統選擇重要變量的方法經常面臨困難。

基於 probit 模型和對回歸系數指定奇異 g-prior 分布，論文的第二部分提出使用貝葉斯變量隨机選擇的方法來選擇重要的基因。基於模擬的 MCMC 方法，我們使用一種有效並且可靠的算法來從後驗分布中抽取樣本。結果顯示這種算法對初始值的選取非常穩健。並且得到每個基因被包含的後驗概率。這些後驗概率可以用於從生物角度的解釋。通過對微陣列文獻中的結腸癌和血癌數據的分析，從選擇的基因和分類的結果將所提出的方法和其他方法進行了比較。

在論文的第三個部分，我們同樣提出使用貝葉斯變量隨机選擇的方法來選擇重要的基因，並且對多分類數據進行分類。這個部分主要基於 multinomial probit 模型和對回歸系數指定廣義 g-prior 分布。基於模擬的 MCMC 方法，我們使用一種有效並且可靠的算法來從後驗分布中抽取樣本。結果顯示這種算法對初始值的選取非常穩健。我們主要通過兩組基因數據-血癌和淋巴瘤數據-來說明我們方法的表現。結果顯示：同其他方法相比較而言，我們的方法可以利用更少的基因來得到具有競爭力的結果。

論文的最後一個部分是關於將來的研究。我們考慮使用基於兩層結構的先驗分布函數的貝葉斯變量隨機選擇的方法來選擇重要基因。這種具有兩層結構的先驗分布擁有更加離散的特點。基於模擬的 MCMC 方法，我們可以使用一種有效並且可靠的算法來從後驗分布中抽取樣本。

Acknowledgement

I owe a debt of thanks to Professor Xin-Yuan Song for her generosity of supervision and encouragement during the course of this research program. I also wish to take this opportunity to express my great appreciation to Professor Sik-Yum Lee for his invaluable advice and helpful comments.

I wish to thank other faculty members in the statistics department for their excellent lectures. I thank my friends for their sincere help. I will always cherish our friendship.

I would like to dedicate this thesis to my family for their boundless love. Finally, I want to thank my wife Ye Chen. Without her love and support, this work would not have been completed. I also have so many thanks to my father-in-law and mother-in-law for their everlasting support and encouragement.

Contents

Abstract i

Acknowledgement iv

1 Introduction 1

2 Bayesian Variable Selection for Disease Classification Using Gene Expression Data 5

2.1 Introduction 5
2.2 Method 8
2.2.1 Model 8
2.2.2 Prior Distribution 10
2.2.3 Computation 12
2.2.4 Classification 15
2.3 Results 16
2.3.1 Colon Cancer Data 17
2.3.2 Leukemia Dataset 20
2.3.3 Computational Time 22
2.4 Discussion 22

3 Multi-class Classification via Bayesian Variable Selection with Gene Expression Data 30

3.1 Introduction 30
3.2 Matrix Variate Distribution 33
3.3 Method 34

3.3.1 Model 34
3.3.2 Prior Specification 36
3.3.3 Computation 38
3.3.4 Classification 41
3.3.5 Misclassification 42
3.4 Real Data Analysis 43
3.4.1 Leukemia Data 43
3.4.2 Lymphoma Data 47
3.4.3 Computational Time 48
3.5 Discussion 48

4 Sparse Bayesian Variable Selection for Classifying High-dimensional Microarray Data **56**
4.1 Introduction 56
4.2 Methods 59
4.2.1 Model 59
4.2.2 Prior Distribution 60
4.2.3 Computation 61
4.2.4 Classification 66

5 Summary and Discussion **67**

A **69**
A.1 Method 69
A.2 Results 74

B **75**
B.1 Matrix Variate Distribution 75
B.2 Method 76

C **82**
C.1 Method 82

Bibliography **89**

List of Figures

2.1 Fig.2.1(a) shows the gene inclusion probabilities (in percentages) versus the gene index, Fig.2.1(b) and Fig.2.1(c) show the number of selected genes and the log relative posterior probabilities of selected genes versus the first 10000 iteration number, respectively. 24

2.2 Fig.2.2 shows the gene inclusion probabilities (in percentages) versus the gene index for leukemia data. 25

3.1 Fig.3.1(a) shows the gene inclusion probabilities (in percentages) versus the gene index, Fig.3.1(b) and Fig.3.1(c) show the number of selected genes and the log relative posterior probabilities of selected genes versus the first 10000 iteration number, respectively. 50

3.2 Genes that distinguish ALL-B, ALL-T and AML. Each column corresponds to a sample array and each row corresponds to a gene. The heat map is generated by using Matrix2png softerware (Pavlidis and Noble, 2003). Genes with expression levels greater than the mean are colored in red and those below the mean are colored in green. 51

3.3 Fig.3.3 shows the gene inclusion probabilities (in percentages) versus the gene index. 51

List of Tables

2.1 Colon cancer data: strongly significant genes for classifying normal and tumor tissues. 26
2.2 Comparison of LOOCV performance of different approaches for Colon cancer data. 27
2.3 Leukemia data: strongly significant genes for discriminating ALL and AML samples. 28
2.4 The comparison of classification methods for the leukemia data. 29

3.1 Significant genes found for discriminating ALL-T, ALL-B and AML. 52
3.2 Error rate results for the training data and test data of Leukemia data, respectively. 53
3.3 The comparison of classification results for Leukemia test data. 53
3.4 Significant genes found for classifying DLBCL, CLL, and FL. 54
3.5 Comparison of LOOCV results of different methods for Lymphoma data. 55

Chapter 1

Introduction

In the practice of statistical modeling, it is often desirable to have an accurate predictive model with a sparse representation. Modern data usually have a large number of predictors. For example, duo to recent advances in information technology, it is possible to access thousands of macroeconomic time series, which have been shown the "value" for signal extraction and forecasting. This is not an issue of mere academic interest. Lar Svensson (2005) described what central bankers do in practice: "Large amounts data about the state of the economy and the rest of world...are collected, processed, and analyzed before each major decision." In an effort to assist in this task, researchers recently have proposed new methods to handle large data sets in the econometrics of forecasting. Also DNA microarray gene expression data usually have the characteristics of fewer observations and larger number of variables. Hence parsimony is especially an important issue. Best-subset selection is a conventional method of variable selection. Standard statistical methods for selecting relevant variables often face difficulties due to the small sample size as it can create an unreliable selection process.

Bayesian stochastic search variable selection is a model- based approach for studying regression models that relate a response y to a vector of candidate explanatory variables $x = (x_1, \cdots, x_p)^T$. In generalized linear models, both the density of y and the mean

function of y conditional on x depend on a linear combination $x^T\beta$ through the regression coefficients $\beta = (\beta_1, \cdots, \beta_p)^T$. Rather than fixing the dimension (the number of selected genes), the SSVS approach uses priors that propose different model γ's and the corresponding sets of regression coefficient β'_γs, where γ indicates the components of covariates that are included in the regression model. This creates additional flexibility as well as the ability to impose a constraint by limiting the dimension. Therefore, the prior works as a penalty to create this constraint.

Bayesian stochastic search variable selection has gained much empirical success in a variety of applications. For example, SSVS is used in basis selection for nonparametric regression (e.g., Smith and Kohn 1996) and in construction of financial index tracking portfolios (e.g., George and McCulloch 1997). Recently, SSVS has been applied to the area of bioinformatics. Lee et al. (2003) applied their multivariate gene selection to microarray data with two classes. Sha et al. (2004) and Zhou et al. (2006) extended the underlying theory to multiple classes data. The multivariate Bayesian model of Lee et al. (2003) and Zhou et al. (2006) used the g-prior (Zellner, 1986) for unknown parameters of regression coefficients associated with the covariates (related genes). However, for situations with high-dimensional covariates, or highly collinear covariates, the covariance matrix involved in the g-prior is nearly singular (Gupta and Ibrahim, 2007), and results in unstable convergence of the algorithm. Moreover, due to the complicated structure of high dimensional distribution, convergence of the algorithm is slow in general. Sha et al. (2004) proposed an algorithm that is based on a multinomial probit model by using adding/deleting and swapping algorithm. According to Lamnisos et al. (2009), this kind of algorithm that randomly chooses to either add or delete a single explanatory variable, or to swap two explanatory variables in the model often leads to high model acceptance rates

when the number of variables is substantially larger than the sample size. Moreover, the Metropolis random walk suggested by Sha et al. (2004) with local proposals and high acceptance rate is often associated with the poor mixing of MCMC chains. Furthermore, as their approach did not capture a priori correlation in the parameters, eliciting a prior covariance matrix with $p > n$ is difficult (Gupta and Ibrahim, 2009). Finally, both Sha et al. (2004) and Zhou et al. (2006) calculated the leave one out cross validation (LOOCV) within the gene selection process. According to Ambroise and McLachlan (2002) and Rocke et al. (2009), a selection bias that optimizes the classification accuracy exists when this internal LOOCV procedure is applied to estimate the prediction error.

Chapter 2 proposes a Bayesian stochastic variable selection approach for gene selection based on a probit regression model with a generalized singular g-prior distribution for regression coefficients. Using simulation-based MCMC methods for simulating parameters from the posterior distribution, an efficient and dependable algorithm is implemented. It is also shown that this algorithm is robust to the choices of initial values, and produces posterior probabilities of related genes for biological interpretation. The performance of the proposed approach is compared with those of other popular methods in gene selection and classification via the well known colon cancer and leukemia data sets in microarray literature.

In Chapter 3, we propose a Bayesian stochastic search variable selection approach for multi-class classification, which can identify relevant genes by assessing sets of genes jointly. We consider a multinomial probit model with a generalized g-prior for the regression coefficients. An efficient algorithm using simulation-based MCMC methods is developed for simulating parameters from the posterior distribution. This algorithm is robust to the choice of initial value, and produces posterior probabilities of

relevant genes for biological interpretation. We demonstrate the performance of the approach with two well-known gene expression profiling data: leukemia data and lymphoma data. Compared with other classification approaches, our approach selects smaller numbers of relevant genes and obtains competitive classification accuracy based on obtained results.

Chapter 4 is about the further research, which presents a stochastic variable selection approach with different two-level hierarchical prior distributions. 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.

□ **End of chapter.**

Chapter 2

Bayesian Variable Selection for Disease Classification Using Gene Expression Data

2.1 Introduction

Class prediction has recently received much attention in the context of DNA microarrays. Its main objective is to classify and predict the diagnostic category of a sample based on its gene expression profile. This problem is challenging because the number of genes is usually much larger than the number of samples available, and only a small subset of genes is relevant in classification. Thus, a critical issue is the identification of genes that contribute most to the classification. Moreover, as emphasized by Dougherty (2001), Li et al. (2002), and Yeung et al. (2005), a small number of relevant genes is essential.

In the past decade, many gene selection approaches have been proposed in the literature. In some published studies, the number of selected genes is large; for example, 2000 genes (Alon et al., 1999), and 1000 or 2000 genes (Furey et al., 2000). Even after performing gene selection, the numbers of selected genes in certain studies are still large compared to the numbers of samples; for example, 50 genes (Golub et al., 1999), 51 genes

(Hendenfalk et al., 1999), 25 to 1000 genes (Furey et al., 2000), 96 genes (Khan et al., 2001), and 231 to 549 genes (Antonov et al., 2004).

In addition, several methods for reducing the number of genes to be considered before using appropriate classification, are univariate methods in the sense that each relevant gene is considered individually. Examples include the weighted voting scheme (Golub et al., 1999), the mixture model algorithm (Pan, 2002), the partial least squares (PLS) (Nguyen and Rocke, 2002), nonparametric methods (Troyanskaya et al., 2002), and the Wilcoxon test statistic (Dettling, 2004). To take into account the dependency between genes for achieving a reduced number of relevant genes, multivariate gene selection procedures, which consider multiple genes simultaneously, have been proposed by Bo and Jonassen (2002), and Jaeger et al. (2003), among others. The Bayesian stochastic search variable selection method (George and McCulloch, 1993) has recently become popular (see Lee et al., 2003, Gupta and Ibrahim, 2007; among others). The multivariate Bayesian model of Lee et al. (2003) used the g-prior (Zellner, 1986) for unknown parameters of regression coefficients associated with the covariates (related genes). However, for situations with high-dimensional covariates, or highly collinear covariates, the covariance matrix involved in the g-prior is nearly singular (Gupta and Ibrahim, 2007), and results in unstable convergence of the algorithm. Moreover, due to the complicated structure of high dimensional distribution, convergence of the algorithm is slow in general. Bae and Mallick (2004) introduced a two level hierarchical Bayesian model with different priors that favor sparseness in terms of number of genes used. They identified the significant genes using the posterior variances of the regression coefficients. However, their methods did not produce the posterior probabilities, which are useful for biomedical interpretation, for the selected genes. Some recent contributions

in the selection of genes for multiclass classification and other important problems can be found in McLachlan et al. (2004, 2008), Le Cao et al. (2008), Le Cao and Chabrier (2008), Rocke et al. (2009), and references therein.

In this chapter, we consider a multivariate Bayesian regression model together with a stochastic search variable selection (SSVS) method for gene selection and classification of diagnostic category. To overcome the problem induced by the possible singularity of the covariance matrix involved in the g-prior distribution of the regression coefficients, we propose a generalized singular g-prior (gsg-prior) on the basis of the Moore-Penrose generalized inverse of matrices. This kind of gsg-prior has been found to be effective for similar statistical problems with large number of genes and small number of samples (West, 2000). Moreover, unlike the method based on approximation, we perform full Bayesian analysis through the Markov chain Monte Carlo (MCMC; Gilks et al., 1996) based on a stochastic search algorithm. In developing our gsg-SSVS algorithm, the efficient sampling scheme suggested by Panagiotelisa and Smith (2008) is implemented. For the posterior analysis associated with this sampling scheme, the unknown intercept and regression coefficients in the Bayesian regression model are integrated out from the joint posterior distribution. This gives a simple and well defined posterior distribution to ensure stable convergence of the resulting MCMC methods. As a result, our algorithm is computationally more stable and efficient compared to the MCMC algorithm in Lee et al. (2003). In addition, the gsg-SSVS approach produces the posterior probabilities for the selected genes, which are 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; McLachlan et al, 2004; Ma et al., 2007; Le Cao et al., 2008; Le Cao and Chabrier, 2008; among others). Our results show that the proposed gsg-SSVS approach reduces the number of selected genes and produces prediction accuracy comparable to those of the existing variable selection and classification methods.

This chapter is organized as follows. In the Method section, we briefly review the model specification based on stochastic search variable selection; we also discuss the related prior distributions and the implementation of the Bayesian method. Discussions on classification are also presented in this section. Results obtained from the analyses of the two published data sets are given in the Results section. Some concluding remarks are presented in the Discussion section. The technical details are provided in Appendix A.

2.2 Method

2.2.1 Model

Suppose that n independent binary random variables $Y_1, \cdots, Y_n$ are observed. 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, the expression levels for a set of genes were measured; hence we have the following data matrix $\mathbf{X}$ of covariates:

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

We define a probit type regression model as $p_i = P(Y_i = 1) = \Phi(\alpha + X_i\beta)$, where α represents the intercept, and $\beta = (\beta_1, \cdots, \beta_p)'$

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 $\alpha + X_i\beta$. According to Albert and Chib (1993), latent variables $Z = (Z_1, Z_2, \cdots, Z_n)'$ are introduced to simplify the structure. More specifically, we define

$$Z_i = \alpha + X_i\beta + \varepsilon_i, \tag{2.1}$$

where the random errors ε_i are independently and identically distributed as $N(0,1)$. The relationship between Y_i and Z_i is

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

Motivated by Lee et al. (2003) in setting a modified model for performing gene selection, we define an indicator vector

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

Given γ, let p_γ be the number of 1 in γ, 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 covariates consisting of all the columns of $\mathbf{X}$ corresponding to those elements of γ that are equal to 1. Hence, for a given γ, the probit regression model (2.1) is reduced to

$$Z_i = \alpha + X_{i,\gamma}\beta_\gamma + \varepsilon_i, \tag{2.2}$$

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

By introducing the latent vector Z and the indicator vector γ, we connect the probit binary regression model for Y_i to a normal linear regression model for Z_i. In the regression model (2.2), the unknowns are $(\alpha, \beta_\gamma, \gamma, Z)$. When $n < p_\gamma$, $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ is not full rank and the conventional approaches encounter serious difficulties. Thus, methods of gene selection for reducing

the dimension of the variable space are needed. As discussed, our gene selection based on (2.2) includes assigning a generalized singular g-prior (gsg-prior) for β_γ to avoid the problem due to a singular or nearly singular $\mathbf{X}_\gamma'\mathbf{X}_\gamma$; integrating α and β_γ out, and drawing γ from the marginal distribution to avoid possible computational difficulties; and estimating the posterior gene inclusion probability, $p(\gamma_i = 1|Y, \mathbf{X})$, by a sufficiently large number of MCMC samples. Genes with high posterior inclusion probabilities are selected for the classification. Therefore, our method updates Z and γ by an efficient MCMC algorithm, and avoids the computation relating to the regression parameters α and β_γ.

2.2.2 Prior Distribution

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

$$\alpha \sim N(0, h), \tag{2.3}$$

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.

Given γ, the prior distribution of the crucial regression coefficient parameters is taken as

$$\beta_\gamma|\gamma \sim N(0, \mathbf{H}_\gamma), \tag{2.4}$$

where $N(0, \mathbf{H}_\gamma)$ is a p_γ-dimensional multivariate normal distribution with mean 0 and covariance matrix $\mathbf{H}_\gamma$. The g-prior (see Zellner, 1986) for β_γ is $N(0, c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1})$, where c is a specified

value. If $n < p_\gamma$, then $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ is not a full rank matrix and $(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}$ does not exist. Moreover, as pointed out by Gupta and Ibrahim (2007), $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ is nearly singular for situations with high-dimensional covariates or highly collinear covariates. However, occurrence of such covariates is common in gene selection problems with large numbers of correlated genes. Taking g-prior for β_γ with such a covariance matrix may lead to the collapse of the MCMC algorithm and other convergence problems, or incorrect simulation of γ or β_γ in the MCMC sampler that may give misleading gene selection results. Here we consider a modified form of the g-prior, namely the generalized singular g-prior (gsg-prior), as follows

$$\beta_\gamma|\gamma \sim N(0, c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+), \tag{2.5}$$

where $(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+$ denotes the Moore-Penrose generalized inverse of $\mathbf{X}_\gamma'\mathbf{X}_\gamma$. This generalized inverse always exists even under situations with high-dimensional covariates, discrete covariates or highly collinear covariates. Moreover, if $\mathbf{X}_\gamma$ is a full column rank matrix, then $(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+ = (\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}$. Hence, the gsg-prior is appropriate for solving the singularity problem.

For $i = 1, \cdots, p$, the prior distributions of γ_i are assumed to be independent, and

$$\gamma_i \sim Bernoulli(\pi_i), \quad 0 \leq \pi_i \leq 1, \tag{2.6}$$

that is $p(\gamma_i = 1) = \pi_i$. We choose small values for π_i, hence restricting the number of genes in the model.

2.2.3 Computation

Let $Y = (Y_1, \cdots, Y_n)$. Under the model and prior specifications in the above sections, the joint posterior distribution is given by

$$p(Z, \alpha, \beta_\gamma, \gamma | Y, \mathbf{X}) \propto \left[\exp\{ -\frac{\sum_{i=1}^{n} (Z_i - \alpha - X_{i,\gamma}\beta_\gamma)^2}{2} \} \prod_{i=1}^{n} \mathrm{I}(A_i) \right] \times \exp(-\frac{\alpha^2}{2h}) \times \left[\exp(-\frac{\beta_\gamma' \mathbf{X}_\gamma' \mathbf{X}_\gamma \beta_\gamma}{2c}) \prod_{i=1}^{m_\gamma} \lambda_i^{-\frac{1}{2}} \right] \times \prod_{i=1}^{p} \pi_i^{\gamma_i} (1 - \pi_i)^{1-\gamma_i}, \tag{2.7}$$

where A_i is either equal to $\{Z_i : Z_i > 0\}$ or $\{Z_i : Z_i \leq 0\}$ corresponding to $Y_i = 1$ or $Y_i = 0$, respectively; $\lambda_1, \cdots, \lambda_{m_\gamma}$ ($m_\gamma \leq p_\gamma$) are the nonzero eigenvalues of $(\mathbf{X}_\gamma' \mathbf{X}_\gamma)^+$, and $\mathrm{I}(\cdot)$ is an indicator function. The MCMC methods can be applied to simulate observations from this intractable joint posterior distribution through the full conditional distributions. It can be shown that the conditional distribution of β_γ given (Z, α, γ) is multivariate normal with a covariance matrix $c(X_\gamma' X_\gamma)^+/(c+1)$. If X_γ is not of full column rank, this covariance matrix is not positive definite and the multivariate normal distribution is degenerated. This may induce convergence problems in the MCMC algorithm. To avoid this problem, we integrate α and β_γ out from the joint posterior distribution. This step can also reduce the strong posterior correlations between Z and β_γ, and β_γ and γ, and thus speeds up the computations. It can be shown that (see Appendix A), the joint posterior distribution of (Z, γ) is

given as follows:

$$p(Z,\gamma|Y,\mathbf{X}) \propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \prod_{i=1}^{n} \mathrm{I}(A_i) \times \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \tag{2.8}$$

where $\mathbf{\Sigma}_\gamma = \mathbf{I_n} + h11' + c\mathbf{X}_\gamma(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+\mathbf{X}_\gamma'$. As $\mathbf{\Sigma}_\gamma$ is positive definite, its inverse exists and $p(Z,\gamma|Y,\mathbf{X})$ is well defined.

The posterior distribution in (2.8) 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. The conditional distributions for implementing the Gibbs sampler are given below:
(*i*) $p(Z|Y,\mathbf{X},\gamma)$: It can be shown that $p(Z|Y,\mathbf{X},\gamma)$ is proportional to $N(0,\mathbf{\Sigma}_\gamma)\prod_{i=1}^{n}\mathrm{I}(A_i)$, which is a multivariate truncated normal distribution. Direct sampling from this distribution is known to be difficult. We follow the method given in Devroye (1986) to simulate samples from the univariate truncated normal distribution $p(Z_i|Z_{(-i)},Y,\mathbf{X},\gamma)$, where $Z_{(-i)}$ is the vector of Z without the i-th element.
(*ii*) $p(\gamma|Y,\mathbf{X},Z)$: This conditional distribution is proportional to $|\mathbf{\Sigma}_\gamma|^{-\frac{1}{2}}\exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$. Inspired by Panagiotelisa and Kohn (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

$$p(\gamma_i|\gamma_{(-i)},Y,\mathbf{X},Z) \propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \tag{2.9}$$

Because γ_i is binary, we can get the conditional probabilities $p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},Z)$ and $p(\gamma_i=0|\gamma_{(-i)},Y,\mathbf{X},Z)$. Denote $\gamma^1 = (\gamma_1,\cdots,\gamma_{i-1},\gamma_i=1,\gamma_{i+1},\cdots,\gamma_p)$ and $\gamma^0 = (\gamma_1,\cdots,\gamma_{i-1},\gamma_i =$

$0, \gamma_{i+1}, \cdots, \gamma_p)$, and similarly define $\mathbf{\Sigma}_{\gamma^1}$ and $\mathbf{\Sigma}_{\gamma^0}$ as the $\mathbf{\Sigma}_\gamma$ in (2.8). It can be shown that (see Appendix A):

$$\begin{aligned} p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z) &= \frac{p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z)}{p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z) + p(\gamma_i = 0|\gamma_{(-i)}, \mathbf{X}, Z)} \\ &= (1 + \frac{1 - \pi_i}{\pi_i}\rho)^{-1}, \end{aligned} \tag{2.10}$$

where

$$\rho = |\mathbf{\Sigma}_{\gamma^1}\mathbf{\Sigma}_{\gamma^0}^{-1}|^{\frac{1}{2}} \exp\left\{\frac{Z'(\mathbf{\Sigma}_{\gamma^1}^{-1} - \mathbf{\Sigma}_{\gamma^0}^{-1})Z}{2}\right\}. \tag{2.11}$$

As a result, an explicit form of the conditional distribution can be derived. In our method, although the dimension of β_γ in equation (2.2) changes in the MCMC iterations, it is not a problem because we integrate α and β_γ out before the Gibbs scheme so that only Z and γ (with a fixed dimension p) are updated. Moreover, by using equation (2.10) our method implements an efficient sampling scheme to do a search over the entire model space during each of iterations, which leads to a more effective algorithm in identifying the significant genes.

To implement the Gibbs sampler, we start with an initial value $(Z^{(0)}, \gamma^{(0)})$, and continue as follows: at the $(k+1)$-th iteration with the k-th value $(Z^{(k)}, \gamma^{(k)})$,
step (a): For $i = 1, 2, \cdots, n$, draw $Z_i^{(k+1)}$ from $p(Z_i^{(k)}|Z_{(-i)}^{(k)}, Y, \gamma^{(k)})$.
step (b): For $i = 1, 2, \cdots, p$, generate a random number u_i from a uniform distribution $U[0, 1]$, calculate the probability $p_i^{(k+1)} = p(\gamma_i^{(k+1)} = 1|\gamma_{(-i)}^{(k)}, Y, \mathbf{X}, Z^{(k+1)})$ via (2.10) and (2.11), and update γ_i as follows:

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

Under mild regularity conditions and for sufficiently large T, $(Z^{(T)}, \gamma^{(T)})$ simulated from the above Gibbs sampler can be regarded as an observation from the joint posterior distribution $p(Z, \gamma|Y, \mathbf{X})$, see Geman and Geman (1984). We collect MCMC samplers $\{(Z^{(k)}, \gamma^{(k)}), k = 1, 2, \cdots, M\}$ 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 $\gamma^{(0)}$ and 0 otherwise. In contrast, Lee et al. (2003) and 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 $(Z^{(k)}, \gamma^{(k)})$, which is important and sufficient for gene selection and classification, while the less important α and β (or β_γ) are not simulated. The relative frequency of each gene can be calculated as

$$\hat{p}(\gamma_i = 1|Y, \mathbf{X}) = \frac{1}{M} \sum_{k=1}^{M} 1[\gamma_i^{(k)} = 1]. \tag{2.12}$$

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 for classification.

2.2.4 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 (Lachenbruch and Mickey,

1968; McLachlan, 1992). Let $Y_{(-i)}$ be the vector of Y without the i-th element. A LOOCV predictive probability for Y_i can be calculated as

$$p(Y_i|Y_{(-i)}, \mathbf{X}) = \Big(\iint p(Y_i|Y_{(-i)}, \mathbf{X}, Z, \gamma)^{-1} p(Z, \gamma|Y, \mathbf{X}) dZ d\gamma \Big)^{-1}. \tag{2.13}$$

Equation (2.13) enables us to use the distribution $p(Z, \gamma|Y, X)$ that was computed with all the data in place of the distribution $p(Z, \gamma|Y_{(i)}, X)$ that is used in the LOOCV context. This replacement is useful to simplify the simulation of Z and γ in the required MCMC iterations and thus significantly reduces the computational and programming efforts in the gene selection problem with a fairly large sample size. An immediate Monte Carlo integration of (2.13) using the generated samples $\{(Z^{(k)}, \gamma^{(k)}), k = 1, 2, \cdots, M\}$ yields:

$$\hat{p}(Y_i|Y_{(-i)}, \mathbf{X}) = \frac{M}{\sum_{k=1}^{M} p(Y_i|Y_{(-i)}, \mathbf{X}, Z^{(k)}, \gamma^{(k)})^{-1}}. \tag{2.14}$$

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

$$p(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}) = \iint p(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}, Z, \gamma) p(Z, \gamma|Y, \mathbf{X}) dZ d\gamma.$$

Similarly, this probability can be approximated by Monte Carlo integration as follows:

$$\hat{p}(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}) = \frac{1}{M} \sum_{k=1}^{M} p(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}, Z^{(k)}, \gamma^{(k)}).$$

2.3 Results

We illustrate the usefulness of the proposed gsg-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 gsg-SSVS approach will be compared with the existing gene selection and classification methods.

2.3.1 Colon Cancer Data

Alon et al. (1999) used Affymetrix Oligonucleotide Array to measure expression levels of 40 tumor and 22 normal colon tissues for 6500 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 2000 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 tumor and 19 normal tissues, while the second contains 35 tumor 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 gsg-SSVS approach, we set $c = 10$, $\pi_i = 0.005, i = 1, \cdots, p$, 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 25 genes for chains 1 and 2, and 30 genes for chain 3 (see Appendix A) from a total of 2000 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 Brown et al. (1998) were used to check convergence. Fig. 2.1(a) shows that the most significant genes, which are determined by the posterior gene inclusion probabilities, are almost the same for three

chains. Fig.2.1(b) plots the number of selected genes versus the iteration number, and Fig.2.1(c) plots the log relative posterior probabilities of selected genes, $\log(p(\gamma|Y, \mathbf{X}, Z))$, versus the iteration number. Fig.2.1(b) and Fig.2.1(c) indicate that the three chains mixed well enough 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 (2.12)).

The 18 most significant genes ranked by the posterior gene inclusion probabilities (see Fig. 2.1(a)) for chain 1 are presented in Table 2.1. Seven of them were also selected by Ben-Dor et al. (2000). On the top of the genes listed in Table 2.1 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 pre-cancerous colon ploys. The second selected gene in Table 2.1 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. Some existing methods in the literature calculated the LOOCV error within the gene selection process. However, as pointed out by Ambroise and McLachlan (2002), 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 Chabrier, 2008), we perform the external LOOCV procedure as follows: 1) omit one observation of the training set, 2) based on the remaining observations, reduce the set of available genes to the top 50 genes as ranked in terms of the t statistic, 3) the p^* most significant genes were re-chosen from the 50 genes by our gsg-SSVS approach, and 4) these p^* genes were used to classify the left out sample. This process was repeated for all observations in the training set until each observation had been held out and predicted exactly once. The performance of our method with $p^* = 6$ and 10 are summarized in Table 2.2. Our method with 6 genes misclassified 5 tumor tissues (T1, T2, T30, T33, T36) and 3 normal tissues (N8, N34, 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 1000 or 2000 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 our proposed gsg-SSVS approaches. Our LOOCV results have been compared with the following classification methods: support vector machine (SVM; Furey et al., 2000); LogitBoost optimal, LogitBoost estimated, LogitBoost 100 iterations, AdaBoost 100 iterations, 1-nearest-neighbor, and Classification tree (Dettling and Bühlmann, 2003); MAVE-LD (Antoniadis et al., 2003) and Supervised group Lasso (SGLasso; Ma et al., 2007). The summary is presented in Table 2.2. 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 c, h, and π. For instance, using $c = 5$ as suggested by Lamnisos et al. (2009) and others, $h = 200$, and $\pi = 0.007$, the identification of the relevant genes and the performance of classification are essentially the same as before. The data set has also been analyzed by using three different chains with different random choices of $\gamma^{(0)}$. We observe that the three sets of the 18 most significant genes associated with different $\gamma^{(0)}$ are almost the same except a minor difference in the rank of gene indices and few non-overlapping genes (see Table A in Appendix A). Moreover, the LOOCV error rates produced by these three chains are the same. Therefore, it seems that the Bayesian results are robust to the choice of $\gamma^{(0)}$.

2.3.2 Leukemia Dataset

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 $p = 6817$ 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 27 are ALL and 11 are AML; and a test set of 34 samples of which 20 are ALL and 14 are AML.

Based on the protocol given in Dudoit et al. (2002), the following preprocessing steps were taken for the data: (i) thresh-

olding: floor of 100 and ceiling of 16000; (ii) filtering: exclusion of genes with max/min ≤ 5 and (max$-$min) ≤ 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 3571 genes. We further transformed the gene expression data to have mean zero and standard deviation one across samples. We applied the Bayesian gsg-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 3571 genes.

The posterior gene inclusion probabilities are presented in Fig.2.2. The relevant genes selected on the basis of these probabilities are reported in Table 2.3, together with the relevant genes selected by Golub et al. (1999) and Ben-Dor et al. (2000). 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 enters 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. 1987), CD63 antigene known as a member of the tranmenbrane 4 superfamily (Smith et al., 1995), and Macmarks known to be involved in growth and metastasis of certain tumors (Spizz and Blackshear, 1997).

The top-ranked 6 genes out of the 18 selected genes were used to conduct the prediction on the test set. The external LOOCV procedure described in Colon Cancer Data section was applied

to get the classification error on the training set. There was 1 training error and 1 test error (the 67-th observation). This 67-th observation was also misclassified in Golub et al. (1999) and Lee et al. (2003). In Table 2.4, we compare our classification results with the following popular classification methods: SVM (Furey et al., 2000); weighted voting machine (WVM) (Golub et al., 1999); MAVE-LD and MAVE-NPLD (Antoniadis et al., 2003); and PLS-LD and PLS-QDA (Nguyen and Rocke, 2002). Our results, with fewer genes, are better than or comparable to those obtained by the above existing methods in the literature. Furthermore, the test set has also been analyzed by the nearest shrunken centroids method (NSCM, Tibshirani et al., 2002) using 21 relevant genes, an iterative BMA algorithm (Yeung et al., 2005) using 20 genes, and the g-prior SSVS method (Lee et al., 2003) using 5 genes. The misclassification error rates made by NSCM, iterative BMA, and g-prior SSVS are 0.0588, 0.0588 and 0.0294, respectively. As no LOOCV error results related to the training set were reported in these analysis, it may not be fair to compare our gsg-SSVS approach with these methods.

2.3.3 Computational Time

The computational times for performing gsg-SSVS in the analysis of Colon Cancer Data and the Leukemia Data are respectively 43 minutes and 47 minutes for 10,000 iterations in a PC with Intel Core2 1.86GHz CPU 1G ram.

2.4 Discussion

We propose a Bayesian probit regression model for gene selection with binary data and then use a small number of the most relevant genes to perform classification. Based on a gsg-prior, a Bayesian SSVS approach using simulation-based MCMC tech-

nique is introduced. In this gsg-SSVS approach, the joint posterior distribution of $(\alpha, \beta_\gamma, \gamma, Z)$ is simplified to a joint posterior distribution of γ and Z after α and β_γ are integrated out. As $(X_\gamma' X_\gamma)^+$ and $\mathbf{\Sigma}_\gamma$ always exist, this posterior distribution is well defined. Moreover, by applying the efficient sampling scheme suggested by Panagiotelisa and Smith (2008), simulating samples from this posterior distribution is simple. At each MCMC iteration, it only requires the generation of Z_i and γ_i from an univariate truncated normal distribution and a binary distribution, respectively. As a result, the proposed algorithm is simple and efficient. 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. Based on the colon cancer and leukemia data sets, we demonstrated that the proposed gsg-SSVS approach compared favorably with other popular methods in performing disease classification.

In this chapter, we considered c and π as known hyperparameters in their prior distributions. This restriction can be relaxed by treating them as unknown parameters and further assigning prior distributions to them. We have not considered the multiclass problem, because the binary case is one of the most common settings. However, the key ideas in this chapter can be applied to handle the multiclass problem. We assume that genes are independent. Extending the model to account for a correlation structure between genes may be helpful for achieving better results.

□ **End of chapter.**

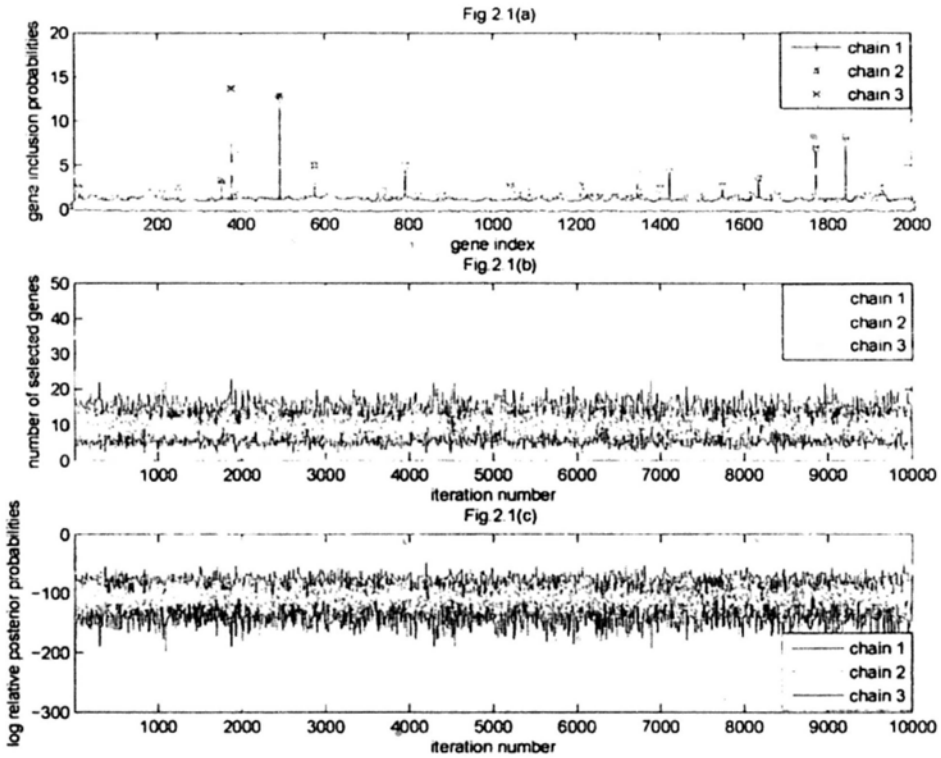

Figure 2.1: Fig.2.1(a) shows the gene inclusion probabilities (in percentages) versus the gene index, Fig.2.1(b) and Fig.2.1(c) show the number of selected genes and the log relative posterior probabilities of selected genes versus the first 10000 iteration number, respectively.

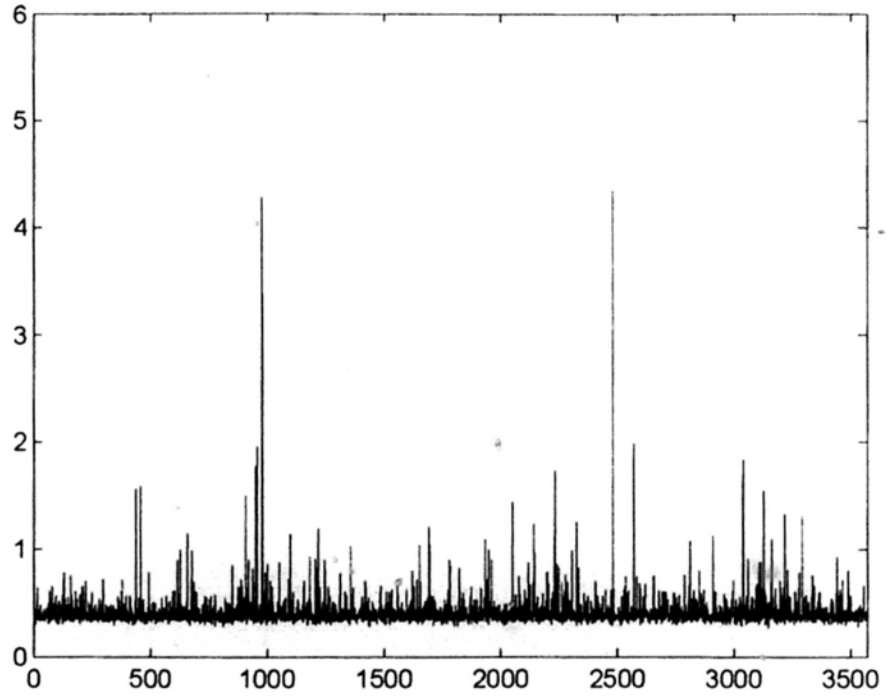

Figure 2.2: Fig.2.2 shows the gene inclusion probabilities (in percentages) versus the gene index for leukemia data.

Table 2.1: Colon cancer data: strongly significant genes for classifying normal and tumor tissues.

Rank	Clone ID	Gene annotation
1	Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor+
2	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE+
3	H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN)+
4	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)+
5	D14812	Human mRNA for ORF, complete cds
6	R88740	ATP SYNTHASE COUPLING FACTOR 6, MITOCHONDRIAL PRECURSOR (HUMAN);
7	J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM(HUMAN);+
8	T62947	60S RIBOSOMAL PROTEIN L24)+
9	M36634	Human vasoactive intestinal peptide (VIP) mRNA,+
10	T57882	MYOSIN HEAVY CHAIN, NONMUSCLE TYPE A
11	R36977	P03001 TRANSCRIPTION FACTOR IIIA;
12	T92451	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE(HUMAN);
13	M63391	Human desmin gene, complete cds.
14	H64807	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
15	R55310	S36390 MITOCHONDRIAL PROCESSING PEPTIDASE;
16	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM(HUMAN);
17	M59040	Human cell adhesion molecule (CD44) mRNA,
18	H11084	VASCULAR ENDOTHELIAL GROWTH FACTOR

+: Ben-Dor et al. (2000)

Table 2.2: 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	gsg-SSVS	10	0.1129
11	gsg-SSVS	6	0.1290

a: Furey et al. (2000);
b: Dettling and Bühlmann (2003);
c: Antoniadis et al. (2003)
d: Ma et al. (2007).

Table 2.3: Leukemia data: strongly significant genes for discriminating ALL and AML samples.

Rank	Gene ID	Gene descriptions
1	X95735	Zyxin^{+*}
2	M27891	CST3 Cystatin C^{+*}
3	Y12670	LEPR Leptin receptor^{+}
4	M23197	CD33 antigen (differentiation antigen)$^{+*}$
5	L09209	APLP2 Amyloid beta (A4) precursor-like protein 2*
6	M22960	PPGB Protective protein for beta-galactosidase*
7	X62654	CD63 antigen*
8	HG1612	Macmarcks*
9	D88422	CYSTATIN A*
10	M27783	ELA2 Elastatse 2, neutrophil
11	M16038	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog^{+*}
12	X04085	Catalase 5'flank and exon 1 mapping to chromosome 11^{+}
13	M83652	PFC Properdin P factor, complement^{+*}
14	X85116	Epb72 gene exon 1^{+*}
15	X74262	RETINOBLASTOMA BINDING PROTEIN P48^{+}
16	X51521	VIL2 Villin 2 (ezrin)*
17	U50136	Leukotriene C4 synthase (LTC4S) gene^{+*}
18	M92287	CCND3 Cyclin D3*

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

Table 2.4: The comparison of classification methods for the leukemia data.

	Method	No. of genes	LOOCV error rate	Test error rate
1	SVM[a]	25 to 1000	0.0526	0.0588 to 0.1176
2	WVM[b]	50	0.0526	0.1471
3	MAVE-LD[c]	50	0.0263	0.0294
4	MAVE-NPLD[c]	50	0.0263	0.0294
5	PLS-LD[d]	50	0.0000	0.0294
6	PLS-QDA[d]	50	0.0000	0.1765
7	gsg-SSVS	6	0.0263	0.0294

a: Furey et al. (2000);
b: Golub et al. (1999);
c: Antoniadis et al. (2003)
d: Nguyen and Rocke (2002).

Chapter 3

Multi-class Classification via Bayesian Variable Selection with Gene Expression Data

3.1 Introduction

In practice, DNA microarray gene expression data usually have the characteristics of fewer samples and larger number of genes. Multi-class classification, based on data with a relatively small number of samples (n) as compared to the number of variables (p) involved, is an important topic in bioinformatics. The problem of high-dimensional multi-class classification is challenging because many noise variables that may not be relevant to classification exist, and these variables can potentially degrade the prediction performance of classification. Moreover, identifying which variables contribute most to the multi-class classification is necessary.

Many variable selection methods related to multi-class classification have been described in the bioinformatics literature. These methods can be classified into univariate and multivariate approaches. Based on the marginal utility of each variable for the classification task, univariate methods consider each variable individually. These methods include parametric and

non-parametric methods. Examples include the weighted voting scheme (Golub et al., 1999), the threshold number of misclassification score (Ben-Dor et al., 2000), the significance analysis of microarray statistic (Tusher et al., 2001), the ratio of between-groups to within-groups sum of squares (Dudoit et al., 2002), the pairwise mean difference (Nguyen and Rocke, 2002), and the Wilcoxon test statistic (Dettling, 2004). Due to their conceptually simple nature, univariate methods have attracted much attention. However, they do not consider the correlations between variables, resulting in a subset of variables that may not be optimal for the considered classification task.

To take into account the dependency between genes for achieving a reduced number of relevant genes, Yeung and Bumgarner (2003) and Jaeger et al. (2003) proposed multivariate gene selection procedures, which do not score each variable individually but determine the combinations of variables that yield high prediction accuracy. The multivariate Bayesian gene selection approach based on the stochastic search variable selection method (George and McCulloch, 1993) has been applied to the multiclass classification problem (see Sha et al., 2004, Zhou et al., 2006). Sha et al. (2004) proposed an algorithm that is based on a multinomial probit model by using adding/deleting and swapping algorithm. According to Lamnisos et al. (2009), this kind of algorithm that randomly chooses to either add or delete a single explanatory variable, or to swap two explanatory variables in the model often leads to high model acceptance rates when the number of variables is substantially larger than the sample size. Moreover, the Metropolis random walk suggested by Sha et al. (2004) with local proposals and high acceptance rate is often associated with the poor mixing of MCMC chains. Furthermore, as their approach did not capture a priori correlation in the parameters, eliciting a prior covariance matrix with $p > n$ is difficult (Gupta and Ibrahim, 2009). Zhou et al. (2006) pro-

posed a multivariate Bayesian model using the g-prior (Zellner, 1986) for the unknown regression coefficients related to relevant genes. For situations with high-dimensional covariates, or highly collinear covariates, the covariance matrix involved in the g-prior is nearly singular (see Gupta and Ibrahim, 2007), resulting in the unstable convergence of the algorithm. Moreover, their methods assumed the covariance matrix of random errors to be an identity matrix. This specification has several limitations. For instance, it entails some symmetry between different classes, and an independence from irrelevant alternatives assumption is not appropriate in some applications (Train, 2003) because this specification postulates independent latent variables. Finally, both Sha et al. (2004) and Zhou et al. (2006) calculated the leave one out cross validation (LOOCV) within the gene selection process. According to Ambroise and McLachlan (2002) and Rocke et al. (2009), a selection bias that optimizes the classification accuracy exists when this internal LOOCV procedure is applied to estimate the prediction error.

In this chapter, we consider a multivariate Bayesian probit model together with a stochastic search variable selection (SSVS) method for the gene selection and the classification of diagnostic category for a multi-class problem. We propose a generalized g-prior (gg-prior) to overcome the problem induced by the possible singularity of the covariance matrix involved in the g-prior distribution of the regression coefficients. We show that this kind of gg-prior is effective in coping situations with a large number of genes and a small number of samples. 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 gg-SSVS algorithm, the efficient sampling scheme suggested by Panagiotelisa and Smith (2008) is implemented. For the posterior analysis associated with this

sampling scheme, the unknown intercept and regression coefficients in the proposed model are integrated out from the joint posterior distribution. This gives a simple and well-defined posterior distribution to ensure stable convergence of the resulting MCMC methods. Hence, our algorithm is more stable and efficient as compared to the MCMC-based algorithm of Sha et al. (2004) and Zhou et al. (2006). In addition, the gg-SSVS approach produces the posterior probability for the selected genes, which is helpful in a diagnostic setting. We illustrate the advantage of our method on two well-known microarray data sets: acute leukemia data (Golub et al., 1999) and lymphoma data (Alizadeh et al., 2000). We compare the performance of the proposed gg-SSVS approach with some other classification procedures in the literature, such as those of Dettling and Büuhlmann (2003) and Yeung et al. (2005), among others. Our results show that the proposed gg-SSVS approach reduces the number of selected genes and produces a prediction accuracy comparable to that of existing methods for variable selection and classification.

The rest of this chapter is structured as follows. The next section provides a brief review of matrix variate distribution. In the Method section, we specify the model on the basis of the stochastic search variable selection procedure. Discussions on the related prior distributions, the implementation of the Bayesian method, and the associated classification are also presented. The results obtained from the analysis of the two published data sets are given in the Results section. Some concluding remarks are presented in the Discussion section. The technical details are provided in Appendix B.

3.2 Matrix Variate Distribution

We follow the notation introduced by Dawid (1981) for matrix variate distribution. $\mathbf{M} + \mathcal{N}(\mathbf{P}, \mathbf{\Sigma})$ will stand for a matrix nor-

mal distribution of $\mathbf{X}$, where $\mathbf{M}$ is the matrix mean of $\mathbf{X}$, and $P_{ii}\mathbf{\Sigma}$ and $\Sigma_{ii}\mathbf{P}$ are the covariance matrices of the i-th row and j-th column of $\mathbf{X}$, respectively. Let $\mathbf{\Sigma} \sim \mathcal{IW}(\delta, \mathbf{Q})$, then the induced marginal distribution for $\mathbf{X}$ is a matrix $\mathcal{T}$ distribution denoted as $\mathcal{T}(\delta; \mathbf{P}, \mathbf{Q})$. The probability density functions of matrix normal distribution and matrix $\mathcal{T}$ distribution are given by Brown (1993) (see Appendix B).

3.3 Method

3.3.1 Model

Suppose we are given a training data set that consists of n samples $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $X_i = (X_{i1}, X_{i1}, \cdots, X_{ip}) \in R^p$ represents covariates or input vectors, and Y_i is a categorical response variable from sample i and takes on values, $0, 1, \cdots, K-1$. Based on the training data, we aim to predict the target values of previously unseen points given a set of new covariates.

Following the standard approach for the multinomial probit model (see Albert and Chib, 1993), we introduce n auxiliary variables $Z_i = (Z_{i1}, \cdots, Z_{iK-1}), i = 1, 2, \cdots, n$ to connect the multinomial probit model to the following multivariate normal linear regression model:

$$Z_i = \alpha + X_i\mathbf{B} + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.1}$$

where α' is a K-1 dimensional vector of intercept, $\mathbf{B}$ is a $p \times (K-1)$ matrix of regression coefficients, and $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{iK-1})$ $i.i.d.$ $\sim N(0, \mathbf{\Sigma})$. The relationship between the auxiliary variables Z_i and the discrete observations Y_i is defined as follows:

$$Y_i = \begin{cases} j & \text{if } \max_{1<k\leq K-1} Z_{ik} > 0, \text{ and } Z_{ij} = \max_{1<k\leq K-1} Z_{ik}, \\ 0 & \text{if } \max_{1<k\leq K-1} Z_{ik} \leq 0. \end{cases} \tag{3.2}$$

Let $\mathbf{Z} = (Z_1', \cdots, Z_n')'$, $\mathbf{X} = (X_1', \cdots, X_n')'$, and $\boldsymbol{\epsilon} = (\epsilon_1', \cdots, \epsilon_n')'$. The multivariate normal regression model (3.1) can be rewritten in matrix form as

$$\mathbf{Z} = 1_n\alpha + \mathbf{XB} + \boldsymbol{\epsilon}, \tag{3.3}$$

where 1_n is an n by 1 vector of ones, $\mathbf{X}$ is an $n \times p$ matrix of covariates, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{I}_n, \boldsymbol{\Sigma})$, in which $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\mathcal{N}(\cdot, \cdot)$ denotes the matrix normal distribution. As introduced in previous section, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{I}_n, \boldsymbol{\Sigma})$ indicates that the mean of $\boldsymbol{\epsilon}$ is an $n \times (K-1)$ matrix of zeros, the covariance matrices of the i-th row and the j-th column of $\boldsymbol{\epsilon}$ are $\boldsymbol{\Sigma}$ and $\sigma_{jj}\mathbf{I}_n$, respectively, and σ_{jj} is the j-th diagonal element of $\boldsymbol{\Sigma}$. This notation has the advantages of maintaining the matrices' structure, avoiding the need to string matrices by row or column as a vector, and using Kronecker product covariance to make the formal Bayesian manipulations much easier.

Let B_i denote the i-th row of $\mathbf{B}$. To model the relationship between the observation Y and a subset of the covariates in $\mathbf{X}$, we introduce an indicator vector $\gamma = (\gamma_1, \cdots, \gamma_p)$ such that

$$\gamma_i = \begin{cases} 1 & \text{if } B_i \neq 0, \\ 0 & \text{if } B_i = 0. \end{cases} \tag{3.4}$$

Here, $\gamma_i = 1$ indicates that the i-th covariate is included in the model, and $\gamma_i = 0$ otherwise. Incorporating γ into (3.3), a model indexed by γ is defined by

$$\mathbf{Z} = 1_n\alpha + \mathbf{X}_\gamma\mathbf{B}_\gamma + \boldsymbol{\epsilon}, \tag{3.5}$$

where $\mathbf{X}_\gamma$ denotes a submatrix of $\mathbf{X}$ with the columns corresponding to $\gamma_i = 0$ being deleted, and $\mathbf{B}_\gamma$ is a submatrix of $\mathbf{B}$ with the rows corresponding to $\gamma_i = 0$ being deleted. Let p_γ denote the number of ones in γ, and the dimension of $\mathbf{X}_\gamma$ and $\mathbf{B}_\gamma$ are $n \times p_\gamma$ and $p_\gamma \times (K-1)$, respectively.

3.3.2 Prior Specification

The unknowns in Model (3.5) are $(\alpha, \mathbf{B}_\gamma, \boldsymbol{\Sigma}, \gamma)$. The choice of prior distributions for these unknown parameters is very important in developing the Bayesian SSVS approach. We consider the following structure of prior distributions for α, $\mathbf{B}_\gamma$, γ, and $\boldsymbol{\Sigma}$:

$$p(\alpha, \mathbf{B}_\gamma, \boldsymbol{\Sigma}, \gamma) = p(\alpha|\boldsymbol{\Sigma})p(\mathbf{B}_\gamma|\boldsymbol{\Sigma}, \gamma)p(\gamma)p(\boldsymbol{\Sigma}). \tag{3.6}$$

Specifically, for α, γ, and $\boldsymbol{\Sigma}$, we propose the prior distributions as follows:

$$\alpha|\boldsymbol{\Sigma} \sim \mathcal{N}(h, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \sim \mathcal{IW}(\rho_0, \mathbf{R}_0), \quad \gamma_i \sim Bernoulli(\pi_i), \tag{3.7}$$

where h is taken to a large value, and $\mathcal{IW}(\cdot, \cdot)$ denotes the inverted Wishart distribution. The scale matrix hyperparameter $\mathbf{R}_0$ is usually taken in the form of $k\mathbf{I}_{K-1}$, in which k is a chosen constant, and $\mathbf{I}_{K-1}$ is a $(K-1) \times (K-1)$ identity matrix. As the expectation of $\boldsymbol{\Sigma}$ is $\mathbf{R}_0/(\rho_0 - 2)$, we generally take $\rho_0 = 3$, which is the smallest integer value such that the expectation of $\boldsymbol{\Sigma}$ exists. For γ, we propose the independent Bernoulli prior distribution with $\pi_i = p(\gamma_i = 1)$, which means that each covariate is selected independently with prior probability π_i, and the value of π_i is usually chosen to be small in order to restrict the number of covariates included in the model.

The prior distribution for the more crucial parameter $\mathbf{B}_\gamma$ is taken as:

$$\mathbf{B}_\gamma|\boldsymbol{\Sigma}, \gamma \sim \mathcal{N}(\mathbf{H}_\gamma, \boldsymbol{\Sigma}), \tag{3.8}$$

where $\mathbf{H}_\gamma$ is a $p_\gamma \times p_\gamma$ dimensional covariance matrix. According to Zellner (1986), the g-prior for $\mathbf{B}_\gamma$ is $\mathcal{N}(c(\mathbf{X}'\mathbf{X})^{-1}, \boldsymbol{\Sigma})$, where c is a specified value. If $n < p_\gamma$, then $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ is not a full rank matrix, and $(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}$ does not exist. Moreover, as pointed out by Gupta and Ibrahim (2007), $\mathbf{X}_\gamma'\mathbf{X}_\gamma$ is nearly singular for situations with high-dimensional covariates or highly collinear

covariates. However, the occurrence of such covariates is common in gene selection problems with large numbers of correlated genes. Taking g-prior for $\mathbf{B}_\gamma$ with such a covariance matrix may lead to the collapse of the MCMC algorithm and other convergence problems, or the incorrect simulation of γ or $\mathbf{B}_\gamma$ in the MCMC sampler which may give misleading gene selection results. Similar to Gupta and Ibrahim (2007), we consider a modified g-prior, the generalized g-prior (gg-prior), as follows:

$$\mathbf{B}_\gamma|\mathbf{\Sigma},\gamma \sim \mathcal{N}((\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma})^{-1}, \mathbf{\Sigma}), \tag{3.9}$$

where τ is a specified scalar similar to the ridge parameter in ridge regression. The advantage of the gg-prior in (3.9) is that it simultaneously stabilizes the prior and posterior simulation of the regression coefficients while possessing the operating characteristics and properties essentially identical to the usual g-prior when high dimensionality and collinearity issues are present. For example, when $p_\gamma > n$, the original matrix $\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c}$ is singular, but $\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{\mathbf{p}_\gamma}$ is not necessarily singular. The ridge parameter τ is generally chosen within a range of values between 0 and $1/c$, leading to maximum stability of the estimated coefficients. As suggested by Gupta and Ibrahim (2007), a fixed value of τ leads to more stable and less variable estimates, and the bias in estimates introduced by τ turns out to be negligible.

3.3.3 Computation

Based on the model and prior specifications, the joint posterior distribution $(\mathbf{Z}, \alpha, \mathbf{B}_\gamma, \mathbf{\Sigma}, \gamma | Y, \mathbf{X})$ is proportional to

$$\exp\Big\{-\frac{\mathrm{tr}[(\mathbf{Z}-1_n\alpha-\mathbf{X}_\gamma\mathbf{B}_\gamma)\mathbf{\Sigma}^{-1}(\mathbf{Z}-1_n\alpha-\mathbf{X}_\gamma\mathbf{B}_\gamma)']}{2}\Big\}\prod_{i=1}^{n}\mathrm{I}(A_i)$$
$$\times\Big|\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c}+\tau\mathbf{I}_{p_\gamma}\Big|^{\frac{K-1}{2}}\exp\Big\{-\frac{\mathrm{tr}[(\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c}+\tau\mathbf{I}_{p_\gamma})\mathbf{B}_\gamma\mathbf{\Sigma}^{-1}\mathbf{B}_\gamma']}{2}\Big\}$$
$$\times\exp\Big\{-\frac{\alpha\mathbf{\Sigma}^{-1}\alpha'}{2h}\Big\}|\mathbf{R}_0|^{\frac{\rho_0+K-2}{2}}\exp\Big\{-\frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}_0)}{2}\Big\}$$
$$\times\prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}|\Sigma|^{-\frac{n+p_\gamma+\rho_0+2K-1}{2}}, \tag{3.10}$$

where A_i is equal to either $\{Z_i : \max_{1<k\leq K-1} Z_{ik} > 0,\ Z_{ij} = \max_{1<k\leq K-1} Z_{ik}\}$ or $\{Z_i : \max_{1<k\leq K-1} Z_{ik} \leq 0\}$ corresponding to $Y_i = j$ or $Y_i = 0$, respectively, and $\mathrm{I}(\cdot)$ is an indicator function. As the joint posterior distribution in (3.10) is intractable, directly simulating observation from it is impossible. Hence, MCMC methods are used to iteratively simulate observations from the full conditional distribution of each component given the others. To reduce the strong posterior correlations among latent quantities $\mathbf{Z}$, $\alpha, \mathbf{B}_\gamma$, and $\mathbf{\Sigma}$ in the MCMC sampling, we integrate out the less important parameters α, $\mathbf{B}_\gamma$, and $\mathbf{\Sigma}$ from the joint posterior distribution, and focus on the most important parameter γ which determines the selected subset of variables in SSVS procedure. After integrating out $\alpha, \mathbf{B}_\gamma$, and $\mathbf{\Sigma}$, the marginal joint posterior distribution of $\mathbf{Z}$ and γ is proportional to (see Appendix B):

$$|\mathbf{P}_\gamma|^{\frac{\rho_0+n-1}{2}}|\mathbf{P}_\gamma+\mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}\prod_{i=1}^{n}\mathrm{I}(A_i)\times\prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \tag{3.11}$$

where $\mathbf{P}_\gamma = \mathbf{I}_n + h1_n1_n' + \mathbf{X}_\gamma(\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma})^{-1}\mathbf{X}_\gamma'$.

The posterior distribution in (3.11) is also intractable; hence the Gibbs sampler (Geman and Geman, 1984) is employed to generate observations from this posterior distribution. The full conditional distributions in implementing the Gibbs sampler are given below (see Appendix B):

$$(i) \qquad \mathbf{Z}|Y, \mathbf{X}, \gamma \sim \mathcal{T}(\rho_0, \mathbf{P}_\gamma, \mathbf{R}_0) \prod_{i=1}^{n} \mathrm{I}(A_i), \tag{3.12}$$

where $\mathcal{T}(\cdot,\cdot,\cdot)$ indicates the truncated matrix student t distribution. As direct sampling from this distribution is difficult, we iteratively simulate each row of $\mathbf{Z}$, Z_i, given the others from the corresponding conditional distributions. Let $\mathbf{Z}_{(-i)}$ be a submatrix of $\mathbf{Z}$ with the i-th row deleted. The conditional distribution of $(Z_i|\mathbf{Z}_{(-i)}, Y, \mathbf{X}, \gamma)$ is the following non-central multivariate truncated t distribution (using the notation of Brown, 1993):

$$\begin{aligned} & Z_i - P_{\gamma,i(-i)}\mathbf{P}^{-1}_{\gamma,(-i)(-i)}\mathbf{Z}_{(-i)} \qquad (3.13) \\ \sim & \mathcal{T}(\rho_0 + n - 1, P_{\gamma,ii.(-i)}, \mathbf{R}_0 + \mathbf{Z}'_{(-i)}P^{-1}_{\gamma,ii}\mathbf{Z}_{(-i)})\mathrm{I}(A_i), \end{aligned}$$

where $P_{\gamma,ii.(-i)} = P_{\gamma,ii} - P_{\gamma,i(-i)}\mathbf{P}^{-1}_{\gamma,(-i)(-i)}P'_{\gamma,i(-i)}$.

$$\begin{aligned} (ii) \qquad & p(\gamma|\mathbf{X}, \mathbf{Z}) \propto |\mathbf{P}_\gamma|^{\frac{\rho_0+n-1}{2}} |\mathbf{P}_\gamma + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}} \qquad (3.14) \\ & \times \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \end{aligned}$$

As γ has support on 2^p values, and p is large, obtaining its posterior by direct enumeration is impractical. Let $\gamma_{(-i)}$ denote the vector of γ without the i-th component. Following Panagiotelisa and Smith (2008), we in turn generate γ_i conditionally on the rest $\gamma_{(-i)}$. The conditional distribution of γ_i given $\gamma_{(-i)}, \mathbf{X}$, and $\mathbf{Z}$ is proportional to:

$$|\mathbf{P}_\gamma|^{\frac{\rho_0+n-1}{2}} |\mathbf{P}_\gamma + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}} \times \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \tag{3.15}$$

Since γ_i is binary, we can calculate the conditional probability of $p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z})$ and $p(\gamma_i = 0|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z})$ exactly. We denote $\gamma^1 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \cdots, \gamma_p)$ and $\gamma^0 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \cdots, \gamma_p)$, then

$$p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z}) = \tag{3.16}$$

$$\frac{p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z})}{p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z}) + p(\gamma_i = 0|\gamma_{(-i)}, \mathbf{X}, \mathbf{Z})} = (1 + \frac{1 - \pi_i}{\pi_i}\rho)^{-1},$$

where $\rho = \frac{|\mathbf{P}_{\gamma^0}|^{\frac{\rho_0+n-1}{2}} |\mathbf{P}_{\gamma^0} + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}}{|\mathbf{P}_{\gamma^1}|^{\frac{\rho_0+n-1}{2}} |\mathbf{P}_{\gamma^1} + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}}$.

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

Step (a): For $i = 1, \cdots, n$, generate $Z_i^{(t+1)}$ from its full conditional distribution (3.13).

Step (b): For $i = 1, \cdots, p$, generate a random number u_i from a uniform distribution $U[0, 1]$ and calculate the probability $p_i^{(t+1)} = p(\gamma_i^{(t+1)} = 1|\gamma_{(-i)}^{(t)}, Y, \mathbf{X}, \mathbf{Z}^{(t+1)})$. Then γ_i is updated as follows:

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

Let $\{(\mathbf{Z}^{(t)}, \gamma^{(t)}), t = 1, 2, \cdots, T\}$ denote the posterior simulation collected after convergence of the Gibbs sampler, where T is a sufficiently large number (Geman and Geman, 1984). The relative frequency of the i-th variable included in the model can be estimated below:

$$\hat{p}(\gamma_i = 1) = \frac{1}{T}\sum_{t=1}^{T} \gamma_i^{(t)}, \quad i = 1, 2, \cdots, p. \tag{3.17}$$

The value of $\hat{p}(\gamma_i = 1)$ provides us an estimate of the posterior variable inclusion probability as a measure of the relative importance of the i-th variable. Our gg-SSVS procedure searches

variables with high posterior inclusion probabilities for classification purpose.

3.3.4 Classification

We check the performance of a class prediction rule 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 (Lachenbruch and Mickey, 1968; McLachlan, 1992; 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

$$p(Y_i = j|Y_{(-i)}) = (\int\int p(Y_i = j|Y_{(-i)}, \mathbf{Z}, \gamma)^{-1} p(\mathbf{Z}, \gamma|Y) d\mathbf{Z} d\gamma)^{-1}. \tag{3.18}$$

An immediate Monte Carlo integration of (3.18) yields

$$\hat{p}(Y_i = j|Y_{(-i)}) = \frac{T}{\sum_{t=1}^{T} p(Y_i = j|Y_{(-i)}, \mathbf{Z}^{(t)}, \gamma^{(t)})^{-1}}, \tag{3.19}$$

where

$$\begin{aligned} & p(Y_i = j|Y_{(-i)}, \mathbf{Z}^{(t)}, \gamma^{(t)}) \\ =& \int p(Y_i = j|Z_i^{(t)}) p(Z_i^{(t)}|Y_{(-i)}, \mathbf{Z}^{(t)}, \gamma^{(t)}) dZ_i^{(t)} \\ =& \int I(Z_{ij}^{(t)} > Z_{ik}^{(t)}, \forall k \neq j) p(Z_i^{(t)}|Y_{(-i)}, \mathbf{Z}^{(t)}, \gamma^{(t)}) dZ_i^{(t)}. \end{aligned} \tag{3.20}$$

If a test set is available, the predictive posterior probability of $Y_{\text{new}} = j$ given the new covariate X_{new} is

$$p(Y_{\text{new}} = j|Y, X_{\text{new}}) = \int\int p(Y_{\text{new}} = j|Y, X_{\text{new}}, \mathbf{Z}, \gamma) p(\mathbf{Z}, \gamma|Y) d\mathbf{Z} d\gamma, \tag{3.21}$$

which can be approximated by the Monte Carlo estimation:

$$\hat{p}(Y_{\text{new}} = j|Y, X_{\text{new}}) = \frac{1}{T}\sum_{t=1}^{T} p(Y_{\text{new}}|Y, X_{\text{new}}, \mathbf{Z}^{(t)}, \gamma^{(t)}), \quad (3.22)$$

with

$$\begin{aligned} & p(Y_{\text{new}} = j|Y, X_{\text{new}}, \mathbf{Z}^{(t)}, \gamma^{(t)}) \\ =& \int p(Y_{\text{new}} = j|Z_{\text{new}}) p(Z_{\text{new}}|Y, X_{\text{new}}, \mathbf{Z}^{(t)}, \gamma^{(t)}) dZ_{\text{new}} \quad (3.23) \\ =& \int I(Z_{\text{newj}} > Z_{\text{newk}},\ \forall k \neq j) p(Z_{\text{new}}|Y, X_{\text{new}}, \mathbf{Z}^{(t)}, \gamma^{(t)}) dZ_{\text{new}}. \end{aligned}$$

Efficient methods for calculating the multivariate integration in Equations (3.21) and (3.23) are described by Genz and Bretz (2002).

3.3.5 Misclassification

When the classification rule determined in previous sections is applied to a multi-class dataset, there are many ways to measure and report the misclassification error rate. The class of each sample is predicted based on the selected variables, then compared against the given label. The overall misclassification error rate, which is the ratio of the total number of misclassification errors over the total sample size, is the simplest type of error rate. For multi-class problem, when the data set is characterized by unbalanced classes with a small number of cases in at least one of the classes, and this "rare" minority class is of particular interest to biologists for its value in diagnosing a disease, it is important and generally more informative to report the error rate for each class. Therefore, in our real applications, we compare our gg-SSVS procedure with other existing methods by reporting several classification error rates, including the overall

misclassification error rate, the average of class error rates, and error rate for each class (Wessels et al. 2005; Wood et al., 2007).

3.4 Real Data Analysis

3.4.1 Leukemia Data

We first applied our classification method to leukemia data, which were originally analyzed by Golub et al. (1999) and are 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 $p = 6817$ 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 were already divided into a training set consisting of 38 samples of which 19 are ALL-B, 8 are ALL-T, and 11 are AML; and a test set of 34 samples, of which 19 are ALL-B, 1 is ALL-T, and 14 are AML.

Following the protocol in Dudoit et al. (2002), preprocessing steps were taken for the data: (i) thresholding: floor of 100 and ceiling of 16000; (ii) filtering: exclusion of genes with max/min ≤ 5 and (max-min) ≤ 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 3571 genes. We further transformed the gene expression data to have mean zero and standard deviation one across samples.

To conduct the Bayesian gg-SSVS procedure, we set $c =$

$10, \pi_i = 0.005, i = 1, \cdots, p, h = 100, \mathbf{R}_0 = 2\mathbf{I}$, and $\rho_0 = 3, \tau = 0.01$. The initial value of $\gamma^{(0)}$ was taken with 25 randomly selected elements set to 1. Three diagnostic plots suggested by Smith and Kohn (1996) and Brown et al. (1998) were used to check convergence. Fig.3.1(a) shows the most significant genes, which are determined by the posterior gene inclusion probabilities. Fig.3.1(b) plots the number of selected genes versus the iteration number, and Fig.3.1(c) plots the log relative posterior probabilities of the selected genes, $\log(p(\gamma|Y, \mathbf{X}, \mathbf{Z}))$, versus the iteration number. Fig.3.1(b) and Fig.3.1(c) show that the three chains mixed well within 10,000 iterations. We collected 50,000 observations after 10,000 burn-in iterations to obtain the estimates of the posterior gene inclusion probabilities (see (3.17)).

Based on the entire training data, the 12 most significant genes, which were ranked by the posterior gene inclusion probabilities, are presented in Table 3.1. The leading gene in Table 3.1 is M27891, which also leads the list of strong genes in the works of Yeung et al. (2005) and Koo et al. (2006). Cystatins (CST3) are endogenous protein inhibitors of cathepsins, and these protease-inhibitor pairs, reported in myeloid cell lines with altered development, might be important in the etiology of AML. Golub et al. (1999) already showed that cystatin C gene is responsible for the subtype classification of leukemia as a two-class (ALL/AML) problem. The CST3 gene was also identified by Antonov et al. (2004) for AML/ALL classification. The relevance of gene X59871 to T-cell ALLs was reported in the biological literature. The gene TCF7 transcription factor 7 (T-cell specific) encodes a transcription factor that is a member of the high-mobility of group protein family. Expression of TCF7 is specific to T-cells, and the gene product was originally designated as TCF-1, a T-cell specific transcription factor. A closely related factor, LEF-1 (lymphocyte transcription factor), is expressed in both T- and B-cell lineages. Both TCF-1 and LEF-1

arise from the same gene, TCF7, by alternative splicing and the use of dual promoters (Kingsmore et al., 1995). We also identified some genes not identified by Yeung et al. (2005) and Koo et al. (2006), such as U05259 and M31523. The MB-1 gene encodes the Ig-alpha protein of the B-cell antigen component but may have other functions in addition to its role in signal transduction in B lineage cells. Ha et al. (1992) reported that MB-1 transcripts could be detected in pre-B cell lines and fetal bone marrow in normal, and mitogen activated- and transformed B cells but not in myeloma plasma cells. Furthermore, MB-1 is located in the 19q13 chromosomal region known to be a site of recurrent abnormalities in ALL. The MB-1 gene was also identified for AML/ALL classification (Gulob et al., 1999; Ben-Dor et al., 2000). Kamps et al. (1990) showed that the heterodimers between tissue-specific basic helix-loop-helix (bHLH) proteins and TCF3 play major roles in determining tissue-specific cell fate during embryogenesis, such as muscle or early B-cell differentiation. They are involved in a form of pre-B-cell acute lymphoblastic leukemia (B-ALL) through a chromosomal translocation which involves TCF3 and PBX1.

We first evaluate the performance of the classification methods for a selected subset of genes with the LOOCV procedure. An external LOOCV procedure proposed by Ambroise and McL achlan (2002) was used to perform the evaluation. Similar to many other multivariate methods, the external LOOCV procedure is challenged by server memory requirements and large computational time. According to the traditional attempts to overcome these problems (see Chu et al., 2005; Le Cao and Chabrier, 2008), we perform the external LOOCV procedure as follows: (1) omit one observation of the training set, (2) based on the remaining observations, reduce the set of available genes to the top 50 genes as ranked in terms of the ratio BSS/WSS (Dudoit et al., 2002), (3) the p^* most significant genes were

re-chosen from the 50 genes by our gg-SSVS approach, (4) the re-chosen p^* genes were used to classify the left out sample, and (5) 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. The misclassification errors of our method with $p^* = 8$, 10, and 12 are summarized in Table 3.2.

We further evaluate the performance of the classification methods for the test data. Our classification on the test data with $p^* = 8$, 10, and 12 genes reported one misclassification error with error rate 0.0294 (see Table 3.2). The test data have also been analyzed by some other multi-class classification methods. For instance, Lee and Lee (2003) reported one test error by multicategory support vector machine procedure using 40 selected genes. Yeung et al. (2005) applied the Bayesian model averaging (BMA) approach and reported one misclassified sample on the test set using 15 genes. This result is one of the most favorable results in the literature. Tan et al. (2005) applied the k-Top Scoring Pairs (k-TSP) to classify the test data. They reported one classification error with 36 genes. Koo et al. (2005) applied the structured polychotomous machine (SPM) to the test data and reported three classification errors using four genes. Our results on the test error rate, together with those given in previously published papers, are summarized in Table 3.3. Our method with fewer genes is shown to be comparable to other popular classification methods.

Whether or not the selected genes serve as legitimate markers for multi-class classification of the test data was further verified by the heat map of the selected genes. By visual inspection of the gene expression of the 12 selected genes, we detect some patterns for classifying ALL-T, ALL-B, and AML. Figure 3.2 illustrates three different patterns of the 12 selected genes in the same fashion as Figure 1 in Lee and Lee (2003) and Figure

5 in Koo et al. (2006).

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 c, π_i, h, $\mathbf{R}_0, \rho_0$, and τ. For instance, using $c = 5$ as suggested by Lamnisos et al. (2009), $\pi_i = 0.007$, $h = 200$, $\mathbf{R}_0 = 4\mathbf{I}, \rho_0 = 6$, and $\tau = 0.005$, the identification of the relevant genes and the performance of classification are essentially the same as before.

3.4.2 Lymphoma Data

The lymphoma data set was previously analyzed by Alizadeh et al. (2000) and are publicly available at `http://llmpp.nih.gov/lymphoma/data/figure1`. This data set contains gene expression levels of 4026 well-measured genes involving three most prevalent adult lymphoid malignancies: diffuse large B-cell lymphoma (DLBCL), chronic lymphocytic leukemia (CLL), and follicular lymphoma (FL). The total sample size is 62, of which 42 samples are DLBCL, 11 samples are CLL, and 9 samples are FL. Some samples contain a number of genes with unreliable or missing data. The following steps (Troyanskaya et al., 2001 and Dudoit et al., 2002) are used to impute the missing data for each gene with missing entries: (i) compute its correlation with all other p -1 genes, and (ii) for each missing entry, identify the five nearest genes having complete data for this entry and impute the missing entry by the average of the corresponding entries for the five neighbors. Each sample is further standardized to have mean zero and variance one across genes. We classify DLBCL, CLL, and FL using our method.

We applied the Bayesian gg-SSVS method with the same input of the hyperparameters as in the first example. The initial value of $\gamma^{(0)}$ is also taken with 25 randomly selected elements set to 1. The posterior gene inclusion probabilities estimated on the

entire training data are presented in Fig. 3.3. The relevant genes selected on the basis of these probabilities are reported in Table 3.4, together with the relevant genes selected by Tibshirani et al. (2003) and Draminski et al. (2008).

Since there is no test set available, the external LOOCV procedure described in Leukemia Data section was applied to obtain the classification error on the training set. In Table 3.5, we compare our classification results with the following popular classification methods: LogitBoost, estimated, AdaBoost, 100 iterations, Classification tree (Dettling and Büuhlmann, 2003), random forest var.sel., SC.s, and NN.vs (Díza-Uriarte and Andés, 2006). We observe from Table 3.5 that our results are comparable to those obtained by the existing methods.

3.4.3 Computational Time

The computational times to run 1 time of the gg-SSVS on the whole set of variables in the leukemia Data and lymphoma data are about 4.5 hours and 5 hours, respectively, for 60,000 iterations in a PC with an Intel Core2 1.86 GHz CPU and 1G ram.

3.5 Discussion

This chapter studies the problem of gene selection and multiclass classification when the sample size is small and the number of genes is large. The auxiliary variables are employed to relate the multinomial probit model to a multivariate regression model. We propose the Bayesian stochastic search variable selection method for gene selection on multi-class microarray data. The gg-prior is employed to solve the singular problem of the covariance matrix involved in the g-prior. We use the algorithm by integrating the regression coefficients out the joint posterior distribution to draw the indicator variable, so that the

MCMC chain will not be reducible. Our method also produces the posterior probabilities for selected genes, which is helpful in biological interpretation. As compared to other approaches on the same multi-class microarray data, our method uses fewer genes and produces comparable classification accuracy.

Chapter 2 (see also Yang and Song, 2010) proposed a hierarchical Bayesian model with a MCMC-based stochastic search algorithm to perform gene selection and classification for a two-class problem. They employed a generalized singular g-prior (gsg-prior) on the basis of the Moore-Penrose generalized inverse of the covariance matrix. We also use the gsg-prior for gene selection and multi-class classification. The gsg-SSVS with $p* = 8,\ 10,$ and 12 all reported a 0.0588 error rate for leukemia test data, which is slightly worse than the current results in Table 3.3, and 0.0323, 0.0323, and 0.0161 LOOCV error rates for lymphoma data, which are the same as the current results in Table 3.5. However, the gsg-SSVS approach is more computationally demanding due to the simulation of the Moore-Penrose generalized inverse of the covariance matrix in each MCMC iteration.

In this chapter, we consider c and π_i as known hyperparameters in their prior distributions. This restriction can be relaxed by treating them as unknown parameters and further assigning prior distributions to them. Extending our framework to account for an interaction structure between genes is also interesting.

□ **End of chapter.**

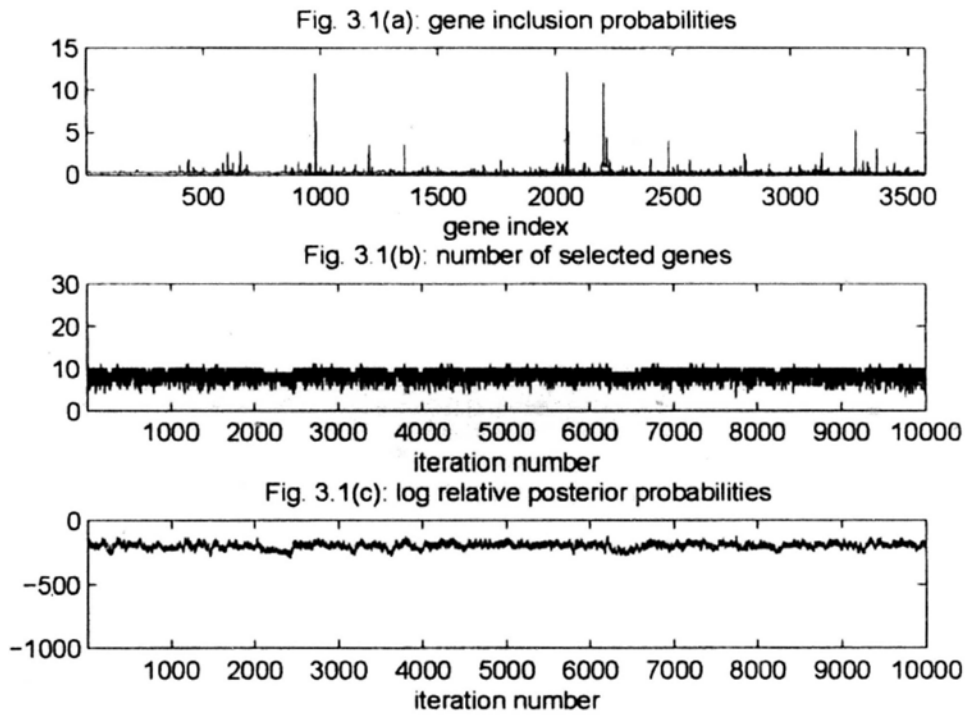

Figure 3.1: Fig.3.1(a) shows the gene inclusion probabilities (in percentages) versus the gene index, Fig.3.1(b) and Fig.3.1(c) show the number of selected genes and the log relative posterior probabilities of selected genes versus the first 10000 iteration number, respectively.

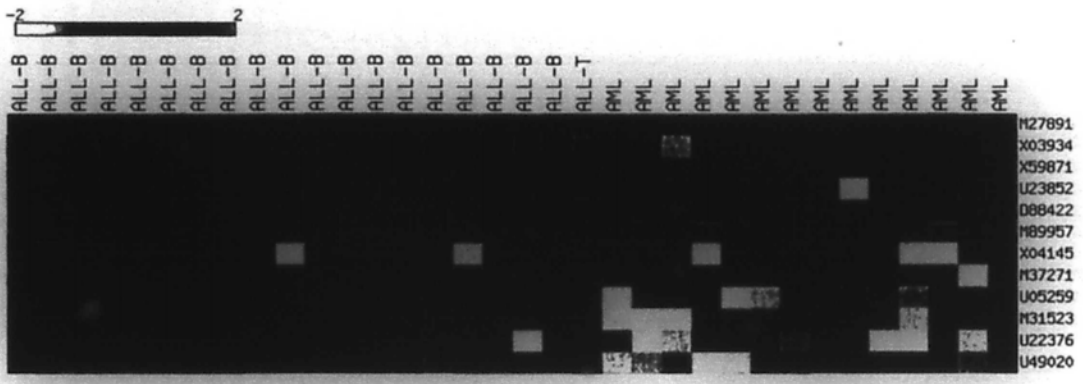

Figure 3.2: Genes that distinguish ALL-B, ALL-T and AML. Each column corresponds to a sample array and each row corresponds to a gene. The heat map is generated by using Matrix2png softerware (Pavlidis and Noble, 2003). Genes with expression levels greater than the mean are colored in red and those below the mean are colored in green.

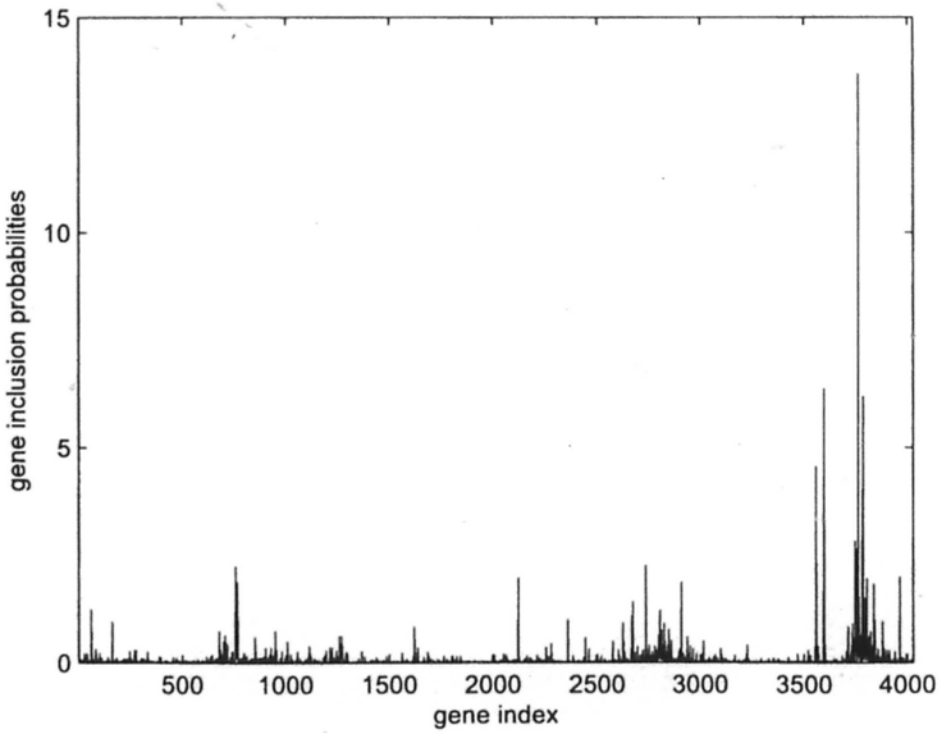

Figure 3.3: Fig.3.3 shows the gene inclusion probabilities (in percentages) versus the gene index.

Table 3.1: Significant genes found for discriminating ALL-T, ALL-B and AML.

Rank	Gene ID	Gene description
1	M27891	CST3 Cystatin C[a,b]
2	X03934	GB DEF = T-cell antigen receptor gene T3-delta[a]
3	X59871	TCF7 Transcription factor 7 (T-cell specific)[a]
4	U23852	GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA
5	D88422	CYSTATIN A
6	M89957	IGB Immunoglobulin-associated beta (B29)
7	X04145	CD3G CD3G antigen, gamma polypeptide
8	M37271	T-CELL ANTIGEN CD7 PRECURSOR
9	U05259	MB-1 gene
10	M31523	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
11	U22376	C-myb gene extracted from Human gene, complete primary cds, and five complete alternatively spliced cds
12	U49020	MEF2A gene (myocyte-specific enhancer factor 2A, C9 form) extracted from Human myocyte-specific enhancer factor 2A gene, first coding

a: Yeung et al. (2005).
b: Koo et al. (2006).

Table 3.2: Error rate results for the training data and test data of Leukemia data, respectively.

	No. of genes	Err^1	Err^2	Err_{ALL-B}	Err_{ALL-T}	Err_{AML}
Training data	8	0.0789	0.0768	0.1053	0.1250	0
	10	0.0526	0.0592	0.0526	0.1250	0
	12	0.0526	0.0592	0.0526	0.1250	0
Test data	8	0.0294	0.0238	0	0	0.0714
	10	0.0294	0.0238	0	0	0.0714
	12	0.0294	0.0238	0	0	0.0714

Err^1: overall error rate;
Err^2: average of the class error rates;
Err_{ALL-B}, Err_{ALL-T} and Err_{AML}: the class-specific error rates for ALL-B, ALL-T and AML, respectively.

Table 3.3: The comparison of classification results for Leukemia test data.

	Method	No. of genes	Overall test error rate
1	multicategory SVM[a]	40	0.0294
2	HC-k-TSP[c]	36	0.0294
3	BMA[b]	15	0.0294
4	PAM[d]	8	0.0588
5	SVM-RFE[d]	6	0.0882
6	SPM[d]	4	0.0882
7	gg-SSVS	8	0.0294
8	gg-SSVS	10	0.0294
9	gg-SSVS	12	0.0294

a: Lee and Lee (2003);
b: Yeung et al. (2005);
c: Tan et al. (2005);
d: Koo et al. (2006);

Table 3.4: Significant genes found for classifying DLBCL, CLL, and FL.

Rank	Gene ID	Gene description
1	GENE1622X	CD63 antigen (melanoma 1 antigen);[a,b] Clone=769861
2	GENE3805X	ISGF3 gamma=IFN alpha/beta-responsive transcription factor ISGF3 gamma[a] subunit (p48); Clone=1372520
3	GENE1644X	(cathepsin L; Clone=345538)[a,b]
4	GENE1775X	(Unknown UG Hs.140483 ESTs;[a] Clone=1319683)
5	GENE1648X	cathepsin B; Clone=297219[a]
6	GENE1647X	cathepsin B; Clone=261517[a,b]
7	GENE1673X	Glutathione peroxidase 1; Clone=712106[a]
8	GENE1610X	Mig=Humig=chemokine targeting T cells;[a,b] Clone=8
9	GENE1795X	CD31=PECAM-1; Clone=359925
10	GENE653X	(Lactate dehydrogenase A; Clone=686889)[a,b]
11	GENE2403X	(Unknown; Clone=1356913)[a,b]
12	GENE30X	(NC2 alpha subunit=repressor of class II gene transcription through specific binding to TBP-promoter complexes via heterodimeric histone fold domains; Clone=1340774)

a: Tibshirani et al. (2003);
b: Draminski et al. (2008).

Table 3.5: Comparison of LOOCV results of different methods for Lymphoma data.

	Method	No. of genes	LOOCV error rate
1	SC.s[b]	2796	0.0330
2	random forest var.sel. (s.e.=0)[b]	73	0.0470
3	random forest var.sel. (s.e.=1)[b]	58	0.0420
4	NN.vs[b]	15	0.0400
5	LogitBoost, estimated[a]	10	0.0323
6	AdaBoost, 100 iterations[a]	10	0.0484
7	Classification tree[a]	10	0.2258
8	gg-SSVS	8	0.0323
9	gg-SSVS	10	0.0323
10	gg-SSVS	12	0.0161

a: Dettling and Büuhlmann (2003);
b: Díza-Uriarte and Andés (2006).

Chapter 4

Sparse Bayesian Variable Selection for Classifying High-dimensional Microarray Data

4.1 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 linear regression models problem in statistics. Among many methods developed in the liter-

ature, several selection methods utilized correlations between genes and class labels. The correlation can be measured by signal-to-noise ratio (Gulob et al., 1999), the Pearson correlation (Hastie et al., 2001), t-statistic (Nguyen and Rocke, 2002), information-based criteria (Liu et al., 2005) and inter-class variations (Yang et al., 2006), or others. These procedures are univariate gene selection methods in the sense that the correlation between genes and disease is examined for each individual gene. Although being useful in practice and being easy to perform, all these methods select one important gene at a time and fail to take into account the correlation between genes. Alternative methods are multivariate approaches that consider multiple genes simultaneously and account for dependency between genes. Some of them are correlation-based approaches, for example, a fast correlation based filter solution (Yu and Liu, 2004) and the Markov blanket filter (Mamitsuka, 2006). Different from the correlation-based approaches, Lee et al. (2003) developed a multivariate Bayesian approach which used a Markov chain Monte Carlo (MCMC)-based stochastic search variable selection algorithm (George and McCulloch, 1993). They adopted the g-prior (Zellner, 1986) for unknown parameters of regression coefficients. However, for situations with high-dimensional covariates, or highly collinear covariates, the covariance matrix involved in the g-prior is nearly singular (Gupta and Ibrahim, 2007).

From a machine learning viewpoint, using support vector machines (SVMs) to deal with high-dimensional and small-sized data is attractive. SVMs have been demonstrated to achieve low test error in classification (Cristianini and Shawe-Taylor, 1999). However, as the standard SVMs utilize all the genes without discrimination, they can suffer from the presence of redundant genes (Hastie et al., 2001). Several methods have been proposed and have reported results on the application of SVMs for per-

forming gene selection. For example, by using generalization bounds from statistical learning theory, Weston et al. (2001) compared feature selection and Fisher score. But these methods need to estimate a trade-off parameter in order to utilize Mercer kernel functions and also lack of probabilistic output. Li et al. (2002) exploited an alternative approach, Bayesian technique of automatic relevance determination (ARD), to perform variable selection. Their approach adopted a zero-mean Gaussian prior with unknown variance for the unknown regression parameter. Compared with SVMs, variable sparsity is naturally incorporated into the algorithm and the optimal number of relevant variables is decided automatically, while SVMs need an additional variable selection procedure and a further criterion to indicate when the best variable set has been found. When applied to gene expression data sets, the ARD approach compared well with alternative kernel-based techniques. The main disadvantage is that the approach sets the value of the coefficient parameters corresponding to irrelevant variables to some small value but not to zero (shrinkage rather than selection). Bae and Mallick (2004) considered a multivariate Bayesian regression model. For the coefficient parameters, they assigned a zero-mean Gaussian prior with three different prior distributions for the unknown variance of the coefficient parameters. They selected the significant genes according to the posterior mean of the variance of the coefficient parameters.

In this chapter, 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)-based stochastic search algorithm. In developing our TH-SSVS algorithm, an efficient sampling scheme is im-

plemented. In addition, the TH-SSVS approach produces the posterior probabilities for the selected genes, which is helpful for achieving better biological interpretation.

4.2 Methods

4.2.1 Model

Suppose the data set has n observations with p predictors. Let $Y = (Y_1, \cdots, Y_n)$ denote the observed binary responses. For each sample i, let x_{ij} be the measurement of the expression level of the j-th gene for the i-th sample. Similar to Section 2.2.1, we define

$$Z_i = \alpha + X_i\beta + \varepsilon_i, \tag{4.1}$$

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

$$Y_i = \begin{cases} 1 & \text{if } Z_i > 0, \\ 0 & \text{if } Z_i \leq 0. \end{cases} \tag{4.2}$$

We introduce a latent binary vector $\gamma = (\gamma_1, \cdots, \gamma_p)$ to index the possible subsets of genes for performing gene selection. 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, given γ, model (4.1) can be rewritten as

$$Z_i = \alpha + X_{i,\gamma}\beta_\gamma + \varepsilon_i, \tag{4.3}$$

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

4.2.2 Prior Distribution

The choice of the prior distributions for the unknown parameters is very important in the Bayesian SSVS approach. Similar to Chapter 2, we consider prior distributions for α, β_γ, and γ with the structure $p(\alpha, \beta_\gamma, \gamma) = p(\alpha)p(\beta_\gamma|\gamma)p(\gamma)$ here.

The prior distribution of α is taken as

$$\alpha \sim N(0, h), \tag{4.4}$$

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 chapter. 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 coefficients 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 regression coefficient β_i given γ_i is assumed to be

$$p(\beta_i|\gamma_i) = (1-\gamma_i)\delta(0) + \gamma_i N(0, \lambda_i), \tag{4.5}$$

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

$$p(\gamma) = \prod_{i=1}^{p} \theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}, \quad 0 \le \theta_i \le 1, \tag{4.6}$$

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.

4.2.3 Computation

Denote $Z = (Z_1, \cdots, Z_n)', \mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_p)$. Under the model and prior specifications in the above sections, the joint

posterior distribution under Model I or Model II is given by

$$p(Z,\alpha,\beta_\gamma,\mathbf{\Lambda},\gamma|Y,\mathbf{X}) \propto \exp\{-\frac{\sum_{i=1}^{n}(Z_i-\alpha-X_{i,\gamma}\beta_\gamma)^2}{2}\}\prod_{i=1}^{n}\mathrm{I}(A_i)$$
$$\times \exp(-\frac{\alpha^2}{2h}) \times \prod_{i=1}^{p}\lambda_i^{-\frac{1}{2}}\exp(-\sum_{i=1}^{p}\frac{\beta_i^2}{2\lambda_i}) \times \prod_{i=1}^{p}\theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i},$$
$$\times \prod_{i=1}^{p}\lambda_i^{-(\frac{a}{2}+1)}\exp(-\sum_{i=1}^{p}\frac{b}{2\lambda_i})\Big\{\text{or}\quad \exp(-\sum_{i=1}^{p}\frac{\tau\lambda_i}{2})\Big\}, \qquad (4.7)$$

where A_i is equal to either $\{Z_i : Z_i > 0\}$ or $\{Z_i : Z_i \leq 0\}$ corresponding to $Y_i = 1$ or $Y_i = 0$, respectively; and $\mathrm{I}(\cdot)$ is an indicator function.

The posterior distribution in (4.7) 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 $Z, \beta_\gamma, \mathbf{\Lambda}$, and γ based on $p(Z,\mathbf{\Lambda}|\mathbf{X},Y,\beta,\gamma) \propto p(Z|\mathbf{X},Y,\mathbf{\Lambda},\gamma)p(\mathbf{\Lambda}|\beta,\gamma)$ and $p(\beta_\gamma,\gamma|\mathbf{X},Z,\mathbf{\Lambda}) \propto p(\beta_\gamma|\mathbf{X},Z,\mathbf{\Lambda},\gamma)$ $p(\gamma|\mathbf{X},Z,\mathbf{\Lambda})$. The conditional distributions for implementing our sampling scheme are given below:

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

$$p(Z|\mathbf{X},Y,\mathbf{\Lambda},\gamma) \propto N(0,\mathbf{\Sigma}_\gamma)\prod_{i=1}^{n}\mathrm{I}(A_i), \qquad (4.8)$$

with $\mathbf{\Sigma}_\gamma = h1_n1_n' + \mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma' + \mathbf{I}_n$, which is a multivariate truncated normal distribution (see Appendix C). In (4.8), β is margin alized out from the posterior distribution $p(Z|\mathbf{X},Y,\beta,\mathbf{\Lambda},\gamma)$ to

reduce autocorrelation between β and Z, thus to improve mixing in the Markov chain. Direct sampling from (4.8) is known to be difficult. We follow the method of Devroye (1986) to simulate samples from the univariate truncated normal distribution $p(Z_i|Z_{(-i)}, \mathbf{X}, Y, \mathbf{\Lambda}, \gamma)$, where $Z_{(-i)}$ is the vector of Z without the i-th element.

(ii) $p(\mathbf{\Lambda}|\beta, \gamma)$: The posterior distribution of the i-th diagonal element of $\mathbf{\Lambda}$, λ_i, under Model I is (see Appendix C)

$$\lambda_i|\beta_i, \gamma_i \sim \text{IG}(\frac{a+1}{2}, \frac{2}{b+\beta_i^2}). \tag{4.9}$$

The posterior distribution of λ_i under Model II is (see Appendix C)

$$\lambda_i^{-1}|\beta_i, \gamma_i \sim \text{InvGauss}(\frac{\sqrt{\tau}}{|\beta_i|}, \tau), \tag{4.10}$$

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

$$\text{InvGauss}(\iota, \kappa) = \sqrt{\frac{\kappa}{2\pi u^3}} \exp\{-\frac{\kappa(u-\iota)^2}{2\iota^2 u}\}, u > 0. \tag{4.11}$$

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

(iii) $p(\beta_\gamma|\mathbf{X}, Z, \mathbf{\Lambda}, \gamma)$: the full conditional distribution for β_γ is (see Appendix C)

$$\beta_\gamma|\mathbf{X}, Z, \mathbf{\Lambda}, \gamma \sim N(\mathbf{\Omega}_\gamma \mathbf{X}_\gamma' \mathbf{\Phi} Z, \mathbf{\Omega}_\gamma), \tag{4.12}$$

where $\mathbf{\Phi} = (h 1_n 1_n' + \mathbf{I}_n)^{-1}$, and $\mathbf{\Omega}_\gamma = (\mathbf{X}_\gamma' \mathbf{\Phi} \mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1})^{-1} = \mathbf{\Lambda}_\gamma - \mathbf{\Lambda}_\gamma \mathbf{X}_\gamma' \mathbf{\Phi} (\mathbf{\Phi} \mathbf{X}_\gamma \mathbf{\Lambda}_\gamma \mathbf{X}_\gamma' \mathbf{\Phi} + \mathbf{\Phi})^{-1} \mathbf{\Phi} \mathbf{X}_\gamma \mathbf{\Lambda}_\gamma$. The matrix inversion for calculating $\mathbf{\Omega}_\gamma$ 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*) $p(\gamma|\mathbf{X}, Z, \mathbf{\Lambda})$: This conditional distribution is proportional to $|\mathbf{\Sigma}_\gamma|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \prod_{i=1}^{p} \theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}$. We marginalize out β from the conditional distribution $p(\gamma|\mathbf{X}, Z, \beta, \mathbf{\Lambda})$ so that the Markov chain would be non-reducible (Panagiotelisa and Kohn, 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

$$p(\gamma_i|\gamma_{(-i)}, \mathbf{X}, Z, \mathbf{\Lambda}) \propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}. \tag{4.13}$$

Because γ_i is binary, we can get the conditional probabilities of $p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z, \mathbf{\Lambda})$ and $p(\gamma_i = 0|\gamma_{(-i)}, \mathbf{X}, Z, \mathbf{\Lambda})$. Denote $\gamma^1 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \cdots, \gamma_p)$ and $\gamma^0 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \cdots, \gamma_p)$, and similarly define $\mathbf{\Sigma}_{\gamma^1}$ and $\mathbf{\Sigma}_{\gamma^0}$ as $\mathbf{\Sigma}_\gamma$ in (4.8). It can be shown that (see Appendix C):

$$p(\gamma_i = 1|\gamma_{(-i)}, \mathbf{X}, Z, \mathbf{\Lambda}) = (1 + \frac{1-\theta_i}{\theta_i}\rho)^{-1}, \tag{4.14}$$

where

$$\rho = |\mathbf{\Sigma}_{\gamma^1}\mathbf{\Sigma}_{\gamma^0}^{-1}|^{\frac{1}{2}} \exp\left\{\frac{Z'(\mathbf{\Sigma}_{\gamma^1}^{-1} - \mathbf{\Sigma}_{\gamma^0}^{-1})Z}{2}\right\}. \tag{4.15}$$

As a result, an explicit form of the conditional distribution in (4.14) can be derived.

To implement the Gibbs sampler, we start with an initial value $(Z^{(0)}, \mathbf{\Lambda}^{(0)}, \beta_\gamma^{(0)}, \gamma^{(0)})$, and continue as follows: at the $(k+1)$-th iteration with the k-th value $(Z^{(k)}, \mathbf{\Lambda}^{(k)}, \beta_\gamma^{(k)}, \gamma^{(k)})$,
step (a): For $i = 1, \cdots, n$, draw $Z_i^{(k+1)}$ from the univariate truncated normal distribution $p(Z_i^{(k)}|Z_{(-i)}^{(k)}, \mathbf{X}, Y, \mathbf{\Lambda}^{(k)}, \gamma^{(k)})$.
step (b): For $i = 1, \cdots, p$, if $\gamma_i = 1$ draw $\lambda_i^{(k+1)}$ from the conditional distribution (4.9) and (4.10) 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 (4.12).
step (d): For $i = 1, \cdots, p$, generate a random number u_i from a uniform distribution $U[0,1]$, calculate the probability $p_i^{(k+1)} = p(\gamma_i^{(k+1)} = 1|\gamma_{(-i)}^{(k)}, \mathbf{X}, Z^{(k+1)}, \mathbf{\Lambda}^{(k+1)})$ via (4.14) and (4.15), and update γ_i as follows:

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

Under mild regularity conditions and for sufficiently large T, $(Z^{(T)}, \mathbf{\Lambda}^{(T)}, \beta_\gamma^{(T)}, \gamma^{(T)})$ simulated from the above Gibbs sampler can be regarded as an observation from the joint posterior distribution $p(Z, \beta_\gamma, \mathbf{\Lambda}, \gamma|Y, \mathbf{X})$, see Geman and Geman (1984). We collect MCMC samplers $\{(Z^{(k)}, \beta_\gamma^{(k)}, \mathbf{\Lambda}^{(k)}, \gamma^{(k)}), k = 1, 2, \cdots, M\}$ 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 $(Z^{(k)}, \beta_\gamma^{(k)}, \mathbf{\Lambda}^{(k)}, \gamma^{(k)})$, 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

$$\hat{p}(\gamma_i = 1|Y, \mathbf{X}) = \frac{1}{M} \sum_{k=1}^{M} \gamma_i^{(k)}. \tag{4.16}$$

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.

4.2.4 Classification

The performance of a classification rule is best assessed by applying the rule created on the training set to the test set. The predictive posterior probability of Y_{new} given the new covariate X_{new} is

$$\begin{aligned} & p(Y_{\text{new}}|Y, X_{\text{new}}) \\ &= \int p(Y_{\text{new}}|Y, X_{\text{new}}, Z, \beta, \mathbf{\Lambda}, \gamma) p(Z, \beta, \mathbf{\Lambda}, \gamma|Y) dZ d\beta d\mathbf{\Lambda} d\gamma. \end{aligned} \tag{4.17}$$

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

$$\hat{p}(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}) = \frac{1}{M} \sum_{k=1}^{M} p(Y_{\text{new}}|Y, \mathbf{X}, X_{\text{new}}, Z^{(k)}, \beta^k, \mathbf{\Lambda}^{(k)}, \gamma^{(k)}). \tag{4.18}$$

□ **End of chapter.**

Chapter 5

Summary and Discussion

The objective of this thsis is to propose new Bayesian approaches in variable selection and diseases classification for applications to high-dimensional data analysis. At first, we have introduced some background of Bayesian variable selection approach and reviewed some related literatures.

Chapter 2 proposes a Bayesian stochastic variable selection approach for gene selection based on a probit regression model with a generalized singular g-prior distribution for regression coefficients. Using simulation-based MCMC methods for simulating parameters from the posterior distribution, an efficient and dependable algorithm is implemented. It is shown that this algorithm is robust to the choices of initial values, and produces posterior probabilities of related genes for biological interpretation. The performance of the proposed approach is compared with those of other popular methods in gene selection and classification via the well known colon cancer and leukemia data sets in microarray literature.

Though we considered c and π as known hyperparameters in their prior distributions. This restriction can be relaxed by treating them as unknown parameters and further assigning prior distributions to them. Furthermore, we assume that genes are independent but in our framework the model can be easily extended to account for a correlation structure between

genes.

In Chapter 3, we propose a Bayesian stochastic search variable selection approach for multi-class classification, which can identify relevant genes by assessing sets of genes jointly. We consider a multinomial probit model with a generalized g-prior for the regression coefficients. An efficient algorithm using simulation-based MCMC methods are developed for simulating parameters from the posterior distribution. This algorithm is robust to the choice of initial value, and produces posterior probabilities of relevant genes for biological interpretation. We demonstrate the performance of the approach with two well-known gene expression profiling data: leukemia data and lymphoma data. Compared with other classification approaches, our approach selects smaller numbers of relevant genes and obtains competitive classification accuracy based on obtained results.

Chapter 4 is about the further research, which presents a stochastic variable selection approach with different two-level hierarchical prior distributions. 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.

Appendix A

A.1 Method

(i) Proof of equation (2.8).

Since the prior distributions for α, β_γ and γ are

$$\alpha \sim N(0,h), \beta_\gamma|\gamma \sim N(0, c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+), \gamma_i \sim \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \text{(A 1)}$$

and conditional on parameters α, β_γ, and γ,

$$Z_i = \alpha + X_{i,\gamma}\beta_\gamma + \varepsilon_i, \quad i = 1, \cdots, n, \qquad \text{(A 2)}$$

we have

$$Z_i|Y, \mathbf{X}, \alpha, \beta_\gamma, \gamma \sim N(\alpha + X_{i,\gamma}\beta_\gamma, 1)\mathrm{I}(A_i), \qquad \text{(A 3)}$$

where A_i is equal to either $\{Z_i : Z_i > 0\}$ or $\{Z_i : Z_i \leq 0\}$ corresponding to $Y_i = 1$ or $Y_i = 0$, respectively; and $\mathrm{I}(\cdot)$ is an indicator function which truncates the univariate normal distribution of Z_i to the appropriate region.

The joint posterior distribution of $(Z, \alpha, \beta_\gamma, \gamma)$ given $(Y, \mathbf{X})$

is

$$
\begin{aligned}
p(Z,\alpha,\beta_\gamma,\gamma|Y,\mathbf{X}) &\propto \prod_{i=1}^{n} p(Z_i|Y,\mathbf{X},\alpha,\beta_\gamma,\gamma)p(\alpha)p(\beta_\gamma|\mathbf{X},\gamma)\prod_{i=1}^{p} p(\gamma_i) \\
&\propto \left[\exp\{-\frac{\sum_{i=1}^{n}(Z_i-\alpha-X_{i,\gamma}\beta_\gamma)^2}{2}\}\prod_{i=1}^{n}\mathrm{I}(A_i)\right] \\
&\quad \times \exp(-\frac{\alpha^2}{2h}) \times \left[\exp(-\frac{\beta_\gamma'\mathbf{X}_\gamma'\mathbf{X}_\gamma\beta_\gamma}{2c})\prod_{i=1}^{m_\gamma}\lambda_i^{-\frac{1}{2}}\right] \\
&\quad \times \prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i},
\end{aligned}
\tag{A 4}
$$

where $\lambda_1,\cdots,\lambda_{m_\gamma}$ $(m_\gamma \leq p_\gamma)$ are the nonzero eigenvalues of $(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^+$. We first integrate out α given Z,β_γ,γ. The exponentiated terms that are associated with α in above equation can be rewritten as follows:

$$
\begin{aligned}
&-\frac{\sum_{i=1}^{n}(Z_i-\alpha-X_{i,\gamma}\beta_\gamma)^2}{2}-\frac{\alpha^2}{2h} \\
&=-\frac{(Z-1\alpha-\mathbf{X}_\gamma\beta_\gamma)'(Z-1\alpha-\mathbf{X}_\gamma\beta_\gamma)}{2}-\frac{\alpha^2}{2h} \\
&=-\frac{(h^{-1}+n)\{\alpha-(h^{-1}+n)^{-1}1'(Z-\mathbf{X}_\gamma\beta_\gamma)\}^2}{2} \\
&\quad -\frac{(1+nh)^{-1}(Z-\mathbf{X}_\gamma\beta_\gamma)'(Z-\mathbf{X}_\gamma\beta_\gamma)}{2}.
\end{aligned}
\tag{A 5}
$$

The exponential of the first term in expression (A 5) forms the kernel of a Gaussian probability density of α and can be integrated out. Thus, the integration of α is done.

Using a special case of binomial inverse theorem (see Woodbury 1950; Plackeett, 1950), the second term of expression (A 5) can be expressed as

$$
-\frac{(Z-\mathbf{X}_\gamma\beta_\gamma)'(\mathbf{I}_n+h11')^{-1}(Z-\mathbf{X}_\gamma\beta_\gamma)}{2}. \tag{A 6}
$$

Turning to the integration of β_γ, the expression (A 6) plus the third term of expression (A 4) can be rewritten as

$$
\begin{aligned}
&-\frac{\beta_\gamma' \mathbf{X}_\gamma' \{(\mathbf{I}_n + h11')^{-1} + c^{-1}\mathbf{I}_n\}\mathbf{X}_\gamma \beta_\gamma - 2\beta_\gamma' \mathbf{X}_\gamma (\mathbf{I}_n + h11')^{-1} Z}{2} \\
&\quad - \frac{Z'(\mathbf{I}_n + h11')^{-1} Z}{2} \\
&= -\frac{(\beta_\gamma - A^{-1}B)' A (\beta_\gamma - A^{-1}B)}{2} - \frac{Z'(\mathbf{I}_n + h11')^{-1} Z - B' A^{-1} B}{2},
\end{aligned}
\tag{A 7}
$$

where $A = \mathbf{X}_\gamma' \{(\mathbf{I}_n + h11')^{-1} + c^{-1}\mathbf{I}_n\}\mathbf{X}_\gamma$, $B = \mathbf{X}_\gamma'(\mathbf{I}_n + h11')^{-1} Z$.

The first term of expression (A 7) is a completed quadratic form in β_γ, which forms a Gaussian probability density and can be integrated out. The second term forms the kernel of a posterior probability density of $Z|\mathbf{X}, \gamma$ as

$$
\begin{aligned}
&-\frac{Z'(\mathbf{I}_n + h11')^{-1} Z}{2} \\
&-\frac{-Z'(\mathbf{I}_n + h11')^{-1}\mathbf{X}_\gamma [\mathbf{X}_\gamma' \{(\mathbf{I}_n + h11')^{-1} + c^{-1}\mathbf{I}_n\}\mathbf{X}_\gamma]^{-1} \mathbf{X}_\gamma' (\mathbf{I}_n + h11')^{-1} Z}{2}.
\end{aligned}
\tag{A 8}
$$

From expression (A 8), we obtain that $p(Z|\mathbf{X}, \gamma) \sim N(0, \mathbf{\Sigma}_\gamma)$, with $\mathbf{\Sigma}_\gamma^{-1} = (\mathbf{I}_n + h11')^{-1}$
$-(\mathbf{I}_n + h11')^{-1}\mathbf{X}_\gamma [\mathbf{X}_\gamma' \{(\mathbf{I}_n + h11')^{-1} + c^{-1}\mathbf{I}_n\}\mathbf{X}_\gamma]^{-1} \mathbf{X}_\gamma' (\mathbf{I}_n + h11')^{-1}$.
Denote $\mathbf{\Sigma}_\gamma^* = \mathbf{I_n} + h11' + c\mathbf{X}_\gamma (\mathbf{X}_\gamma' \mathbf{X}_\gamma)^+ \mathbf{X}_\gamma'$. Then

$$
\begin{aligned}
\mathbf{\Sigma}_\gamma^{-1} \mathbf{\Sigma}_\gamma^* = \{&(\mathbf{I}_n + h11')^{-1} \\
&-(\mathbf{I}_n + h11')^{-1}\mathbf{X}_\gamma [\mathbf{X}_\gamma' \{(\mathbf{I}_n + h11')^{-1} + c^{-1}\mathbf{I}_n\}\mathbf{X}_\gamma]^{-1} \mathbf{X}_\gamma' (\mathbf{I}_n + h11')^{-1}\} \\
&\times [(\mathbf{I_n} + h11') + c\mathbf{X}_\gamma (\mathbf{X}_\gamma' \mathbf{X}_\gamma)^+ \mathbf{X}_\gamma']
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{I}_n + c(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&\quad -(\frac{1}{1+nh} + \frac{1}{c})^{-1}(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&\quad -(\frac{1}{1+nh} + \frac{1}{c})^{-1}(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'}(\mathbf{I_n} + h11^{'})^{-1}c\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + c(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&\quad -(\frac{1}{1+nh} + \frac{1}{c})^{-1}(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&\quad -(\frac{1}{1+nh} + \frac{1}{c})^{-1}\frac{c}{1+nh}(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + c(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} \\
&\quad -c(\mathbf{I_n} + h11^{'})^{-1}\mathbf{X}_\gamma(\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma)^{+}\mathbf{X}_\gamma^{'} = \mathbf{I}_n.
\end{aligned}
$$

Therefore, $\mathbf{\Sigma}_\gamma = \mathbf{\Sigma}_\gamma^{*}$ and

$$
p(Z|\mathbf{X}, \gamma) \sim N(0, \mathbf{\Sigma}_\gamma). \qquad \text{(A 9)}
$$

Hence, the joint posterior distribution of $(Z, \gamma|Y, \mathbf{X})$ is

$$
\begin{aligned}
p(Z, \gamma|Y, \mathbf{X}) &\propto p(Z|Y, \mathbf{X}, \gamma)p(\gamma) \\
&\propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}}\exp(-\frac{Z^{'}\mathbf{\Sigma}_\gamma^{-1}Z}{2})\prod_{i=1}^{n}\mathrm{I}(A_i) \times \prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}.
\end{aligned}
\qquad \text{(A 10)}
$$

(ii) Proof of equation (2.10).

From equations (A 1) and (A 10), we have

$$
\begin{aligned}
p(\gamma_i|\gamma_{(-i)}, Y, \mathbf{X}, Z) &\propto p(Z|\mathbf{X}, \gamma)p(\gamma_i) \\
&\propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}}\exp(-\frac{Z^{'}\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i},
\end{aligned}
\qquad \text{(A 11)}
$$

and

$$p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z) \propto \frac{1}{|\mathbf{\Sigma}_{\gamma^1}|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \pi_i, \quad \text{(A 12)}$$

$$p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z) \propto \frac{1}{|\mathbf{\Sigma}_{\gamma^0}|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^0}^{-1}Z}{2}) \times (1 - \pi_i), \quad \text{(A 13)}$$

where $\gamma^1 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \cdots, \gamma_p)$ and $\gamma^0 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \cdots, \gamma_p)$. As γ_i is binary, we have

$$p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z) + p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z) = 1. \quad \text{(A 14)}$$

From equations (A 12)-(A 14), we get

$$\begin{aligned}
& p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z) \\
&= \frac{p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z)}{p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z) + p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z)} \\
&= \frac{|\mathbf{\Sigma}_{\gamma^1}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \pi_i}{|\mathbf{\Sigma}_{\gamma^1}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \pi_i + |\mathbf{\Sigma}_{\gamma^0}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^0}^{-1}Z}{2}) \times (1 - \pi_i)} \\
&= (1 + \frac{1 - \pi_i}{\pi_i}\rho)^{-1},
\end{aligned}$$

where

$$\rho = |\mathbf{\Sigma}_{\gamma^1}\mathbf{\Sigma}_{\gamma^0}^{-1}|^{\frac{1}{2}} \exp\{\frac{Z'(\mathbf{\Sigma}_{\gamma^1}^{-1} - \mathbf{\Sigma}_{\gamma^0}^{-1})Z}{2}\}. \quad \text{(A 15)}$$

A.2 Results

Table A: Initial (randomly selected) gene indices, selected gene indices by gsg-SSVS, and the LOOCV error rates for three different chains in the analysis of Colon Cancer Data.

Chain No.	Initial (randomly selected) gene index	selected gene index by gsg-SSVS
Chain 1	203 1879 385 1743 100 1621 513 1021 1642 531 213 909 1172 933 1254 346 1583 1207 949 1329 493 227 16 1306 846	377 493 1843 1772 576 792 1423 1346 1635 353 1042 **822** 249 1924 **1210** 14 1400 1549
Chain 2	1023 980 726 1946 657 1702 1224 1522 807 1773 1144 240 1573 825 1272 728 1962 1790 771 889 1219 1504 941 1724 1684	377 493 1843 1772 576 792 1423 1346 1635 353 1549 1042 **1210** 249 1924 1400 14 **625**
Chain 3	388 1459 683 1572 882 718 1723 1950 275 1024 1569 1755 342 29 778 633 1080 734 1365 1286 116 82 557 1372 1207 406 95 1769 985 1398	377 493 1843 1772 1423 792 576 1635 353 1346 249 1549 1924 1400 1042 **625** 14 **739**

*The gene indices in boldface indicate non-overlapping genes in the three sets of the 18 most significant genes. Note that the ten top-ranked selected genes are the same, only minor differences appeared in relation to genes with lower ranks.

Appendix B

B.1 Matrix Variate Distribution

(i) Matrix normal distribution

Let $\mathbf{Z}(p \times q)$ be a random matrix. $\mathbf{Z}$ is said to follow a matrix normal distribution, $\mathbf{M} + \mathcal{N}(\mathbf{P}, \mathbf{Q})$, if $\mathbf{M}$ is the mean of $\mathbf{Z}$, and $P_{ii}\mathbf{Q}$ and $Q_{ii}\mathbf{P}$ are the covariance matrices of the i-th row and the j-th column of $\mathbf{Z}$, respectively.

If $\mathbf{P}$ and $\mathbf{Q}$ are positive definite, the probability density function of the matrix normal distribution can be represented by

$$p(\mathbf{Z}) = (2\pi)^{-\frac{pq}{2}}|\mathbf{P}|^{-\frac{q}{2}}|\mathbf{Q}|^{-\frac{p}{2}} \exp\Big\{ - \frac{\mathrm{tr}[\mathbf{P}^{-1}(\mathbf{Z} - \mathbf{M})\mathbf{Q}^{-1}(\mathbf{Z} - \mathbf{M})']}{2}\Big\}.$$

(ii) Inverse Wishart distribution

Let $\mathbf{U}(q \times q)$ be a matrix. $\mathbf{U}$ is said to follow an inverse Wishart distribution, $\mathbf{U} \sim IW(\delta; \mathbf{Q})$, if for $\delta > 0$, the density function is defined as

$$p(\mathbf{U}) = c(q, \delta)|\mathbf{Q}|^{\frac{\delta+q-1}{2}}|\mathbf{U}|^{-\frac{\delta+2q}{2}} \exp\{-\frac{\mathrm{tr}(\mathbf{U}^{-1}\mathbf{Q})}{2}\}, \quad \mathbf{U} > 0,$$

with $c(q, \delta) = 2^{-\frac{q(\delta+q-1)}{2}}/\Gamma_q[\frac{\delta+q-1}{2}]$.

(iii) Matrix Student T-distribution

Let $\Phi \sim IW(\delta; \mathbf{Q})$ and given Φ, $\mathbf{T} \sim \mathcal{N}(\mathbf{P}, \mathbf{\Phi})$. The induced marginal distribution for $\mathbf{T}(p \times q)$ is a matrix T-distribution denoted as $\mathcal{T}(\delta; \mathbf{P}, \mathbf{Q})$.

The density function of the matrix T-distribution exists if $\delta > 0, \mathbf{P} > 0$, and $\mathbf{Q} > 0$. The density function of the matrix T-distribution is given by

$$p(\mathbf{T}) = c(p,q,\delta)|\mathbf{P}|^{\frac{\delta+p-1}{2}}|\mathbf{Q}|^{-\frac{p}{2}}|\mathbf{P}+\mathbf{T}\mathbf{Q}^{-1}\mathbf{T}'|^{-\frac{\delta+p+q-1}{2}},$$

where $c(p,q,\delta) = \pi^{-\frac{pq}{2}}\Gamma_q[\frac{\delta+p+q-1}{2}]/\Gamma_q[\frac{\delta+q-1}{2}]$.

Marginal and conditional distributions of the matrix T-distrib ution are also matrix-T. For example, if $\mathbf{T}$ is partitioned into $\mathbf{T}' = (\mathbf{T}_1', \mathbf{T}_2')$ with $\mathbf{T}_i(p_i \times q)$, $i = 1,2$, and $p_1 + p_2 = p$, then marginally $\mathbf{T}_2 \sim \mathcal{T}(\delta; \mathbf{P}_{22}, \mathbf{Q})$, and the conditional distribution of $\mathbf{T}_1$ given $\mathbf{T}_2$ is $\mathbf{T}_1 - \mathbf{P}_{12}\mathbf{P}_{22}^{-1}\mathbf{T}_2 \sim \mathcal{T}(\delta + p_2; \mathbf{P}_{11.2}, \mathbf{Q} + \mathbf{T}_2'\mathbf{P}_{11}^{-1}\mathbf{T}_2)$.

B.2 Method

(i) Proof of equation (3.13)

Since the prior distributions for α, $\mathbf{B}_\gamma$, $\mathbf{\Sigma}$ and γ are

$$\begin{aligned} &\alpha|\mathbf{\Sigma} \sim \mathcal{N}(h, \mathbf{\Sigma}), \quad \mathbf{B}_\gamma|\mathbf{\Sigma} \sim \mathcal{N}((\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma})^{-1}, \mathbf{\Sigma}), \\ &\mathbf{\Sigma} \sim \mathcal{IW}(\rho_0, \mathbf{R}_0), \quad p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \end{aligned} \tag{B 1}$$

and conditional on parameters α, $\mathbf{B}_\gamma$ $\mathbf{\Sigma}$ and γ,

$$\mathbf{Z} = 1_n\alpha + \mathbf{X}_\gamma\mathbf{B}_\gamma + \boldsymbol{\epsilon}, \tag{B 2}$$

Then,

$$(\mathbf{Z}|Y, \mathbf{X}, \alpha, \mathbf{B}_\gamma, \mathbf{\Sigma}, \gamma) - (1_n\alpha + \mathbf{X}_\gamma\mathbf{B}_\gamma) \sim \mathcal{N}(\mathbf{I}_n, \mathbf{\Sigma}) \prod_{i=1}^{n} \mathrm{I}(A_i), \tag{B 3}$$

where A_i is equal to either $\{Z_i : \max_{1<k\leq K-1} Z_{ik} > 0$, and $Z_{ij} = \max_{1<k\leq K-1} Z_{ik}\}$ or $\{Z_i : \max_{1<k\leq K-1} Z_{ik} \leq 0\}$ corresponding to $Y_i = j$ or $Y_i = 0$, respectively, and $\mathrm{I}(A_i)$ is an indicator function.

The joint posterior distribution of $(\mathbf{Z}, \alpha, \mathbf{B}_\gamma, \mathbf{\Sigma}, \gamma | Y, \mathbf{X})$ is proportional to

$$|\mathbf{\Sigma}|^{-\frac{n}{2}} \exp\{-\frac{\mathrm{tr}[(\mathbf{Z} - 1_n\alpha - \mathbf{X}_\gamma\mathbf{B}_\gamma)\mathbf{\Sigma}^{-1}(\mathbf{Z} - 1_n\alpha - \mathbf{X}_\gamma\mathbf{B}_\gamma)']}{2}\} \prod_{i=1}^{n} \mathrm{I}(A_i)$$

$$\times |\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma}|^{\frac{K-1}{2}} |\mathbf{\Sigma}|^{-\frac{p_\gamma}{2}} \exp\{-\frac{\mathrm{tr}[(\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma})\mathbf{B}_\gamma\mathbf{\Sigma}^{-1}\mathbf{B}_\gamma']}{2}\}$$

$$\times |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{\alpha\mathbf{\Sigma}^{-1}\alpha'}{2h}\} \times \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \times |\Sigma|^{-\frac{\rho_0+2K-2}{2}} |\mathbf{R_0}|^{\frac{\rho_0+K-2}{2}}$$

$$\times \exp\{-\frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}_0)}{2}\}. \qquad \text{(B 4)}$$

We first integrate out α given $\mathbf{Z}$, $\mathbf{B}_\gamma$, $\mathbf{\Sigma}$, and γ. Using $\mathrm{tr}(\mathbf{MN}) = \mathrm{tr}(\mathbf{NM})$, the exponentiated terms that are associated with α in expression (B 4) can be rewritten as:

$$-\frac{(\mathbf{Z} - 1_n\alpha - \mathbf{X}_\gamma\mathbf{B}_\gamma)'(\mathbf{Z} - 1_n\alpha - \mathbf{X}_\gamma\mathbf{B}_\gamma)}{2} - \frac{\alpha'\alpha}{2h}$$

$$= -\frac{(h^{-1}+n)\{\alpha - (h^{-1}+n)^{-1}1_n'(Z - \mathbf{X}_\gamma\mathbf{B}_\gamma)\}\{\alpha - (h^{-1}+n)^{-1}1_n'(Z - \mathbf{X}_\gamma\mathbf{B}_\gamma)\}'}{2}$$

$$-\frac{(1+nh)^{-1}(\mathbf{Z} - \mathbf{X}_\gamma\mathbf{B}_\gamma)'(\mathbf{Z} - \mathbf{X}_\gamma\mathbf{B}_\gamma)}{2}. \qquad \text{(B 5)}$$

The exponential of the first term in expression (B 5), together with the factors $\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1})$ and $|\mathbf{\Sigma}|^{-\frac{1}{2}}$, forms the kernel of a matrix variate normal probability density of α and can be integrated out. Thus, the integration of α is done.

Using a special case of binomial inverse theorem (see Woodbury, 1950 and Plackeett, 1950), the second term of expression (B 5) can be written as

$$-\frac{(\mathbf{Z} - \mathbf{X}_\gamma\mathbf{B}_\gamma)'(\mathbf{I}_n + h1_n1_n')^{-1}(\mathbf{Z} - \mathbf{X}_\gamma\mathbf{B}_\gamma)}{2}. \qquad \text{(B 6)}$$

Denote $\mathbf{H}_\gamma = (\frac{\mathbf{X}_\gamma'\mathbf{X}_\gamma}{c} + \tau\mathbf{I}_{p_\gamma})^{-1}$. Turning to the integration of $\mathbf{B}_\gamma$, the expression (B 6) plus the second term of expression (B 4) can be rewritten as

$$\begin{aligned}
&-\frac{\mathbf{B}_\gamma'\{\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}\}\mathbf{B}_\gamma - \mathbf{B}_\gamma'\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z}}{2} \\
&-\frac{-\mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma\mathbf{B}_\gamma + \mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z}}{2} \qquad \text{(B 7)} \\
&= -\frac{(\mathbf{B}_\gamma - \mathbf{M}^{-1}\mathbf{N})'\mathbf{M}(\mathbf{B}_\gamma - \mathbf{M}^{-1}\mathbf{N})}{2} - \frac{\mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z} - \mathbf{M}'\mathbf{N}^{-1}\mathbf{M}}{2},
\end{aligned}$$

where $\mathbf{M} = \{\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}\}$, and $\mathbf{N} = \mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z}$.

The first term of expression (B 7) is the completed quadratic form of $\mathbf{B}_\gamma$. Together with the factors $\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1})$ and $|\mathbf{\Sigma}|^{-\frac{1}{2}}$, this term forms a matrix normal probability density and can be integrated out. The second term of (B 7) forms the kernel of the posterior probability density of $\mathbf{Z}|\mathbf{X}, \gamma$ as follows:

$$\begin{aligned}
&-\frac{\mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z}}{2} \qquad \text{(B 8)} \\
&-\frac{\mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z}}{2}.
\end{aligned}$$

From expression (B 8), $p(\mathbf{Z}|\mathbf{X}, \mathbf{\Sigma}, \gamma) \sim \mathcal{N}(\mathbf{P}_\gamma^*, \mathbf{\Sigma})$, with

$$\begin{aligned}
&\mathbf{P}_\gamma^{*-1} = (\mathbf{I}_n + h1_n1_n')^{-1} \\
&-(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}.
\end{aligned}$$

Denote $\mathbf{P}_\gamma = \mathbf{I_n} + h1_n1_n^{'} + \mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'}$. Then,

$$
\begin{aligned}
&\mathbf{P}_\gamma^{*-1}\mathbf{P}_\gamma \\
&= \{(\mathbf{I}_n + h1_n1_n^{'})^{-1} - (\mathbf{I}_n + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}]^{-1} \\
&\quad \mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n^{'})^{-1}\}(\mathbf{I_n} + h1_n1_n^{'} + \mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'}) \\
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}]^{-1} \\
&\quad \mathbf{X}_\gamma^{'}(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma}{nh+1} + \mathbf{H}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma}{nh+1} + \mathbf{H}_\gamma^{-1}]^{-1}\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma\mathbf{H}_\gamma}{nh+1}\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'} - (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{H}_\gamma\mathbf{X}_\gamma^{'} = \mathbf{I}_n.
\end{aligned}
$$

Therefore, $\mathbf{P}_\gamma = \mathbf{P}_\gamma^{*}$ and

$$
\begin{aligned}
p(\mathbf{Z}|Y, \mathbf{X}, \mathbf{\Sigma}, \gamma) &\sim \mathcal{N}(\mathbf{P}_\gamma, \mathbf{\Sigma}) \\
&\propto |\mathbf{P}_\gamma|^{-\frac{K-1}{2}}|\mathbf{\Sigma}|^{-\frac{n}{2}}\exp\{-\frac{\mathrm{tr}[\mathbf{\Sigma}^{-1}\mathbf{Z}^{'}\mathbf{P}_\gamma^{-1}\mathbf{Z}]}{2}\}\prod_{i=1}^{n}\mathrm{I}(A_i).
\end{aligned}
\tag{B 9}
$$

Since the form of the last term of expression (B 4) is the same as expression (B 9), their product is proportional to

$$
|\mathbf{P}_\gamma|^{-\frac{K-1}{2}}|\mathbf{R_0}|^{\frac{\rho_0+K-2}{2}}|\mathbf{\Sigma}|^{-\frac{n+\rho_0+2(K-1)}{2}}\exp\{-\frac{\mathrm{tr}[\mathbf{\Sigma}^{-1}(\mathbf{Z}^{'}\mathbf{P}_\gamma^{-1}\mathbf{Z} + \mathbf{R}_0)]}{2}\}\prod_{i=1}^{n}\mathrm{I}(A_i).
\tag{B 10}
$$

Given γ, $\mathbf{\Sigma}$ can be integrated out of expression (B 10). Thus, we have

$$p(\mathbf{Z}|Y,\mathbf{X},\gamma) \propto |\mathbf{P}_\gamma|^{\frac{n+\rho_0-1}{2}}|\mathbf{R_0}|^{-\frac{n}{2}}|\mathbf{P}_\gamma+\mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}\prod_{i=1}^{n}\mathrm{I}(A_i), \tag{B 11}$$

which is the probability density function of the truncated matrix student T-distribution $\mathcal{T}(\rho_0;\mathbf{P}_\gamma,\mathbf{R}_0)$.

(ii) Proof of equation (3.16)

From equations (B 4) and (B 11),

$$\begin{aligned} p(\gamma_i|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z}) &\propto p(\mathbf{Z}|\mathbf{X},\gamma)p(\gamma_i) \\ &\propto |\mathbf{P}_\gamma|^{\frac{n+\rho_0-1}{2}}|\mathbf{R_0}|^{-\frac{n}{2}}|\mathbf{P}_\gamma+\mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}} \\ &\quad\times \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \end{aligned} \tag{B 12}$$

and

$$\begin{aligned} &p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z}) \\ &\propto |\mathbf{P}_{\gamma^1}|^{\frac{n+\rho_0-1}{2}}|\mathbf{R_0}|^{-\frac{n}{2}}|\mathbf{P}_{\gamma^1}+\mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}\times\pi_i, \qquad \text{(B 13)} \\ &p(\gamma_i=0|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z}) \\ &\propto |\mathbf{P}_{\gamma^0}|^{\frac{n+\rho_0-1}{2}}|\mathbf{R_0}|^{-\frac{n}{2}}|\mathbf{P}_{\gamma^0}+\mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}\times(1-\pi_i), \qquad \text{(B 14)} \end{aligned}$$

where $\gamma^1=(\gamma_1,\cdots,\gamma_{i-1},\gamma_i=1,\gamma_{i+1},\cdots,\gamma_p)$ and $\gamma^0=(\gamma_1,\cdots,\gamma_{i-1},\gamma_i=0,\gamma_{i+1},\cdots,\gamma_p)$. As γ_i is binary, we have

$$p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z})+p(\gamma_i=0|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z})=1. \tag{B 15}$$

From equations (B 13)-(B 15), we get

$$\begin{aligned} &p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z}) \qquad \text{(B 16)} \\ &=\frac{p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z})}{p(\gamma_i=1|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z})+p(\gamma_i=0|\gamma_{(-i)},Y,\mathbf{X},\mathbf{Z})} \\ &=(1+\frac{1-\pi_i}{\pi_i}\rho)^{-1}, \end{aligned}$$

where

$$\rho = \frac{|\mathbf{P}_{\gamma^0}|^{\frac{\rho_0+n-1}{2}}|\mathbf{P}_{\gamma^0} + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}}{|\mathbf{P}_{\gamma^1}|^{\frac{\rho_0+n-1}{2}}|\mathbf{P}_{\gamma^1} + \mathbf{Z}\mathbf{R}_0^{-1}\mathbf{Z}'|^{-\frac{\rho_0+n+K-2}{2}}}. \qquad \text{(B 17)}$$

Appendix C

C.1 Method

(i) Proof of equation (4.8).

Since the prior distributions for α, β_γ, λ and γ are

$$\begin{aligned} &\alpha \sim N(0,h), \quad \beta_i|\gamma_i \sim N(0,\lambda_i), \quad \gamma_i \sim \theta_i{}^{\gamma_i}(1-\theta_i)^{1-\gamma_i}, \\ &\lambda_i \sim \mathrm{IG}(\frac{a}{2},\frac{2}{b}) \quad (\text{or} \quad \lambda_i \sim Ga(1,\frac{\tau}{2})) \end{aligned} \tag{C 1}$$

and conditional on parameters α, β_γ, and γ,

$$Z_i = \alpha + X_{i,\gamma}\beta_\gamma + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{C 2}$$

we have

$$Z_i|Y, \mathbf{X}, \alpha, \beta_\gamma, \gamma \sim N(\alpha + X_{i,\gamma}\beta_\gamma, 1)\mathrm{I}(A_i), \tag{C 3}$$

where A_i is equal to either $\{Z_i : Z_i > 0\}$ or $\{Z_i : Z_i \leq 0\}$ corresponding to $Y_i = 1$ or $Y_i = 0$, respectively; and $\mathrm{I}(\cdot)$ is an indicator function which truncates the univariate normal distribution of Z_i to the appropriate region.

The joint posterior distribution of $(Z, \alpha, \beta_\gamma, \mathbf{\Lambda}, \gamma)$ given $(Y, \mathbf{X})$

is

$$
\begin{aligned}
& p(Z, \alpha, \beta_\gamma, \Lambda, \gamma | Y, \mathbf{X}) \\
& \propto \prod_{i=1}^{n} p(Z_i | Y, \mathbf{X}, \alpha, \beta_\gamma, \gamma) p(\alpha) p(\beta_\gamma | \mathbf{X}, \mathbf{\Lambda}, \gamma) \prod_{i=1}^{p} p(\lambda_i) p(\gamma_i) \\
& \propto \left[\exp\{ - \frac{\sum_{i=1}^{n} (Z_i - \alpha - X_{i,\gamma} \beta_\gamma)^2}{2} \} \prod_{i=1}^{n} \mathrm{I}(A_i) \right] \times \exp(-\frac{\alpha^2}{2h}) \\
& \times |\mathbf{\Lambda}_\gamma|^{-\frac{1}{2}} \exp(-\frac{\beta_\gamma' \mathbf{\Lambda}_\gamma^{-1} \beta_\gamma}{2c}) \times \prod_{i=1}^{p} \theta_i^{\gamma_i} (1 - \theta_i)^{1 - \gamma_i}, \\
& \times \prod_{i=1}^{p} \lambda_i^{-(\frac{a}{2}+1)} \exp(-\sum_{i=1}^{p} \frac{b}{2\lambda_i}) \{ \text{or} \quad \exp(-\sum_{i=1}^{p} \frac{\tau \lambda_i}{2}) \}
\end{aligned}
\tag{C 4}
$$

We first integrate out α given $Z, \beta_\gamma, \mathbf{\Lambda}, \gamma$. The exponentiated terms that are associated with α in above equation can be rewritten as follows:

$$
\begin{aligned}
& - \frac{\sum_{i=1}^{n} (Z_i - \alpha - X_{i,\gamma} \beta_\gamma)^2}{2} - \frac{\alpha^2}{2h} \\
& = - \frac{(Z - 1\alpha - \mathbf{X}_\gamma \beta_\gamma)' (Z - 1\alpha - \mathbf{X}_\gamma \beta_\gamma)}{2} - \frac{\alpha^2}{2h} \\
& = - \frac{(h^{-1} + n) \{ \alpha - (h^{-1} + n)^{-1} 1' (Z - \mathbf{X}_\gamma \beta_\gamma) \}^2}{2} \\
& \quad - \frac{(1 + nh)^{-1} (Z - \mathbf{X}_\gamma \beta_\gamma)' (Z - \mathbf{X}_\gamma \beta_\gamma)}{2}.
\end{aligned}
\tag{C 5}
$$

The exponential of the first term in expression (C 5) forms the kernel of a Gaussian probability density of α and can be integrated out. Thus, the integration of α is done.

Using a special case of binomial inverse theorem (see Woodbury 1950; Plackeett, 1950), the second term of expression (C 5) can be expressed as

$$
- \frac{(Z - \mathbf{X}_\gamma \beta_\gamma)' (\mathbf{I}_n + h 11')^{-1} (Z - \mathbf{X}_\gamma \beta_\gamma)}{2}. \tag{C 6}
$$

Turning to the integration of β_γ, the expression (C 6) plus the third term of expression (C 4) can be rewritten as

$$
\begin{aligned}
&-\frac{\beta_\gamma'\{\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_{n'})^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}\}\beta_\gamma - \beta_\gamma'\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}Z}{2} \\
&-\frac{-Z'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma\beta_\gamma + Z'(\mathbf{I}_n + h1_n1_n')^{-1}Z}{2} \\
&= -\frac{(\beta_\gamma - \mathbf{\Omega\Phi})'\mathbf{\Omega}^{-1}(\beta_\gamma - \mathbf{\Omega\Phi})}{2} - \frac{\mathbf{Z}'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{Z} - \mathbf{\Phi}'\mathbf{\Omega\Phi}}{2},
\end{aligned}
\tag{C 7}
$$

where $\mathbf{\Omega} = [\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1}$, $\mathbf{\Phi} = \mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}Z$.

The first term of expression (C 7) is a completed quadratic form in β_γ, which forms a Gaussian probability density and can be integrated out. The second term forms the kernel of a posterior probability density of $Z|\mathbf{X}, \mathbf{\Lambda}, \gamma$ as

$$
\begin{aligned}
&-\frac{Z'(\mathbf{I}_n + h1_n1_n')^{-1}Z}{2} \\
&-\frac{-Z'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}Z}{2}.
\end{aligned}
\tag{C 8}
$$

From expression (C 8), we obtain that $p(Z|\mathbf{X}, \mathbf{\Lambda}, \gamma) \sim N(0, \mathbf{\Sigma}_\gamma)$, with $\mathbf{\Sigma}_\gamma{}^{-1} = (\mathbf{I}_n + h1_n1_n')^{-1}$
$-(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}$.
Denote $\mathbf{\Sigma}_\gamma{}^* = \mathbf{I_n} + h1_n1_n' + \mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma'$. Then

$$
\begin{aligned}
&\mathbf{\Sigma}_\gamma{}^{-1}\mathbf{\Sigma}_\gamma{}^* \\
&= \{(\mathbf{I}_n + h1_n1_n')^{-1} - (\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1} \\
&\quad \mathbf{X}_\gamma'(\mathbf{I}_n + h1_n1_n')^{-1}\}(\mathbf{I_n} + h1_n1_n' + \mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma')
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n{'})^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\mathbf{X}_\gamma^{'}(\mathbf{I}_n + h1_n1_n{'})^{-1}\mathbf{X}_\gamma + \mathbf{\Lambda}_\gamma^{-1}]^{-1} \\
&\quad \mathbf{X}_\gamma^{'}(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma}{nh+1} + \mathbf{\Lambda}_\gamma^{-1}]^{-1}\mathbf{X}_\gamma^{'} \\
&\quad -(\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma[\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma}{nh+1} + \mathbf{\Lambda}_\gamma^{-1}]^{-1}\frac{\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma}{nh+1}\mathbf{X}_\gamma^{'} \\
&= \mathbf{I}_n + (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^{'} - (\mathbf{I_n} + h1_n1_n^{'})^{-1}\mathbf{X}_\gamma\mathbf{\Lambda}_\gamma\mathbf{X}_\gamma^{'} = \mathbf{I}_n.
\end{aligned}
$$

Therefore, $\mathbf{\Sigma}_\gamma = \mathbf{\Sigma}_\gamma{}^*$ and

$$p(Z|\mathbf{X}, \mathbf{\Lambda}, \gamma) \sim N(0, \mathbf{\Sigma}_\gamma). \qquad \text{(C 9)}$$

Hence, the joint posterior distribution of $(Z, \gamma|Y, \mathbf{X}, \mathbf{\Lambda})$ is

$$
\begin{aligned}
p(Z, \gamma|Y, \mathbf{X}, \mathbf{\Lambda}) &\propto p(Z|Y, \mathbf{X}, \mathbf{\Lambda}, \gamma)p(\gamma) \\
&\propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}}\exp(-\frac{Z^{'}\mathbf{\Sigma}_\gamma^{-1}Z}{2})\prod_{i=1}^{n}\mathrm{I}(A_i) \times \prod_{i=1}^{p}\theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}.
\end{aligned}
\qquad \text{(C 10)}
$$

(ii) Proof of equation (4.9) and equation (4.10).

From equation (C 1), we get

$$
\begin{aligned}
p(\lambda_i|\beta_i, \gamma_i) &\propto p(\beta_i|\lambda_i, \gamma_i)p(\lambda_i) \\
&\propto \lambda_i^{-\frac{1}{2}}\exp(-\frac{\beta_i^2}{2\lambda_i}) \times \lambda_i^{-(\frac{a}{2}+1)}\exp(-\frac{b}{2\lambda_i}) \qquad \text{(C 11)} \\
&\propto \lambda_i^{-(\frac{a+1}{2}+1)}\exp(-\frac{b+\beta_i^2}{2\lambda_i})
\end{aligned}
$$

Hence, the joint posterior distribution of $p(\lambda_i|\beta_i,\gamma_i)$ is $\mathrm{IG}(\frac{a+1}{2}, \frac{2}{b+\beta_i^2})$ for model I.

From equation (C 1), we also can get

$$\begin{aligned} p(\lambda_i|\beta_i,\gamma_i) &\propto p(\beta_i|\lambda_i,\gamma_i)p(\lambda_i) \\ &\propto \lambda_i^{-\frac{1}{2}} \exp(-\frac{\beta_i^2}{2\lambda_i}) \times \frac{\tau}{2} \exp(-\frac{\tau\lambda_i}{2}) \end{aligned} \tag{C 12}$$

Denote $\eta_i = \lambda_i^{-1}$, then $\frac{d\lambda_i}{d\eta_i} = -\frac{1}{\eta_i^2}$, and

$$\begin{aligned} p(\eta_i|\beta_i,\gamma_i) &\propto \frac{\tau}{\sqrt{\frac{2\pi}{\eta_i}}} \exp(-\frac{\beta_i^2\eta_i^2+\tau}{2\eta_i}) \times |-\frac{1}{\eta_i^2}| \\ &\propto \frac{\tau}{\sqrt{2\pi\eta_i^3}} \exp(-\frac{\eta_i^2+\frac{\tau}{\beta_i^2}}{\frac{2\eta_i}{\beta_i^2}}) \\ &\propto \frac{\tau}{\sqrt{2\pi\eta_i^3}} \exp(-\frac{(\eta_i-\frac{\sqrt{\tau}}{\beta_i})^2}{\frac{2\eta_i}{\beta_i^2}}) \\ &\propto \frac{\tau}{\sqrt{2\pi\eta_i^3}} \exp(-\frac{\tau}{2(\frac{\sqrt{\tau}}{\beta_i})^2} \frac{(\eta_i-\frac{\sqrt{\tau}}{\beta_i})^2}{\eta_i}) \end{aligned} \tag{C 13}$$

Hence, the joint posterior distribution of $p(\lambda_i^{-1}|\beta_i,\gamma_i)$ is $\mathrm{InvGauss}(\frac{\sqrt{\tau}}{|\beta_i|},\tau)$ for model II.

(iii) Proof of equation (4.12).

From equation (C 7), we can get equation (4.12).

(iv) Proof of equation (4.14).

From equations (C 1) and (C 10), we have

$$p(\gamma_i|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) \propto p(Z|\mathbf{X}, \mathbf{\Lambda}, \gamma)p(\gamma_i)$$
$$\propto \frac{1}{|\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_\gamma^{-1}Z}{2}) \times \theta_i^{\gamma_i}(1-\theta_i)^{1-\gamma_i}, \tag{C 14}$$

and

$$p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) \propto \frac{1}{|\mathbf{\Sigma}_{\gamma^1}|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \theta_i, \tag{C 15}$$

$$p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) \propto \frac{1}{|\mathbf{\Sigma}_{\gamma^0}|^{\frac{1}{2}}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^0}^{-1}Z}{2}) \times (1-\theta_i), \tag{C 16}$$

where $\gamma^1 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \cdots, \gamma_p)$ and $\gamma^0 = (\gamma_1, \cdots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \cdots, \gamma_p)$. As γ_i is binary, we have

$$p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) + p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) = 1. \tag{C 17}$$

From equations (C 12)-(C 14), we get

$$p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda})$$
$$= \frac{p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda})}{p(\gamma_i = 1|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda}) + p(\gamma_i = 0|\gamma_{(-i)}, Y, \mathbf{X}, Z, \mathbf{\Lambda})}$$
$$= \frac{|\mathbf{\Sigma}_{\gamma^1}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \theta_i}{|\mathbf{\Sigma}_{\gamma^1}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^1}^{-1}Z}{2}) \times \theta_i + |\mathbf{\Sigma}_{\gamma^0}|^{-\frac{1}{2}} \exp(-\frac{Z'\mathbf{\Sigma}_{\gamma^0}^{-1}Z}{2}) \times (1-\theta_i)}$$
$$= (1 + \frac{1-\theta_i}{\theta_i}\rho)^{-1},$$

where

$$\rho = |\boldsymbol{\Sigma}_{\gamma^1} \boldsymbol{\Sigma}_{\gamma^0}^{-1}|^{\frac{1}{2}} \exp\{\frac{Z'(\boldsymbol{\Sigma}_{\gamma^1}^{-1} - \boldsymbol{\Sigma}_{\gamma^0}^{-1})Z}{2}\}. \qquad \text{(C 18)}$$

Bibliography

Albert,J. and Chib,S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Stat. Assoc.*, **88**, 669-679.

Alizadeh, A.A., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. 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.

Antonov,A.V. *et al.* (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644-652.

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

Ben-Dor,A. *et al.* (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559-583.

Bo,T. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biology*, **3**, 1-17.

Brown,P.J. (1993) *Measurement, Regression, and Calibration.* Oxford: Clarendon.

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

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

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

Dettling,M. (2004) BagBoosting for Tumor Classification with Gene Expression Data. *Bioinformatics*, **20**, 3583-3593.

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

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

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

Draminski,M. et al. (2008) Monte Carlo feature selection for supervised classification. *Bioinformatics*, **24**, 110-117.

Díza-Uriarte and Andés (2006) Gene selection and classification of microarray data using random forest. *BMC Bioformatics*, **7**, 3.

Dudoit,Y., Yang,H., Callow,M. and Speed,T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77-87.

Furey,T. *et al.* (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*. London: Chapman and Hall, pp.145-158.

Genz,A. and Bretz,F. (2002) Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, **11**, 950-971.

George,E.I. and McCulloch,R.E. (1993) Variable Selection via Gibbs Sampling. *J. Amer. Stat. Assoc.*, **88**, 881-889.

George,E.I. and McCulloch,R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica*, **7**, 339-373.

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

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

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

Gupta,M. and Ibrahim,J.G. (2007) Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Amer. Stat. Assoc.*, **102**, 867-880.

Gupta,M. and Ibrahim,J.G. (2009) An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. *Statistica Sinica*, **19**, 1641-1663.

Ha, H.J., Kubagawa, H., Burrows, P.D. (1992) Molecular cloning and expression pattern of a human gene homologous to the murine mb-1 gene. *J. Immunol.*, **148**, 1526-1531.

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

Hedelfank,I. *et al.* (2001) Gene expression profiles in hereditary breast cancer. *N. Eng. J. Med.*, **344**, 539-548.

Jaeger, J., Sengupta, R. and Ruzzo, W.L. (2003) Improved gene selection for classification of microarrays. *Pac. Symp. Biocomput.*, **8**, 53-64.

Kamps,M.P., Murre,C., Sun,X.-H. and Baltimore,D. (1990) A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in pre-B ALL. *Cell*, **6**, 547-555.

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.

Khan,J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673-679.

Kingsmore,S.F.,Watson,M.L. and Seldin,M.F. (1995) Genetic mapping of the T lymphocyte-specific transcription factor 7 gene on mouse Chromosome 11. *Mamm. Genome*, **6**, 378.

Koo.J.Y., Sohn,I., Kim,S., and Lee,J.W. (2006) Structured polychotomous machine diagnosis of multiple cancer types using gene expression. *Bioinformatics*, **22**, 950-958.

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

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.

Le Cao K-A., Rossouw,D. Robert-Grani,C. and Besse,P. (2008) A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, **28**, 1-16.

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

Lee,Y. and Lee,C.K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132-1139.

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.

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.

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*. New York: Wiley, pp.342.

McLachlan,G.J., Do,K.-A. and Ambroise,C. (2004) *Analyzing microarray gene expression data*. Hoboken, New Jersey: Wiley.

McLachlan, G.J., Chevelu, J. and Zhu, J. (2008). Correcting for selection bias via cross-validation in the classification of microarray data. in N.Balakrishnan, E.Pena and M.J.Silvapulle (eds.), *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honour of Professor Pranab K. Sen*. Hayward, California: IMS Collections, Vol. 1, pp.383-395.

Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

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

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546-554.

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.

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

Plackett, R. L. (1950).Some theorems in least squares. *Biometrika*, **37**, 149-157.

Rocke,D.R., Ideker,T., Troyanskaya,O., Quackenbush,J. and Dopazo,J. (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701-702.

Sha,N., et al. (2004) Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**, 812-819.

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

Svensson,L.E.O. (2005) Monetary policy with judgment: forecast targeting. *International Journal of Central Banking*, **1**, 1-54.

Tan,A.C., Naiman,D.Q., Xu,L., Winslow,R.L. and Geman,D. (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896-3904.

Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567-6572.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104-117.

Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001) Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, **17**, 520-525.

Troyanskaya,O.G., Garber,M.E., Brown,P.O., Botstein,D. and Altman,R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1361.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci. USA*, **98**, 5116-5121.

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.

Wessels,L.F.A. et al. (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755-3762.

West,M. (2000) Bayesian factor regression models in the large p small n paradigm. in Bernardo,J.M., Bayarri,M.J., Berger,J.O., Dawid,A.P., Heckerman,D., Smith,A.F.M. and West,M. (ed.), *Bayesian Statistics 7*. Oxford: Oxford University Press, pp.733-742.

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.

Wood,I.A., Visscher,P.M. and Mengersen,K.L. (2007) Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, **23**, 1363-1370.

Woodbury, M. (1950). Inverting modified matrices. Technical Report, **42**. Statistical Research Group, Princeton University.

Yam,J., Chan,K. and Hsiao,W. (2001) Suppression of the tumorigenicity of mutant p53- transformed rat embryo fibrob-

lasts through expression of a newly cloned rat nonmuscle myosin heavy chain-B. *Oncogene*, **20**, 58-68.

Yang,A.J. and Song, X.Y. (2010) Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **26**, 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. and Bumgarner,R.E. (2003) Multi-class classification of microarray data with repeated measurements: application to cancer. *Genome Biology*, **4**, R83.

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

Yuan,M. and Lin,Y. (2005) Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, **100**, 1215-1225.

Yu,L. and Liu,H. (2004) Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.*, **5**, 1205-1224.

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

Zhou, X., Wang, X. and Dougherty, E. R. (2006) Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. *IEE Proceedings-Systems Biology*, **153**, 70-78.