Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 37 (2004) 249–259

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Cancer classification and prediction using logistic regression with Bayesian gene selection

Xiaobo Zhou[a,b], Kuang-Yu Liu[a,b], Stephen T.C. Wong[a,b,*]

[a] *Harvard Center for Neurodegeneration and Repair—Center for Bioinformatics, Harvard Medical School, 220 Longwood Avenue, Boston, MA 02115, USA*

[b] *Radiology Department, Harvard Medical School and Brigham and Women's Hospital, 77 Francis Street, Boston, MA 02115, USA*

Received 13 June 2004
Available online 11 September 2004

Abstract

In microarray-based cancer classification and prediction, gene selection is an important research problem owing to the large number of genes and the small number of experimental conditions. In this paper, we propose a Bayesian approach to gene selection and classification using the logistic regression model. The basic idea of our approach is in conjunction with a logistic regression model to relate the gene expression with the class labels. We use Gibbs sampling and Markov chain Monte Carlo (MCMC) methods to discover important genes. To implement Gibbs Sampler and MCMC search, we derive a posterior distribution of selected genes given the observed data. After the important genes are identified, the same logistic regression model is then used for cancer classification and prediction. Issues for efficient implementation for the proposed method are discussed. The proposed method is evaluated against several large microarray data sets, including hereditary breast cancer, small round blue-cell tumors, and acute leukemia. The results show that the method can effectively identify important genes consistent with the known biological findings while the accuracy of the classification is also high. Finally, the robustness and sensitivity properties of the proposed method are also investigated.

Keywords: Gene microarray; Logistic regression; Bayesian gene selection; Cancer classification

1. Introduction

Cancer classification and prediction has become one of the most important applications of DNA microarray due to their potentials in cancer diagnostic and prognostic prediction [2,9,11,13]. Given the thousands of genes and the small number of data samples involved in microarray-based classification, gene selection is an important research problem [17]. Many gene selection algorithms have been proposed in the literature for gene classification; for example, support vector machines [10], genetic algorithms [16], Bayesian variable selection [14,33–35], the minimum description length principle for model selection [12], and the logistic regression model [3]. The logistic regression model, also known as logit in the literature, is one of the most common models for prediction, regression, and classification of disease data [6,22].

Logit-based methods have been successfully applied to cancer classification [3,19]; nevertheless, gene selection and classification based on the same logit method does not exist. Certain variable selection schemes for the logistic regression model have been proposed [4,23], but they are not suitable for microarray-like problems having large numbers of variables and small sample sizes.

Our observation is that no closed form expression for the posterior distribution of the selected genes exists for logistic regression, whereas such a closed form

* Corresponding author.
E-mail address: stephen_wong@hms.harvard.edu (S.T.C. Wong).

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2004.07.009

expression exists for linear probit regression [14]. Motivated by the Bayesian variable selection based on probit regression, we thus propose a new Bayesian method to both gene selection and classification using the logistic regression model. The basic idea of our Bayesian method is in conjunction with a logistic regression model to relate the gene expression with the class labels. Rather than fixing the number of selected genes or features, we assign a prior distribution over it. The new Bayesian-based logit method allows flexibility compared to the existing logit methods by adding pre-determined constraints, such as reducing the selected number of genes for consideration using a prior distribution. This method uses Gibbs sampling and Markov chain Monte Carlo (MCMC) algorithms to discover important genes. To implement Gibbs Sampler or MCMC search, however, we need to derive a posterior distribution of the selected genes given the observed data. Unfortunately, no closed form of such a distribution exists for logistic regression models. Thus, we adapt an approximated posterior distribution in our calculation.

After the important genes are identified, the same Bayesian-based logit method in turn will be used for cancer classification and prediction. We have evaluated the proposed method against several published microarray disease data sets, including those of hereditary breast cancer, small round blue-cell tumors, and acute leukemia. The experimental results show that the proposed Bayesian method can effectively identify important genes consistent with the known biological findings while the accuracy of the classification is high. In addition, the robustness and sensitivity properties for the Bayesian-based logit method are also investigated.

The remainder of this paper is organized as follows. In Section 2, we first formulate the problem of gene selection and classification for logistic regression, then we provide the Bayesian gene selection algorithm using Gibbs sampling and MCMC algorithm. Section 3 provides experimental results on the three different microarray data sets. Section 4 presents the conclusions. Finally, the detailed derivation of the proposed Bayesian-based logical regression method and certain procedures for efficient implementation of the proposed Bayesian method are discussed in Appendices A and B.

2. Material and methods

2.1. Material preparation for DNA microarray

DNA microarrays work by hybridization of labeled RNA or DNA in solution to DNA molecules attached at specific locations on a surface. Such arrays are often made of high-density arrays of oligonucleotide [18] or complementary DNA (cDNA) [25,26]. Such an arrangement allows a highly parallel monitoring of gene expression (mRNA abundance) patterns for thousands of genes at the same time in a single experiment. These transcription profiling techniques have been applied to study the patterns of gene expression across many experiments that survey a wide variety of cellular responses, phenotypes, conditions, and often through observations at multiple time points [7,15,18,21,25,28,29,31]. Such studies often involve two major objectives, class discover and class prediction, which can be used to develop a more complete understanding of the function, regulation, and interactions of genes and their products at RNA and protein levels. The more comprehensive knowledge may then help to delineate the underlying etiology of many diseases and improve their diagnosis and prognosis [9,11,13]. Further details of experimental procedures in sample preparation, hybridization, and washing, are documented in the above mentioned references.

2.2. Problem formulation

Assume we are interested in classifying whether a particular cancer is present or not. Let $\boldsymbol{z} = [z_1, \ldots, z_m]^{\mathrm{T}}$ denote the class labels, where $z_i = 1$ indicates sample i has the cancer, and $z_i = 0$ indicates sample i does not have the cancer. Denote $x_1, \ldots, x_n$ as the expression levels of n genes. Let $x_{i,j}$ be the measurement of the expression level of the jth gene for the ith sample. Let $\boldsymbol{X} = (x_{i,j})_{m,n}$ denote the expression levels of all genes, i.e.,

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}. \quad (1)$$

Let $\boldsymbol{X}_i \triangleq [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]$ denote the ith row of the above matrix. We model $v_i \triangleq P(z_i = 1|\boldsymbol{X})$ by using a logistic regression model given by

$$\log \frac{v_i}{1 - v_i} = \boldsymbol{X}_i \boldsymbol{\beta} \triangleq x_{i,1}\beta_1 + \cdots + x_{i,n}\beta_n, \quad i = 1, \ldots, m, \quad (2)$$

where $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_n]^{\mathrm{T}}$ contains the regression coefficients. According to [1,20], the logistical model can be rewritten as $z_i = 1(y_i > 0)$, $y_i = X_i\boldsymbol{\beta} + \log(F(t_i)/(1 - F(t_i)))$, $t_i \sim \mathcal{N}(0, \sigma^2)$, and $\sigma^2 \sim \Gamma(v/2, v/2)$, where F denotes the cumulative distribution function of the t distribution with mean 0 and variance 1. This model can be approximated by [20] $z_i = 1(y_i > 0)$, $y_i|\boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, a\sigma^{-2}I_v)$, $\sigma^2 \sim \Gamma(v/2, v/2)$, where $a \triangleq \pi^2(v - 2)/3v$ with $v = 7.3$ and I is an identity matrix or unit matrix. Then posterior computation for parameters $\boldsymbol{\beta}$, σ^2 under that approximation can be accomplished using MCMC algorithm. Motivated by this, we derived approximated posterior distribution of a indicator vector for gene selection.

Define $\boldsymbol{\gamma}$ as the $n \times 1$ indicator vector with the jth element γ_j such that $\gamma_j = 0$ if $\beta_j = 0$ (the variable is not

selected) and $\gamma_j = 1$ if $\beta_j \neq 0$ (the variable is selected). The Bayesian variable selection is to estimate $\boldsymbol{\gamma}$ from the posterior distribution $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$. Given $\boldsymbol{\gamma}$, let $\boldsymbol{\beta}_\gamma$ consists of all non-zero elements of $\boldsymbol{\beta}$ and let $\boldsymbol{X}_\gamma$ be the columns of $\boldsymbol{X}$ corresponding to those $\boldsymbol{\gamma}$ that are equal to 1. Now the problem is how to estimate $\boldsymbol{\gamma}$ and the corresponding $\boldsymbol{\beta}_\gamma$. Note that no closed form expression exists for the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{y}, X)$, and neither for $p(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{X})$.

2.3. Bayesian gene selection based on logistic regression

The indicator vector $\boldsymbol{\gamma}$ can be modeled as a realization from any prior $p(\boldsymbol{\gamma})$ on the 2^n possible values of $\boldsymbol{\gamma}$ given by $p(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \pi_i^{\gamma_i}(1-\pi_i)^{(1-\gamma_i)}$, where $\pi_i = P(\gamma_i = 1)$ is a prior probability to select the jth gene. This form is actually a Bernoulli distribution for selecting each gene.

We make the following assumptions on the priors of the parameters. First, given $\boldsymbol{\beta}_\gamma$, $\boldsymbol{X}_\gamma$, and σ^2, the likelihood of $\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \boldsymbol{X}_\gamma, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, a\sigma^2\mathcal{I})$. Then, given $\boldsymbol{\gamma}$ and σ^2, the prior for $\boldsymbol{\beta}_\gamma$ is $\boldsymbol{\beta}_\gamma \sim \mathcal{N}(0, \sigma^2\Sigma_\gamma)$, where Σ_γ is set as $(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$ in this study. Given $\boldsymbol{\gamma}$, the prior of σ^2 is assumed to be an conjugate inverse-Gamma distribution $p(\sigma^2|\boldsymbol{\gamma}) \propto \mathcal{IG}(\frac{v}{2}, \frac{v}{2})$. Moreover, the $\{\gamma_j\}_{j=1}^n$ are assumed to be independent with $p(\gamma_j = 1) = \pi_j$, $j = 1,\ldots,n$. In this paper we set $\pi_j = 15/n$ for all genes, based on the total sample number $m = 22$. If π_j is chosen to take a larger value, then we found that often times $(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$ is singular.

Here we introduce the Bayesian variable selection principle [5]. A Gibbs sampler is employed to estimate the parameters. Denote

$$S(\boldsymbol{\gamma}, \boldsymbol{y}) \triangleq v + a^{-1}\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$, $P_\gamma \triangleq \mathcal{I} - \boldsymbol{X}_\gamma M_\gamma \boldsymbol{X}_\gamma^{\mathrm{T}}$ with $M_\gamma \triangleq (\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + a\Sigma_\gamma^{-1})^{-1} = (1+a)^{-1}(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$. Define $n_\gamma = \sum_{i=1}^{n}\gamma_i$. In Appendix A, we show that

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\gamma}) &\propto \int_\sigma \left\{ \int_{\beta_\gamma} p(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}_\gamma|\sigma^2)\, \mathrm{d}p(\boldsymbol{\beta}_\gamma) \right\} p(\sigma^2)\, \mathrm{d}\sigma^2 \\ &\propto a^{-\frac{m+n_\gamma}{2}}(1+a)^{-\frac{n_\gamma}{2}} S(\boldsymbol{\gamma}, \boldsymbol{y})^{-\frac{m+v}{2}}. \end{aligned} \tag{4}$$

Then the posterior distribution of $\boldsymbol{\gamma}$ is

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) \\ &\propto a^{-\frac{m+n_\gamma}{2}}(1+a)^{-\frac{n_\gamma}{2}} S(\boldsymbol{\gamma}, \boldsymbol{y})^{-\frac{m+v}{2}} \prod_{j=1}^{n} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}. \end{aligned} \tag{5}$$

In Appendix A, we also show the posterior distribution of σ^2 and $\boldsymbol{\beta}$ are, respectively, given by

$$p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}_\gamma) \propto \mathcal{IG}\left(\frac{m+v}{2}, \frac{S(\boldsymbol{\gamma}, \boldsymbol{y})}{2}\right), \tag{6}$$

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}_\gamma, \sigma^2) \propto \mathcal{N}(H_\gamma, a\sigma^2 M_\gamma), \tag{7}$$

where $H_\gamma \triangleq M_\gamma \boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y} = (1+a)^{-1}(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y}$. Based on the posterior distribution (5), a Gibbs sampler can be employed to estimate all the parameters. We use the following Gibbs sampling algorithm to estimate $\{\boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2\}$.

- Draw $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma}|\boldsymbol{y})$ in (5). In fact, we sample each γ_j independently from

$$p(\gamma_j|\boldsymbol{y}, \gamma_{i\neq j}) \propto a^{-\frac{m+n_\gamma}{2}}(1+a)^{-\frac{n_\gamma}{2}} S(\boldsymbol{\gamma}, \boldsymbol{y})^{-\frac{m+v}{2}} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}, \quad j = 1, \ldots, n. \tag{8}$$

- Draw σ^2 from $p(\sigma^2|\boldsymbol{y},\boldsymbol{\gamma})$ in (6).
- Draw $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{y})$ in (7).
- Draw z_i, $i = 1, \ldots, m$ from a truncated normal distribution as follows [24]: $y_i|\boldsymbol{\beta}, z_i = 1 \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, a\sigma^2)1_{\{y_i>0\}}$, $y_i|\boldsymbol{\beta}, z_i = 0 \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, a\sigma^2)1_{\{y_i<0\}}$.

In this study, 25,000 Gibbs iterations are implemented with the first 5000 as burn-in period. We obtain the Monte Carlo samples as $\{\boldsymbol{\gamma}^{(t)}, t = 1, \ldots, T\}$, where $T = 25{,}000$. Finally, we count the number of times that each gene appears in $\{\boldsymbol{\gamma}^{(t)}, t = 5001, \ldots, 25{,}000\}$. We define the appearance frequency of a gene as the number of appearances of this gene divided by the total iteration (i.e., 20,000 here). The genes with the highest appearance frequencies play the strongest role in predicting the target gene. We will discuss some implementation issues in the Appendix B.

2.4. Cancer classification and prediction using the strongest genes

Now assume the genes corresponding to non-zero $\boldsymbol{\gamma}$ are the strongest genes obtained by the above Bayesian variable-selection algorithm. We still use X_γ to denote the profiles of these strongest genes. For fixed $\boldsymbol{\gamma}$, we again use a Gibbs sampler to estimate the linear regression coefficients $\boldsymbol{\beta}$ as follows: First draw $\boldsymbol{\beta}_\gamma$ according to (7), then draw σ^2 according to (6) and iterate the two steps. In this study, 1500 iterations are implemented with the first 500 as the burn-in period. Thus we obtain the Monte Carlo samples $\{\tilde{\boldsymbol{\beta}}_\gamma^{(t)}, \tilde{\sigma}^{2(t)}, t = 1, \ldots, \tilde{T}\}$. We then predict the tested sample by $P(y = 1|X_\gamma) = \frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}} \exp\{X_\gamma\tilde{\boldsymbol{\beta}}_\gamma^{(t)}\}/(1+\exp\{X_\gamma\tilde{\boldsymbol{\beta}}_\gamma^{(t)}\})$. If we consider computational complexity, an alternative approach is iterative methods such as the Newton–Raphson procedure which can be adopted to obtain the maximum likelihood estimate of $\boldsymbol{\beta}_\gamma$ [8], then we can predict the tested sample by

$$P(y = 1|X_\gamma) = \frac{\exp\{X_\gamma\boldsymbol{\beta}_\gamma\}}{1+\exp\{X_\gamma\boldsymbol{\beta}_\gamma\}}. \tag{9}$$

2.5. Pre-selection method

If there are 3000 gene variables, then for each iteration we have to estimate $\boldsymbol{\beta}_\gamma$ 3000 times because we need to sample γ_j for each gene according to (8). The computational complexity of the Bayesian gene selection algorithm in the previous section is very high. Hence, some fast algorithms must be developed to speed up the computation. Here we pre-select some genes using the following procedure. The other fast implementation issues are discussed in Appendix B.

Suppose that the total number of genes is p, and we will only consider $n < p$ candidates in the Bayesian selection algorithm. We next discuss how to pre-select the n genes using an F-test. In pattern recognition, we usually adopt the following criterion: the smaller is the sum of squares within groups and the bigger is the sum of squares between groups, then it is expected a better classification accuracy. Therefore we can define a score using the above two statistics to pre-select genes, i.e., the ratio of the between-group to within-group sum of squares:

$$R(j) \triangleq \frac{\sum_{i=1}^{m}\sum_{k=0}^{K-1} 1_{(y_i=k)}(\bar{x}_{k,j}-\bar{x}_j)^2}{\sum_{i=1}^{m}\sum_{k=0}^{K-1} 1_{(y_i=k)}(x_{i,j}-\bar{x}_{k,j})^2}, \quad 1 \leqslant j \leqslant p, \tag{10}$$

where K the number of classes; p is the total number of original genes (Note that the number of genes n used in the Bayesian selection procedure is much smaller than p); $\bar{x}_j$ denotes the average expression level of gene j across all samples; and $\bar{x}_{k,j}$ denotes the average expression level of gene j across the samples belonging to class k where class k corresponds to $\{y_i = k\}$; and the indicator function 1_Ω is equal to one if event Ω is true and zero otherwise. We select a threshold $\aleph$ and keep those genes j such that $R(j) \geqslant \aleph$. The pre-selection procedure yields n genes such that $R(j) \geqslant \aleph$.

3. Experimental results

3.1. Breast cancer data

In our first experiment, we will focus on hereditary breast cancer data, which can be downloaded from the web page for the original paper [11]. In [11], cDNA microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. Twenty-two breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3226 genes for each tumor sample. We use our methods to classify BRCA1, BRCA2, and sporadic. The ratio data have been truncated from below at 0.1 and above at 20. Log of the ratio data are employed to test the proposed gene selection method. The cross-validation (leave-one-out) method is employed to compute all classification errors in this paper. The number of preselected genes are 473 in this data set.

Table 1 lists the strongest genes using the proposed approximate Bayesian gene selection method. Gene 10 (Clone ID: 26184, phosphofructokinase, platelet) is the strongest gene. The gene TOB1 (Clone ID 823940) is the top 3 gene listed in Table 1 [14]. Gene 1008 (Clone ID: 897781, keratin 8) is also listed in the top 20 genes. These results are consistent with other references [11].

Table 1
The top 20 important genes using the proposed gene selection algorithm for breast cancer data ($\pi_i = 15/n$)

Gene No.	Frequency	Index No. (Clone ID)	Gene description
1	0.3103	10 (26184)	Phosphofructokinase, platelet
2	0.1621	118 (47542)	Small nuclear ribonucleoprotein D1 polypeptide (16kD)
3	0.1399	336 (823940)	Transducer of ERBB2, 1 (TOB1)
4	0.1338	2699 (44180)	α-2-macroglobulin
5	0.1335	2761 (47884)	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)
6	0.1330	742 (183200)	Fumarylacetoacetate
7	0.1305	2382 (21652)	Catenin (cadherin-associated protein), α1 (102 kDa)
8	0.1289	2018 (139354)	ESTs
9	0.1279	157 (809981)	Glutathione peroxidase 4 (phospholipid hydroperoxidase)
10	0.1260	739 (214068)	GATA-binding protein 3
11	0.1260	1120 (841617)	Human mRNA for ornithine decarboxylase antizyme, ORF 1 and ORF 2
12	0.1251	2272 (309583)	ESTs
13	0.1250	1620 (137638)	ESTs
14	0.1246	1999 (247818)	ESTs
15	0.1243	1859 (307843)	ESTs
16	0.1241	439 (160793)	Discs, large (*Drosophila*) homolog 1
17	0.1234	2734 (46019)	Minichromosome maintenance deficient (*S. cerevisiae*) 7
18	0.1233	247 (725680)	Transcription factor AP-2γ (activating enhancer-binding protein 2γ)
19	0.1233	3009 (366647)	Butyrate response factor 1 (EGF-response factor 1)
20	0.1230	2423 (26082)	Butyrate response factor 1 (EGF-response factor 1)

Using the top 5, 10, and 15 genes for classification, it is seen that the classification error based 5 genes and 10 genes is zero. Note that there is one error in the original paper [11]. There is one error using 15 genes, which is likely due to the small sample size. The conditional probabilities based the three criteria using top 10 genes are listed in Table 2. These are very close to the true label values (namely, 0 and 1). The Eq. (9) is employed to predict cancers in this study.

Table 2
The estimated probabilities of each sample for breast cancer data using the proposed Bayesian gene algorithm ($\pi_i = 15/n$)

Sample index No.	True label	$P(y = 1\|X)$	Prediction
1	0	0.0060	0
2	0	0.0000	0
3	0	0.0000	0
4	0	0.0021	0
5	0	0.0000	0
6	0	0.0237	0
7	1	1.0000	1
8	1	0.9709	1
9	1	1.0000	1
10	1	1.0000	1
11	1	0.9915	1
12	1	1.0000	1
13	1	1.0000	1
14	1	1.0000	1
15	1	0.9999	1
16	1	0.9908	1
17	1	0.9359	1
18	0	0.0000	0
19	1	1.0000	1
20	1	1.0000	1
21	1	1.0000	1
22	1	1.0000	1
No. of misclassification			0

Table 4
The estimated probabilities of each sample for SRBCT data using the proposed algorithm ($\pi_i = 15/n$)

Sample index No.	True label	$P(y = 1\|X)$	Prediction
1	0	0.0000	0
2	0	0.0001	0
3	0	0.0001	0
4	0	0.0000	0
5	0	0.0000	0
6	0	0.0000	0
7	0	0.0000	0
8	0	0.0000	0
9	0	0.0000	0
10	0	0.0060	0
11	0	0.0032	0
12	0	0.0025	0
13	0	0.0000	0
14	0	0.0000	0
15	0	0.0115	0
16	0	0.0000	0
17	0	0.0001	0
18	0	0.0001	0
19	0	0.0000	0
20	0	0.0000	0
21	0	0.0042	0
22	0	0.0002	0
23	0	0.0002	0
24	1	0.9928	1
25	1	1.0000	1
26	1	0.9954	1
27	1	1.0000	1
28	1	1.0000	1
29	1	1.0000	1
30	1	1.0000	1
31	1	1.0000	1
32	1	1.0000	1
33	1	1.0000	1
34	1	1.0000	1
35	1	0.9999	1
No. of misclassification			0

Table 3
The top 20 important genes selected using the proposed Bayesian gene selection algorithm for SRBCT data ($\pi_i = 15/n$)

Gene No.	Frequency	Index No. (Clone ID)	Gene description
1	0.1403	1389 (770394)	Fc fragment of IgG, receptor, transporter, α
2	0.1310	52 (50359)	Mannose phosphate isomerase
3	0.1250	1873 (166195)	Ribonuclease/angiogenin inhibitor
4	0.1240	1914 (824704)	Mannose phosphate isomerase
5	0.1195	545 (1435862)	Antigen identified by monoclonal antibodies 12E7, F21 and O13
6	0.1138	842 (262231)	No name
7	0.1128	1093 (812965)	v-myc avian myelocytomatosis viral oncogene homolog
8	0.1108	246 (377461)	Caveolin 1, caveolae protein, 22kD
9	0.1105	812 (166236)	Glucose-6-phosphate dehydrogenase
10	0.1088	153 (383188)	Recoverin
11	0.1050	137 (486175)	Solute carrier family 16 (monocarboxylic acid transporters), member 1
12	0.1000	2157 (244637)	*Homo sapiens* mRNA full length insert cDNA clone EUROIMAGE 45620
13	0.0995	1088 (85171)	ADP-ribosylation factor 4
14	0.0985	2050 (295985)	ESTs
15	0.0985	976 (786084)	Chromobox homolog 1 (*Drosophila* HP1 β)
16	0.0985	742 (812105)	Transmembrane protein
17	0.0978	1601 (629896)	Microtubule-associated protein 1B
18	0.0960	823 (134748)	Glycine cleavage system protein H (aminomethyl carrier)
19	0.0933	255 (325182)	Cadherin 2, N-cadherin (neuronal)
20	0.0922	1862 (789376)	Thioredoxin reductase 1

3.2. Small round blue-cell tumors

This experiment focuses on the small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS) in [13]. We classify the rhabdomyosarcoma and neuroblastoma tumors. The data set for the two cancers is composed of 2308 genes, and the sample consists of 35 tumors, 23 for RMS, and 12 for NB. The ratio data has been truncated from below at 0.01. The number of preselected genes are 282 in this data set.

Table 3 lists the strongest genes using the proposed approximate Bayesian gene selection method. Gene 1389 (clone ID 770394) is top important gene in the list. It is also an important gene listed in [13]. A number of other previously noted genes also appear [13,32]: gene 545 (Clone ID 1435862), gene 246 (Clone ID 377461), gene 153 (Clone ID 383188), gene 2050 (Clone ID 295985), gene 742 (Clone ID 812105), gene 1601 (Clone ID 629896), and gene 255 (clone ID 325182). The conditional probabilities based the three criteria using top 10 genes are listed in Table 4. These conditional probabilities are very close to the true label values. Using the top 5, 10, and 15 genes for classification based on the three criteria, no error is found.

3.3. Acute leukemia data

The leukemia data of [9] is publicly available the paper's original website. The microarray data contains 7129 human genes, sampled from 72 cases of cancer, of which 38 are of type B-cell ALL, 9 are of type T-cell ALL, and 25 of type AML. The data are preprocessed as recommended in [30]: gene values are truncated from

Table 6
The estimated probabilities of each sample for leukemia data using the proposed algorithm ($\pi_i = 15/n$)

Sample index No.	True label	$P(y = 1\|X)$	Prediction
1	0	0.0006	0
2	0	0.0004	0
3	0	0.0000	0
4	0	0.0000	0
5	0	0.0000	0
6	0	0.0000	0
7	0	0.0000	0
8	0	0.0000	0
9	0	0.0000	0
10	0	0.0000	0
11	0	0.0000	0
12	0	0.0006	0
13	0	0.0000	0
14	0	0.0007	0
15	0	0.0000	0
16	0	0.0000	0
17	0	0.0025	0
18	0	0.0005	0
19	0	0.0000	0
20	0	0.0000	0
21	0	0.0010	0
22	0	0.0007	0
23	0	0.0029	0
24	0	0.0000	0
25	0	0.0001	0
26	0	0.0000	0
27	0	0.0000	0
28	1	0.9963	1
29	1	0.9992	1
30	1	1.0000	1
31	1	1.0000	1
32	1	0.9931	1
33	1	0.9935	1
34	1	1.0000	1
35	1	1.0000	1
36	1	0.9999	1
37	1	1.0000	1
38	1	1.0000	1
No. of misclassification			0

Table 5
The top 20 important genes selected using the proposed Bayesian gene selection algorithm for acute leukemia data ($\pi_i = 15/n$)

Gene No.	Frequency	Index No.	Gene description
1	0.1153	4211	Vascular endothelial growth factor receptor 1 precursor
2	0.1108	5772	C-myb gene extracted from Human (c-myb) gene, complete primary cds
3	0.1093	2354	CCND3 cyclin D3
4	0.1083	1144	SPTAN1 spectrin, α, non-erythrocytic 1 (α-fodrin)
5	0.1038	1928	Oncoprotein 18 (Op18) gene
6	0.1035	4167	ALDR1 aldehyde reductase 1 (low K_m aldose reductase)
7	0.1027	804	Macmarcks
8	0.1008	6281	MYL1 myosin light chain (alkali)
9	0.1008	4398	DNMT DNA methyltransferase
10	0.1003	1630	Inducible protein mRNA
11	0.1000	1882	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)
12	0.0983	1834	CD33 CD33 antigen (differentiation antigen)
13	0.0978	5501	TOP2B topoisomerase (DNA) II β (180 kDa)
14	0.0978	2348	ACADM acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain
15	0.0978	1120	SNRPN small nuclear ribonucleoprotein polypeptide N
16	0.0975	5039	LEPR leptin receptor
17	0.0970	6855	TCF3 transcription factor 3 (E2A immunoglobulin enhancer-binding factors E12/E47)
18	0.0963	6279	GB DEF = PTX3 gene promotor region
19	0.0958	3258	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
20	0.0948	1704	ADA adenosine deaminase

below at 100 and from above at 16,000; genes having the ratio of the maximum over the minimum less than 5 or the difference between the maximum and the minimum less than 500 are excluded; and finally the base-10 logarithm is applied to the 3571 remaining genes. Here we consider the 38-tumor sample, splitting it between ALL (11) and AML (27). The number of preselected genes are 356 in this data set.

Table 5 lists the 20 strongest genes based on the proposed approximate Bayesian gene selection. The index number is the Clone ID in this data set. The top ten genes are gene 4211, gene 5772, gene 2354, gene 1144, gene 1928, gene 4167, gene 804, gene 6281, gene 4398, and gene 1630. Genes 5772, gene 1882, gene 1834, gene 1630, and gene 5772 are also listed in [9]. The conditional probabilities based the three criteria using top 10 genes are listed in Table 6. Again we saw that these conditional probabilities are very close to the true label values. Using the top 5, 10, and 15 genes for classification, no error is found. We also test the 34 samples, and one error is found, but there are four error in original paper [9].

4. Sensitivity and robustness

To check the sensitivity and robustness of our algorithms, we have added white Gaussian noise with different variances to the breast cancer data and re-applied our algorithms to the contaminated data. The strongest genes are listed in Table 7. It is seen gene 10 (phosphofructokinase, platelet) and gene 336 TOB1 remain very important for different noise levels. The results indicate that the proposed methods are not sensitive to the different noise levels.

To check the sensitivity to the prior distributions, we have re-run the algorithms for $\pi_i = 10/n$. According to Table 8, most of the selected genes are the same as the gene list in Table 1, hence it is seen that the proposed gene selection method is robust to the prior setting.

Table 8
The top 20 important genes selected using the proposed Bayesian gene selection algorithm for breast cancer data with different prior $\pi_i = 10/n$

Gene No.	Frequency	Index No.	Clone ID
1	0.1409	10	26184
2	0.1210	118	47542
3	0.0980	336	823940
4	0.0970	858	783729
5	0.0927	258	324210
6	0.0920	733	134748
7	0.0885	2699	44180
8	0.0870	2018	139354
9	0.0867	955	950682
10	0.0862	1417	825478
11	0.0850	1466	767817
12	0.0830	2428	26184
13	0.0820	1120	841617
14	0.0818	272	47681
15	0.0808	1859	307843
16	0.0808	1008	897781
17	0.0805	1200	811930
18	0.0800	116	754998
19	0.0798	955	950682
20	0.0795	1531	711826

Table 7
The top 20 important genes selected using the proposed Bayesian gene selection algorithm for breast cancer data for different noise levels ($\pi_i = 15/n$)

Gene No.	$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.5$	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.1411	10 (26184)	0.1356	10 (26184)	0.1467	10 (26184)
2	0.1273	118 (47542)	0.1213	118 (47542)	0.1315	118 (47542)
3	0.0980	336 (823940)	0.1000	336 (823940)	0.0975	272 (47681)
4	0.0920	955 (950682)	0.0940	2699 (44180)	0.0970	336 (823940)
5	0.0907	2699 (44180)	0.0935	955 (950682)	0.0943	258 (324210)
6	0.0905	1443 (566887)	0.0927	2428 (26184)	0.0920	2428 (26184)
7	0.0887	2428 (26184)	0.0922	858 (783729)	0.0910	733 (134748)
8	0.0885	2259 (814270)	0.0887	1179 (788721)	0.0907	858 (783729)
9	0.00882	733 (134748)	0.0860	1443 (566887)	0.0902	1120 (841617)
10	0.0882	585 (41356)	0.0855	585 (41356)	0.0895	1620 (137638)
11	0.0865	1179 (788721)	0.0845	1999 (247818)	0.0885	1443 (566887)
12	0.0850	496 (376516)	0.0845	258 (324210)	0.0875	2734 (46019)
13	0.0845	3009 (366647)	0.0838	733 (134748)	0.0848	1446 (767817)
14	0.0845	1999 (247818)	0.0825	1466 (767817)	0.0848	1179 (788721)
15	0.0845	1008 (897781)	0.0823	1531 (711826)	0.0838	2699 (44180)
16	0.0843	3010 (366824)	0.0823	1120 (841617)	0.0835	809 (810899)
17	0.0843	2387 (22230)	0.0820	2734 (46019)	0.0810	556 (212198)
18	0.0838	2259 (814270)	0.0820	2423 (26082)	0.0808	496 (376516)
19	0.0830	258 (324210)	0.0810	272 (47681)	0.0805	1008 (897781)
20	0.0828	1766 (239958)	0.0800	1008 (897781)	0.0805	1466 (767817)

4.1. Comparisons

Various gene selection methods and classifiers for cancer classification have been proposed. In particular, there is strong evidence that Bayesian gene selection is effective [14]. Regarding classification, the linear probit (LProbit) [14], and logistical regression with AIC (BIC and MDL) gene selection [36] have proved effective. Using the breast-cancer data set, we will compare the performance of these classifiers when used in conjunction with the previously proposed Bayesian gene-selection methods and the logistic method developed in this paper. We summarize linear probit based and logistical regression with AIC-based gene selection, along with the corresponding classifiers.

Probit gene selection and classification [14]: the relation between the class label y_i and the gene expression levels x_i is modeled by using a probit regression model which yields $P(y_i = 1|x_i) = \Phi(\boldsymbol{x}_i\boldsymbol{\beta})$, $i = 1, \ldots, m$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)^{\mathrm{T}}$ is the vector of regression parameters and Φ is the standard normal cumulative distribution function. Gene selection based on probit regression is similar to that of logistic regression using Gibbs sampling. The difference is the posterior distribution of $p(\gamma|z)$, see [14]. After obtaining the strongest genes, we can estimate $P(y = 1|\boldsymbol{X})$ using Gibbs sampling for the probit regression classifier. The model of logistical regression with AIC-based gene selection and classification is similar to the model used in this paper. The difference is that we approximated $p(\gamma|z)$ by using information theory AIC criterion.

Table 9 lists the top 20 genes selected by probit regression and logistic regression with AIC-based gene selection, respectively. Gene 10 (phosphofructokinase, platelet) and gene 336 TOB1 are important genes for all methods, and also quite a few genes are the same for all methods. The misclassification numbers using the three classifiers (Logit, LProbit, and Logit-AIC) for three gene selection methods (logit, probit, Logit-AIC) with 5, 10, and 15 top genes are as follows: No error is found for all of the classifiers based on Small Round Blue-cell Tumors and acute leukemia data, but one error for all of the classifiers based on breast cancer data. Therefore, we conclude that the proposed approach in this study is at least comparative with the other two.

5. Conclusion

In this work, we proposed a Bayesian approach to gene selection using the logistic regression model. The basic idea of our approach is in conjunction with a logistic regression model to relate the gene expression with the class labels. Rather than fixing the number of selected genes or features, we assigned a prior distribution over it. The approach creates additional flexibility by allowing the imposition of constraints, such as not allowing the dimension to be too big by using this prior. We use Gibbs sampling and MCMC methods to discover important genes. In order to implement Gibbs Sampler and MCMC search, we derived a posterior distribution of selected genes given the observed data. Once important genes are identified, the same logistic regression model was employed for cancer classification. For practicality, we

Table 9
The top 20 important genes selected using linear probit regression [14] and logistic regression with Bayesian gene selection based on AIC criterion [36] for breast cancer data

Gene No.	Probit		Logit-AIC	
	Frequency	Index No. (Clone ID)	Frequency	Index No. (Clone ID)
1	0.0860	1008 (897781)	0.1454	1008 (897781)
2	0.0840	336 (823940)	0.1394	496 (376516)
3	0.0780	10 (26184)	0.1340	336 (823940)
4	0.0750	1068 (840702)	0.1331	2699 (44180)
5	0.0710	496 (376516)	0.1240	2761 (47884)
6	0.0690	118 (47542)	0.1167	742 (183200)
7	0.0660	3009 (366647)	0.1044	2382 (21652)
8	0.0660	585 (293104)	0.1003	2018 (139354)
9	0.0620	523 (28012)	0.9025	157 (809981)
10	0.0610	556 (212198)	0.0545	739 (214068)
11	0.0590	1999 (247818)	0.0483	1120 (841617)
12	0.0550	2423 (26082)	0.0473	2272 (309583)
13	0.0540	498 (667598)	0.0472	1620 (137638)
14	0.0520	140 (30093)	0.0463	1999 (247818)
15	0.0510	1277 (73531)	0.0433	1859 (307843)
16	0.0500	955 (950682)	0.0426	439 (160793)
17	0.0500	272 (47681)	0.0424	2734 (46019)
18	0.0490	2734 (46019)	0.0419	247 (725680)
19	0.0490	1859 (307843)	0.0414	3009 (366647)
20	0.0480	555 (548957)	0.0405	2423 (26082)

also investigated efficient implementational issues of these methods.

The proposed Bayesian based logit method was tested on data sets including hereditary breast cancer data, small round blue-cell tumor data, and acute leukemia tumor data. The experimental results show that the proposed method can effectively find genes that are consistent with the existing biological knowledge and with high accuracy. Our experimental results also show that the proposed gene selection has robustness and sensitivity properties. Note that for the breast cancer data and acute leukemia tumor data, the classification accuracy of the proposed Bayesian based logit method is much better than that reported in the original papers, and for small round blue-cell tumor data, the proposed logit approach has a perfect classification result.

Acknowledgment

This research was supported by the Center of Bioinformatics Program grant of Harvard Center of Neurodegeneration and Repair, Harvard Medical School, Boston, USA.

Appendix A. Derivation of (4), (6), and (7)

For notational convenience, we denote $\eta \triangleq \sigma^2$ and $a \triangleq \pi^2(v-2)/3v$, where $v = 7.3$. Since

$$\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \eta, \boldsymbol{X}_\gamma \sim \mathcal{N}(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, a\eta\boldsymbol{I}_m),$$

$$\beta_\gamma|\eta \sim \mathcal{N}(0, \eta\Sigma_\gamma), \text{ and } \eta \sim \mathcal{IG}\left(\frac{v}{2}, \frac{v}{2}\right),$$

where Σ_γ is set as $(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$ in this study, $\mathcal{IG}$ is the inverse Gamma distribution, then we have

$$p(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \eta, \boldsymbol{X}_\gamma) \propto a^{-\frac{m}{2}}\eta^{-\frac{m}{2}} \times \exp\left\{-\frac{1}{2a\eta}(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\right\}, \tag{A.1}$$

$$p(\boldsymbol{\beta}_\gamma|\eta,\gamma) \propto \eta^{-\frac{n_\gamma}{2}}|\Sigma_\gamma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\eta}\boldsymbol{\beta}_\gamma^{\mathrm{T}}(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)\boldsymbol{\beta}_\gamma\right\}, \tag{A.2}$$

$$p(\eta) \propto b(v)\eta^{-(v/2+1)}\exp\left(-\frac{v}{2\eta}\right), \tag{A.3}$$

where $b(v) \triangleq \left(\frac{v}{2}\right)^{\frac{v}{2}}(\Gamma(\frac{v}{2}))^{-1}$. According to the Bayesian theorem, we have

$$p(\boldsymbol{\beta}_\gamma, \eta|\boldsymbol{X}_\gamma, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \eta, \boldsymbol{X}_\gamma)p(\boldsymbol{\beta}_\gamma, \eta) = p(\boldsymbol{y}|\boldsymbol{\beta}_\gamma, \eta, \boldsymbol{X}_\gamma)p(\boldsymbol{\beta}_\gamma|\eta)p(\eta). \tag{A.4}$$

Using ((A.1)–(A.4)), we have

$$p(\boldsymbol{\beta}_\gamma, \eta|\boldsymbol{X}_\gamma, \boldsymbol{y}) \propto a^{-\frac{m}{2}}\eta^{-\frac{m+v+n_\gamma+2}{2}}|\Sigma_\gamma|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2a\eta}\left[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma - (\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\mathrm{T}}\boldsymbol{y}\right.\right. \tag{A.5}$$

$$\left.\left. + (\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\mathrm{T}}(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma) + a\boldsymbol{\beta}_\gamma\Sigma_\gamma^{-1}\boldsymbol{\beta}_\gamma\right] - \frac{v}{2\eta}\right\}. \tag{A.6}$$

Define

$$M_\gamma^{-1} \triangleq \boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + a\Sigma_\gamma^{-1} \Longleftrightarrow M_\gamma = \left(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + a\Sigma_\gamma^{-1}\right)^{-1}, \tag{A.7}$$

$$H_\gamma \triangleq M_\gamma\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y} = \left(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + a\Sigma_\gamma^{-1}\right)^{-1}\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y}, \tag{A.8}$$

$$P_\gamma \triangleq \mathcal{I} - \boldsymbol{X}_\gamma M_\gamma \boldsymbol{X}_\gamma^{\mathrm{T}}. \tag{A.9}$$

Then we have the following equations

$$M_\gamma^{-1}H_\gamma = \boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y},$$

$$\boldsymbol{\beta}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{\beta}_\gamma^{\mathrm{T}}M_\gamma^{-1}H_\gamma,$$

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma H_\gamma = H_\gamma M_\gamma^{-1}H_\gamma.$$

Note that the following equality holds:

$$\begin{aligned}
&\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma - (\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\mathrm{T}}\boldsymbol{y} + (\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^{\mathrm{T}}(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma) + a\boldsymbol{\beta}_\gamma\Sigma_\gamma^{-1}\boldsymbol{\beta}_\gamma \\
&= \boldsymbol{y}^{\mathrm{T}}\left(\mathcal{I} - \boldsymbol{X}_\gamma M_\gamma\boldsymbol{X}_\gamma^{\mathrm{T}}\right)\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma H_\gamma - H_\gamma^{\mathrm{T}}M_\gamma^{-1}\boldsymbol{\beta}_\gamma \\
&\quad - \boldsymbol{\beta}_\gamma^{\mathrm{T}}M_\gamma^{-1}H_\gamma + \boldsymbol{\beta}_\gamma^{\mathrm{T}}\left(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma + a\Sigma_\gamma^{-1}\right)\boldsymbol{\beta}_\gamma \\
&= \boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma H_\gamma - H_\gamma^{\mathrm{T}}M_\gamma^{-1}\boldsymbol{\beta}_\gamma - \boldsymbol{\beta}_\gamma^{\mathrm{T}}M_\gamma^{-1}H_\gamma + \boldsymbol{\beta}_\gamma^{\mathrm{T}}M_\gamma\boldsymbol{\beta}_\gamma \\
&= \boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y} + (\boldsymbol{\beta}_\gamma - H_\gamma)^{\mathrm{T}}M_\gamma^{-1}(\boldsymbol{\beta}_\gamma - H_\gamma).
\end{aligned} \tag{A.10}$$

Then (A.6) becomes

$$\begin{aligned}
p(\boldsymbol{\beta}_\gamma, \eta|\boldsymbol{X}_\gamma, \boldsymbol{y}) &\propto a^{-\frac{m}{2}}\eta^{-\frac{m+v+n_\gamma+2}{2}}|\Sigma|^{-\frac{1}{2}} \\
&\times \exp\left\{-\frac{1}{2a\eta}\left[\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y} + (\boldsymbol{\beta}_\gamma - H_\gamma)^{\mathrm{T}}M_\gamma^{-1}(\boldsymbol{\beta}_\gamma - H_\gamma)\right] - \frac{v}{2\eta}\right\} \\
&= a^{-\frac{m}{2}}\eta^{-\frac{m+v+n_\gamma+2}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{a^{-1}\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y} + v}{2\eta}\right\} \\
&\times \exp\left\{-\frac{1}{2a\eta}(\boldsymbol{\beta}_\gamma - H_\gamma)^{\mathrm{T}}M_\gamma^{-1}(\boldsymbol{\beta}_\gamma - H_\gamma)\right\}.
\end{aligned} \tag{A.11}$$

After integrating out $\boldsymbol{\beta}_\gamma$ in (A.11), we have

$$\begin{aligned}
p(\eta|\boldsymbol{X}_\gamma, \boldsymbol{y}) &\propto \int_{\boldsymbol{\beta}_\gamma} p(\boldsymbol{\beta}_\gamma, \eta|\boldsymbol{X}_\gamma, \boldsymbol{y})\,\mathrm{d}\boldsymbol{\beta}_\gamma \\
&\propto a^{-\frac{m+n_\gamma}{2}}\eta^{-\frac{m+v+2}{2}}|\Sigma_\gamma|^{-\frac{1}{2}}|M_\gamma|^{\frac{1}{2}} \\
&\times \exp\left\{-\frac{a^{-1}\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y} + v}{2\eta}\right\}.
\end{aligned} \tag{A.12}$$

That is $\eta|\boldsymbol{X}_\gamma,\boldsymbol{y} \propto \mathcal{IG}(\frac{m+v}{2}, \frac{v+a^{-1}\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y}}{2})$. Hence (6) holds. Note that

$$p(\boldsymbol{\beta}_\gamma,\eta|\boldsymbol{X}_\gamma,\boldsymbol{y}) = p(\boldsymbol{\beta}_\gamma|\eta,\boldsymbol{X}_\gamma,\boldsymbol{y})p(\eta|\boldsymbol{X}_\gamma,\boldsymbol{y}). \tag{A.13}$$

Comparing (A.11), (A.12), and (A.13), we have

$$p(\boldsymbol{\beta}_\gamma|\eta,\boldsymbol{y},\boldsymbol{X}_\gamma) \propto a^{-\frac{n_\gamma}{2}}\eta^{-\frac{n_\gamma}{2}}|M_\gamma|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2a\eta}(\boldsymbol{\beta}_\gamma - H_\gamma)^{\mathrm{T}}M_\gamma^{-1}(\boldsymbol{\beta}_\gamma - H_\gamma)\right\}. \tag{A.14}$$

That is $\boldsymbol{\beta}_\gamma|\eta,\boldsymbol{y},\boldsymbol{X}_\gamma \sim \mathcal{N}(H_\gamma, a\eta M_\gamma)$. Therefore, (7) holds. After integrating out η in (A.12), we have

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\gamma}) &= \int_\eta \left\{\int_{\boldsymbol{\beta}_\gamma} p(\boldsymbol{y}|\boldsymbol{\beta}_\gamma,\eta,\boldsymbol{X}_\gamma)p(\boldsymbol{\beta}_\gamma|\eta)p(\eta)\mathrm{d}\boldsymbol{\beta}_\gamma\right\}\mathrm{d}\eta \\ &\propto a^{-\frac{m+n_\gamma}{2}}|\Sigma_\gamma|^{-\frac{1}{2}}|M|^{\frac{1}{2}}\int_\eta \eta^{-\frac{m+v+2}{2}} \\ &\quad\times \exp\left\{-\frac{a^{-1}\boldsymbol{y}^{\mathrm{T}}P_\gamma\boldsymbol{y}+v}{2\eta}\right\}\mathrm{d}\eta \\ &\propto a^{-\frac{m+n_\gamma}{2}}|\Sigma_\gamma|^{-\frac{1}{2}}|M|^{\frac{1}{2}}(v+a^{-1}\boldsymbol{y}P_\gamma\boldsymbol{y})^{-\frac{m+v}{2}}. \end{aligned} \tag{A.15}$$

We set $\Sigma_\gamma \triangleq (\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$, then $M_\gamma = (1+a)^{-1}(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}$, and then $|\Sigma_\gamma|^{-\frac{1}{2}}|M_\gamma|^{\frac{1}{2}} = (1+a)^{-\frac{n_\gamma}{2}}$. Hence, $p(\boldsymbol{y}|\boldsymbol{\gamma}) \propto a^{-\frac{m+n_\gamma}{2}}(1+a)^{-\frac{n_\gamma}{2}}S(\boldsymbol{\gamma},\boldsymbol{y})^{-\frac{m+v}{2}}$, i.e., (4) holds.

Appendix B. Fast computation

B.1. Computation of $p(\gamma_j|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j})$ in (8)

Because γ_j only takes 0 or 1, we can re-consider $p(\gamma_j = 1|\boldsymbol{y},\boldsymbol{X},i\neq j)$ and $p(\gamma_j = 0|\boldsymbol{y},\boldsymbol{X},i\neq j)$. Let $\boldsymbol{\gamma}^1 = (\gamma_1,\ldots,\gamma_{j-1},\gamma_j = 1,\ \gamma_{j+1},\ldots,\gamma_n)$ and $\boldsymbol{\gamma}^0 = (\gamma_1,\ldots,\gamma_{j-1},\gamma_j = 0,\ \gamma_{j+1},\ldots,\gamma_n)$. According to (8), we have

$$p(\gamma_j = 1|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j}) \propto a^{-\frac{m+n_{\gamma^1}}{2}}(1+a)^{-\frac{n_{\gamma^1}}{2}}S(\boldsymbol{\gamma}^1|\boldsymbol{y},\boldsymbol{X})^{-\frac{m+v}{2}}\pi_j,$$

$$p(\gamma_j = 0|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j}) \propto a^{-\frac{m+n_{\gamma^0}}{2}}(1+a)^{-\frac{n_{\gamma^0}}{2}}S(\boldsymbol{\gamma}^0|\boldsymbol{y},\boldsymbol{X})^{-\frac{m+v}{2}}(1-\pi_j).$$

Since $p(\gamma_j = 1|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j}) + p(\gamma_j = 0|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j}) = 1$, some straightforward computation yields

$$p(\gamma_j = 1|\boldsymbol{y},\boldsymbol{X},\gamma_{i\neq j}) \propto \frac{1}{1+h}, \tag{B.1}$$

$$\text{with } h = \frac{1-\pi_j}{\pi_j}(a+a^2)^{\frac{1}{2}}\left[\frac{S(\boldsymbol{\gamma}^1|\boldsymbol{y}\boldsymbol{X})}{S(\boldsymbol{\gamma}^0|\boldsymbol{y},\boldsymbol{X})}\right]^{\frac{m+v}{2}}. \tag{B.2}$$

If $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ before γ_j is generated, meaning we have obtained $S(\boldsymbol{\gamma}^0|\boldsymbol{y},\boldsymbol{X})$, then we only need to compute $S(\boldsymbol{\gamma}^1|\boldsymbol{y},\boldsymbol{X})$, and vice versa.

B.2. Fast computation of $S(\boldsymbol{\gamma},\boldsymbol{y})$ in (3)

The key to speed up the whole computation is to compute $S(\boldsymbol{\gamma},\boldsymbol{y})$ fast where a gene variable is added or removed from $\boldsymbol{\gamma}$. Denote

$$E(\boldsymbol{\gamma},\boldsymbol{y}) \triangleq \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{X}_\gamma)^{-1}\boldsymbol{X}_\gamma^{\mathrm{T}}\boldsymbol{y}. \tag{B.3}$$

The (B.3) can be computed using the fast QR decomposition, QR-delete and QR-insert algorithms when a variable is added or removed [27] (Ch. 10.1.1b). Now we want to estimate $S(\boldsymbol{\gamma},\boldsymbol{y})$ in (3). After straightforward computation, $S(\boldsymbol{\gamma},\boldsymbol{y})$ is given by

$$S(\boldsymbol{\gamma},\boldsymbol{y}) = v + \frac{a\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + E(\boldsymbol{\gamma},\boldsymbol{y})}{a(1+a)}. \tag{B.4}$$

Thus, after computing $E(\boldsymbol{\gamma},\boldsymbol{y})$ using QR decomposition, QR-delete or QR-insert algorithms, we then can obtain $S(\boldsymbol{\gamma},\boldsymbol{y})$.

References

[1] Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 1993;88:669–79.

[2] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

[3] Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 2003;19:563–70.

[4] Chen M-H, Ibrahim JG, Yiannoutsos C. Prior elicitation, variable selection, and Bayesian computation for logistic regression models. Journal of the Royal Statistical Society, Series B 1999;61:223–42.

[5] Chipman H, George EI, McCulloch R. The practical implementation of Bayesian model selection 2002. Avilable from: http://www-stat.wharton.upenn.edu/~dgeorge/Research_papers/ims.pdf.

[6] Cruickshanks KJ, Vadheim CM, Moss SE, Roth MP, Riley WJ, Maclaren NK, et al. Genetic marker associations with proliferative retinopathy in persons diagnosed with diabetes before 30 year of age. Diabetes 1992;41:879–85.

[7] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–6.

[8] Greene WH. Econometric analysis. Upper Saddle River, NJ: Prentice Hall; 1997.

[9] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

[10] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46:389–422.

[11] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene expression profiles in hereditary breast cancer. The New England Journal of Medicine 2001;344:539–48.

[12] Jornsten R, Yu B. Simultaneous gene clustering and subset selection for classification via MDL. Bioinformatics 2003;19:1100–9.

[13] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001;7:673–9.
[14] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–7.
[15] Lashkari DA, DeRisi JL, McCusker JH, Gentile AF, Hwang SY, Brown PO, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America 1997;94:13057–62.
[16] Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17:1131–42.
[17] Li W, Yang Y. How many genes are needed for a discriminant microarray data analysis?. In: Lin SM, Johnson KF, editors. Methods of microarray data analysis. Kluwer Academic; 2002. p. 137–50.
[18] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MY, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 1996;14:1675–80.
[19] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18:39–50.
[20] O'Brien SM, Dunson DB. Bayesian multivariate logistic regression 2004. Avialable from: http://citeseer.nj.nec.com/566151.html.
[21] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proceedings of the National Academy of Sciences of the United States of America 1994;91:5022–6.
[22] Pulkstenis E, Robinson TJ. Links goodness-of-fit tests for ordinal response regression models. Statistical Medicine 2004;23:999–1014.
[23] Qian G, Field C. Using MCMC for logistic regression model selection involving large number of candidate models. In: The Proceeding of the 4th International Conference on Monte Carlo and Quasi-Monte Carlo methods in Scientific Computing, Nov. 27–Dec. 1, Hong Kong. 2000.
[24] Robert C. Simulation of truncated normal variables. Statistics and Computing 1995;5:121–5.
[25] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–70.
[26] Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proceedings of the National Academy of Sciences of the United States of America 1996;93:10614–9.
[27] Seber GAF. Multivariate observations. New York: Wiley; 1984.
[28] Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Research 1996;6:639–45.
[29] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 1998;9:3273–970.
[30] Tabus I, Rissenan J, Astola J. Classification and feature gene selection using the normalized maximum likelihood for discrete regression. Signal Processing 2003;83:713–27.
[31] Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology 1997;15:1359–67.
[32] Zhou X, Wang X, Dougherty ER. Binarization of microarray data based on a mixture model. Molecular Cancer Therapeutics 2003;2:679–84.
[33] Zhou X, Wang X, Dougherty ER. Missing value estimation based on linear and nonlinear regression with Bayesian gene selection. Bioinformatics 2003;19:2302–7.
[34] Zhou X, Wang X, Dougherty ER. Gene prediction using multinomial probit regression with Bayesian variable selection. EURASIP Journal of Applied Signal Processing, special issue on Genomic Signal Processing 2004;3:115–24.
[35] Zhou X, Wang X, Dougherty ER. A Bayesian approach to nonlinear probit gene selection and classification. Journal of Franklin Institute, special issue on Genomics, Signal Processing and Statistics 2004;341:137–56.
[36] Zhou X, Wang X, Dougherty ER, Wong ST, Gene selection using logistic regressions based on AIC, BIC and MDL criteria, IEEE Transaction on Computational Biology and Bioinformatics 2004 [accepted to be published].