BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 2 2010, pages 215–222
doi:10.1093/bioinformatics/btp638

Gene expression

Bayesian variable selection for disease classification using gene expression data

Yang Ai-Jun* and Song Xin-Yuan*

Department of Statistics, The Chinese University of Hong Kong, Hong Kong, P.R.China

Received on March 26, 2009; revised on October 30, 2009; accepted on November 7, 2009

Advance Access publication November 17, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: An important application of gene expression microarray data is the classification of samples into categories. Accurate classification depends upon the method used to identify the most relevant genes. Owing to the large number of genes and relatively small sample size, the selection process can be unstable. Modification of existing methods for achieving better analysis of microarray data is needed.

Results: We propose 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 Markov chain Monte Carlo 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 other popular methods in gene selection and classification via the well-known colon cancer and leukemia datasets in microarray literature.

Availability: A free Matlab code to perform gene selection is available at http://www.sta.cuhk.edu.hk/xysong/geneselection/.

Contact: ajyang81@gmail.com; xysong@sta.cuhk.edu.hk.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 with 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 231–549 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), non-parametric 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 (SSVS) method (George and McCulloch, 1993) has recently become popular (Gupta and Ibrahim, 2007; Lee *et al*., 2003; 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 article, we consider a multivariate Bayesian regression model together with a 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

*To whom correspondence should be addressed.

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 with 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 datasets: 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 (Le Cao and Chabrier, 2008; Le Cao *et al*., 2008; Ma *et al*., 2007; McLachlan *et al*., 2004; Nguyen and Rocke, 2002; among others). Our results show that the proposed gsg-SSVS approach reduces the number of selected genes and produces prediction accuracy comparable with those of the existing variable selection and classification methods.

This article is organized as follows. In Section 2, we briefly review the model specification based on SSVS; 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 datasets are given in Section 3. Some concluding remarks are presented in Section 4. The technical details are provided in the Supplementary Material.

2 METHODS

2.1 Model

Suppose that n independent binary random variables $Y_1,\ldots,Y_n$ 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:

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

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 $\beta=(\beta_1,\ldots,\beta_p)'$ is a p by one-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 $Z=(Z_1,Z_2,\ldots,Z_n)'$ are introduced to simplify the structure. More specifically, we define

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

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

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

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 non-zero 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 (1) is reduced to

$$Z_i=\alpha+X_{i,\gamma}\beta_\gamma+\varepsilon_i, \tag{2}$$

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), 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) includes assigning a 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 Prior distribution

The choice of the prior distributions for the unknown parameters is very important in the Bayesian SSVS approach. In this article, 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

$$\alpha\sim N(0,h), \tag{3}$$

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

$$\beta_\gamma|\gamma\sim N(0,\mathbf{H}_\gamma), \tag{4}$$

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 (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 gsg-prior, as follows

$$\beta_\gamma|\gamma\sim N(0,c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{+}), \tag{5}$$

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

$$\gamma_i\sim \text{Bernoulli}(\pi_i), \quad 0\le\pi_i\le 1, \tag{6}$$

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

2.3 Computation

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

$$p(Z,\alpha,\beta_\gamma,\gamma|Y,\mathbf{X})\propto\left[\exp\left\{-\frac{\sum_{i=1}^n(Z_i-\alpha-X_{i,\gamma}\beta_\gamma)^2}{2}\right\}\prod_{i=1}^n \mathrm{I}(A_i)\right] \quad (7)$$

$$\times\exp\left(-\frac{\alpha^2}{2h}\right)\times\left[\exp\left(-\frac{\beta_\gamma^{'}\mathbf{X}_\gamma^{'}\mathbf{X}_\gamma\beta_\gamma}{2c}\right)\prod_{i=1}^{m_\gamma}\lambda_i^{-\frac{1}{2}}\right]\times\prod_{i=1}^p\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i},$$

where A_i is either equal to $\{Z_i:Z_i>0\}$ or $\{Z_i:Z_i\le 0\}$ corresponding to $Y_i=1$ or $Y_i=0$, respectively; $\lambda_1,\ldots,\lambda_{m_\gamma}$ $(m_\gamma\le p_\gamma)$ are the non-zero 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 (Supplementary Material), the joint posterior distribution of (Z,γ) is given as follows:

$$p(Z,\gamma|Y,\mathbf{X})\propto\frac{1}{|\Sigma_\gamma|^{\frac{1}{2}}}\exp\left(-\frac{Z^{'}\Sigma_\gamma^{-1}Z}{2}\right)\prod_{i=1}^n\mathrm{I}(A_i)$$
$$\times\prod_{i=1}^p\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}, \quad (8)$$

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

The posterior distribution in (8) cannot be expressed in an explicit form; therefore, we use an MCMC technique, namely the Gibbs sampler (Geman and Geman, 1984), to generate observations from this posterior distribution. 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 $N(0,\Sigma_\gamma)\prod_{i=1}^n\mathrm{I}(A_i)$, which is a multivariate truncated normal distribution. Direct sampling from this distribution is known to be difficult. In this article, 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 $|\Sigma_\gamma|^{-\frac{1}{2}}\exp\left(-Z^{'}\Sigma_\gamma^{-1}Z\right)/2\times\prod_{i=1}^p\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}$. 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

$$p(\gamma_i|\gamma_{(-i)},Y,\mathbf{X},Z)\propto\frac{1}{|\Sigma_\gamma|^{\frac{1}{2}}}\exp\left(-\frac{Z^{'}\Sigma_\gamma^{-1}Z}{2}\right)\times\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}. \quad (9)$$

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 $\gamma^1=(\gamma_1,\ldots,\gamma_{i-1},\gamma_i=1,\gamma_{i+1},\ldots,\gamma_p)$ and $\gamma^0=(\gamma_1,\ldots,\gamma_{i-1},\gamma_i=0,\gamma_{i+1},\ldots,\gamma_p)$, and similarly define Σ_{γ^1} and Σ_{γ^0} as the Σ_γ in (8). It can be shown that (Supplementary Material):

$$p(\gamma_i=1|\gamma_{(-i)},\mathbf{X},Z)= \quad (10)$$
$$\frac{p(\gamma_i=1|\gamma_{(-i)},\mathbf{X},Z)}{p(\gamma_i=1|\gamma_{(-i)},\mathbf{X},Z)+p(\gamma_i=0|\gamma_{(-i)},\mathbf{X},Z)}=\left(1+\frac{1-\pi_i}{\pi_i}\rho\right)^{-1},$$

where

$$\rho=|\Sigma_{\gamma^1}\Sigma_{\gamma^0}^{-1}|^{\frac{1}{2}}\exp\left\{\frac{Z'(\Sigma_{\gamma^1}^{-1}-\Sigma_{\gamma^0}^{-1})Z}{2}\right\}. \quad (11)$$

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) 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 (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 $i=1,2,\ldots,n$, draw $Z_i^{(k+1)}$ from $p(Z_i^{(k)}|Z_{(-i)}^{(k)},Y,\mathbf{X},\gamma^{(k)})$.

Step (b): for $i=1,2,\ldots,p$, 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 (10) and (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 $\{(Z^{(k)},\gamma^{(k)}),k=1,2,\ldots,M\}$ 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

$$\hat{p}(\gamma_i=1|Y,\mathbf{X})=\frac{1}{M}\sum_{k=1}^M 1[\gamma_i^{(k)}=1]. \quad (12)$$

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.

2.4 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 samplebased 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 (Supplementary Material)

$$p(Y_i|Y_{(-i)},\mathbf{X})=\left(\iint p(Y_i|Y_{(-i)},\mathbf{X},Z,\gamma)^{-1}p(Z,\gamma|Y,\mathbf{X})\mathrm{d}Z\mathrm{d}\gamma\right)^{-1}. \quad (13)$$

Equation (13) enables us to use the distribution $p(Z,\gamma|Y,X)$, which 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 (13) using the generated samples $\{(Z^{(k)},\gamma^{(k)}),k=1,2,\ldots,M\}$ yields:

$$\hat{p}(Y_i|Y_{(-i)},\mathbf{X})=\frac{M}{\sum_{k=1}^{M}p(Y_i|Y_{(-i)},\mathbf{X},Z^{(k)},\gamma^{(k)})^{-1}}. \quad (14)$$

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

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

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

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

3 RESULTS

We illustrate the usefulness of the proposed gsg-SSVS approach via two well-known datasets: 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.

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 $c=10$, $\pi_i=0.005, i=1,\ldots,p$ and $h=100$. To check convergence, three chains with different initial values of Z and γ are run based on the entire dataset. The initial values $\gamma^{(0)}$ were obtained based on randomly selecting 25 genes for chains 1 and 2, and 30 genes for chain 3 (Supplementary Material) 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. Figure1a shows that the most significant genes, which are determined by the posterior gene inclusion probabilities, are almost the same for three chains. Figure 1b plots the number of selected genes versus the iteration number, and Figure 1c plots the log relative posterior probabilities of selected genes, $\log(p(\gamma|Y,\mathbf{X},Z))$, versus the iteration number. Figure 1b and c indicates 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 [Equation (12)].

Based on the entire dataset, the 18 most significant genes ranked by the posterior gene inclusion probabilities (Fig. 1a) for chain 1 are presented in Table 1. Seven of them were also selected by Ben-Dor *et al*. (2000). On the top of the genes listed in Table 1 is uroguanylin precursor Z50753. Notterman *et al*. (2001) showed that a reduction of uroguanylin might be an indication of colon tumors; and Shailubhai *et al*. (2000) reported that treatment with uroguanylin has a positive therapeutic significance to the reduction in pre-cancerous colon ploys. The second selected gene in Table 1

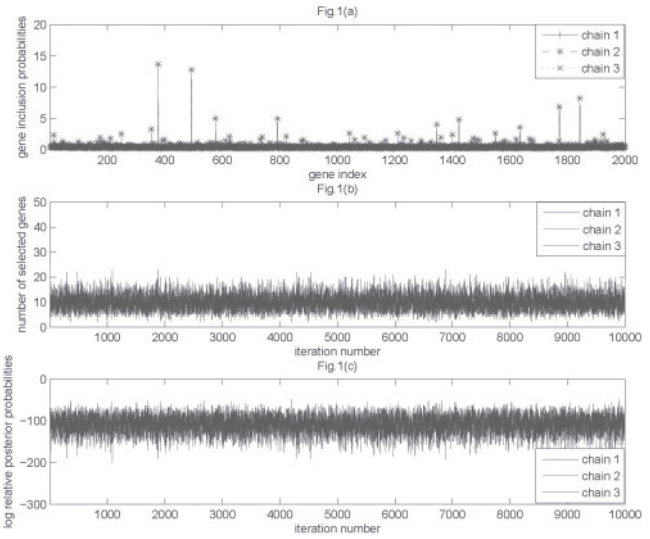

Fig. 1. (**a**) The gene inclusion probabilities (in percentages) versus the gene index; (**b** and **c**) the number of selected genes and the log-relative posterior probabilities of selected genes versus the iteration number, respectively.

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

No.	Clone ID	Gene annotation
1	Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor[a]
2	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE[a]
3	H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN)[a]
4	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)[a]
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)[a]
8	T62947	60S RIBOSOMAL PROTEIN L24)[a]
9	M36634	Human vasoactive intestinal peptide (VIP) mRNA[a]
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

[a]Ben-Dor *et al*. (2000).

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

	Method	No. of genes	LOOCV error rate
1	SVM[a]	1000 or 2000	0.0968
2	LogitBoost, optimal[b]	2000	0.1290
3	Classification tree[b]	200	0.1452
4	MAVE-LD[c]	50	0.1613
5	1-Nearest-neighbor[b]	25	0.1452
6	LogitBoost, estimated[b]	25	0.1935
7	SGLasso[c]	19	0.1290
8	LogitBoost, 100 iterations[b]	10	0.1452
9	AdaBoost, 100 iterations[b]	10	0.1613
10	gsg-SSVS	14	0.1129
11	gsg-SSVS	10	0.1129
12	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).

is R87126 (myosin heavy chain, non-muscle). The isoform B of R87126 acts as a tumor suppressor 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 the referees, this internal LOOCV procedure is biased and provides optimistic results. Therefore, an external LOOCV procedure proposed by Ambroise and McLachlan (2002) was used in our analysis. Similar to many other multivariate methods, this procedure is challenged by server memory requirements and large computational time. According to the traditional attempts to overcome these problems (Antoniadis *et al*., 2003; Le Cao and Chabrier, 2008), we perform the external LOOCV procedure as follows: (i) omit one observation of the training set; (ii) based on the remaining observations, reduce the set of available genes to the top 50 genes as ranked in terms of the t-statistic; (iii) the p^* most significant genes were rechosen from the 50 genes by our gsg-SSVS approach; and (iv) these p^* genes were used to classify the left out samples. This process was repeated for all observations in the training set until each observation had been held out and predicted exactly once. Based on the LOOCV strategy, the performance of our method with $p^*=6$, 10 and 14 are summarized in Table 2. Our method with six genes misclassified five tumor tissues (T1, T2, T30, T33, T36) and three normal tissues (N8, N34, N36). Alon *et al*. (1999), using a muscle index based on the average intensity of ESTs, misclassified five tumor tissues (T2, T30, T33, T36, T37) and three normal tissues (N8, N12, N34). Furey *et al*. (2000), applying the support vector machine (SVM) with 1000 or 2000 genes, misclassified three tumor tissues (T30, T33, T36) and three 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 gsg-SSVS approaches. Our LOOCV results have been compared with the following classification methods: 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. It is clear from the comparison that our method, which used fewer genes, is better than or comparable with 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 dataset 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 dataset 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 the Supplementary Material). 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 $\gamma^{(0)}$.

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) thresholding: floor of 100 and ceiling of 16 000; (ii) filtering: exclusion of genes with $\max/\min \leq 5$ and $(\max-\min) \leq 500$, where max and min refer, respectively, to the maximum and minimum expression levels of a particular gene across samples; and (iii) base-10 logarithmic transformation. The filtering resulted in 3571 genes. We further transformed the gene expression data to have mean 0 and SD 1 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 estimated on the entire training set are presented in Figure 2. The relevant genes selected on the basis of these probabilities are reported in Table 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

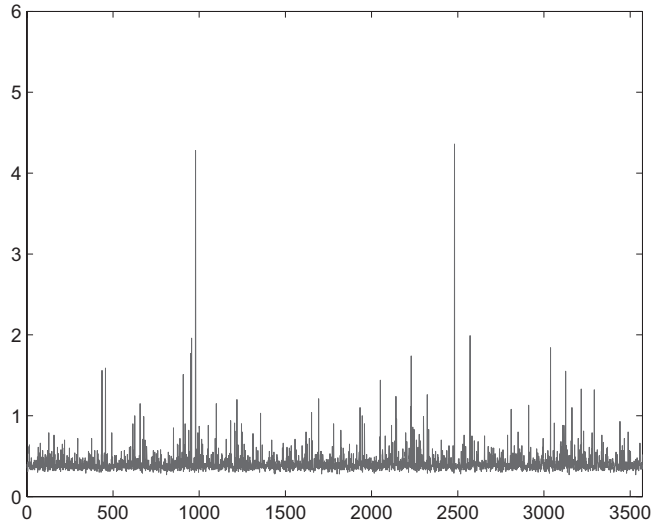

Fig. 2. shows the gene inclusion probabilities (in percentages) versus the gene index for leukemia data.

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

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

[a]Golub *et al*. (1999).
[b]Ben-Dor *et al*. (2000).

by intrinsic leucine-rich 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).

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

	Method	No. of genes	Training error rate	Test error rate
1	SVM[a]	25–1000	0.0526	0.0588–0.1176
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	14	0.0263	0.0294
8	gsg-SSVS	10	0.0263	0.0294
9	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).

The top-ranked 6, 10 and 14 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, in which the significant genes used for classification were selected from the 50 preselected genes in each LOOCV loop, was applied to get the classification error on the training set. There was one training error and one test error (the 67th observation). This 67th observation was also misclassified in Golub *et al*. (1999) and Lee *et al*. (2003). In Table 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 with 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 Bayesian model average (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 rate results related to the training set in these three methods were reported, it is not possible to provide a direct comparison of our gsg-SSVS approach with these methods.

3.3 Computational time

The computational times to run once the gsg-SSVS approach on the whole set of variables in colon cancer data and the leukemia data are, respectively, about 258 and 282 min for 60 000 iterations in a PC with Intel Core2 1.86 GHz CPU 1 GB RAM.

4 DISCUSSION AND CONCLUSION

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 technique 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 Σ_γ 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 iterations, 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 datasets, we demonstrated that the proposed gsg-SSVS approach compared favorably with other popular methods in performing disease classification.

In this article, 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, we believe that the key ideas in this article can be applied to handle multiclass problem. We assume that genes are independent. In our future research, we will extend the model to account for a correlation structure between genes.

ACKNOWLEDGEMENTS

The authors would like to thank two referees, the editor and the associate editor for their constructive comments which have substantially improved this article.

Funding: Hong Kong Special Administration Region (GRF 450508).

Conflict of Interest: none declared.

REFERENCES

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

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. *et al.* (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 Biol.*, **3**, 1–17.

Brown,P.J. *et al.* (1998) Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. B*, **60**, 627–641.

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*. Springer, New York.

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

Dudoit,Y. *et al.* (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.

George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **6**, 721–741.

Gilks,W. *et al.* (1996) *Markov Chain Monte Carlo in Practise*. Chapman and Hall, London.

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. Am. Stat. Assoc.*, **102**, 867–880.

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

Jaeger,J. *et al.* (2003) Improved gene selection for classification of microarrays. *Pac. Symp. Biocomput.*, **8**, 53–64.

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

Keely,P. *et al.* (1998) Integrins and GTPases in tumour cell growth, motility and invasion. *Trends Cell Biol.*, **8**, 101–107.

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

Lamnisos,D. *et al.* (2009) Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *J. Comput. Graph. Stat.*, **18**, 592–612.

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

Le Cao K.-A. *et al.* (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 35.

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

Li,Y. *et al.* (2002) Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, **18**, 1332–1339.

Ma,S. *et al.* (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.

McLachlan,G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, p. 342.

McLachlan,G.J. *et al.* (2004) *Analyzing Microarray Gene Expression Data*. Wiley, Hoboken, New Jersey.

McLachlan,G.J. *et al.* (2008) Correcting for selection bias via cross-validation in the classification of microarray data. In Balakrishnan,N. *et al.* (eds.) *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honour of Professor Pranab K. Sen.*, Vol. 1. IMS Collections, Hayward, CA, 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 Res.*, **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. *J. Econom.*, **143**, 291–316.

Rocke,D.R. *et al.* (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

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 Res.*, **60**, 5151–5157.

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

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

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

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

Troyanskaya,O.G. *et al.* (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1361.

Wang,Y. and Gilmore,T. (2003) Zyxin and paxillin proteins: focal adhesion plaque LIM domain proteins go nuclear. *Biochim. Biophys. Acta*, **1593**, 115–120.

West,M. (2000) Bayesian factor regression models in the large p small n paradigm. In Bernardo,J.M. *et al.* (ed.) *Bayesian Statistics 7*. Oxford University Press, Oxford, pp.733–742.

Yam,J. *et al.* (2001) Suppression of the tumorigenicity of mutant p53- transformed rat embryo fibroblasts through expression of a newly cloned rat nonmuscle myosin heavy chain-B. *Oncogene*, **20**, 58–68.

Yeung,K.Y. *et al.* (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, **21**, 2394–2402.

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