BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3423–3430
doi:10.1093/bioinformatics/bth419

Gene selection using a two-level hierarchical Bayesian model

Kyounghwa Bae and Bani K. Mallick[1,*]

[1]*Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*

Received on March 7, 2004; revised on July 9, 2004; accepted on July 10, 2004
Advance Access publication July 15, 2004

ABSTRACT

Summary: The fundamental problem of gene selection via cDNA data is to identify which genes are differentially expressed across different kinds of tissue samples (e.g. normal and cancer). cDNA data contain large number of variables (genes) and usually the sample size is relatively small so the selection process can be unstable. Therefore, models which incorporate sparsity in terms of variables (genes) are desirable for this kind of problem. This paper proposes a two-level hierarchical Bayesian model for variable selection which assumes a prior that favors sparseness. We adopt a Markov chain Monte Carlo (MCMC) based computation technique to simulate the parameters from the posteriors. The method is applied to leukemia data from a previous study and a published dataset on breast cancer.

Contact: bmallick@stat.tamu.edu

Supplementary information: http://stat.tamu.edu/people/faculty/bmallick.html

INTRODUCTION

Microarray experiments typically measure the expression levels of several thousands of genes simultaneously. In cDNA data, it is common to have a large number of genes and a relatively small sample size. By removing redundant variables (genes), it would be possible to highlight those genes that are most relevant for certain events (for instance, certain diseases or a certain type of tumor).

Several approaches for finding the differentially expressed genes have been proposed: the t-test (e.g. Devore and Peck, 1997), a regression modeling approach (Thomas *et al*., 2001), mixture model approach (Pan, 2002) and non-parametric methods (Troyanskaya *et al*., 2002). All of these are univariate gene selection methods and hence suffer from the fact that no correlations between the genes are considered in the selection procedure. Recently, Lee *et al*. (2003) developed a multivariate Bayesian model to perform variable selection. Their method made use of mixture prior distribution, which is very sensitive toward the choice of some hyper-parameters like the mixing probability π. In general, the algorithm is slow due to complicated mixing structure of the posterior distribution.

From a machine learning viewpoint, high dimensionality and sparsity of data points suggest the use of support vector machines (SVMs) (Campbell, 2002). Usually SVMs achieve low test error despite small sample sizes. Several papers have reported results on the application of SVMs for performing variable selection (Guyon *et al*., 2002; Weston *et al*., 2001). However, this method has a number of disadvantages, such as the absence of probabilistic output and the necessity of estimating a trade-off parameter in order to utilize Mercer kernel functions. An alternative approach is to exploit the Bayesian technique of automatic relevance determination (ARD). An ARD approach has been used previously for constructing a sparse classifier using the relevance vector machine (RVM) of Tipping (2000, 2001). Li *et al*. (2002) utilized ARD to perform variable selection rather than using generalization bounds from statistical learning theory. Their variable selection method has a performance similar to SVMs when applied to gene expression datasets from cDNA microarray data. The advantage of their approach is that variable sparsity is naturally incorporated into the algorithm—the optimal number of relevant variables is decided automatically. In contrast, for an SVM an additional variable selection procedure has to be added and a further criterion must be used to indicate when the best variable set has been found. In terms of practical application, Li *et al*. (2002) highlight the importance of a small number of influential genes. They use a zero-mean Gaussian prior with unknown variance for the unknown regression parameter β that favors sparseness in estimating β. This choice of prior for β shows very good performances (Williams, 1998; Williams and Barber, 1998) but the main disadvantage is that it does not control the structural complexity of the resulting functions. That is, if one of the components of β happens to be irrelevant, a Gaussian prior will not set it exactly to zero but instead to some small value (shrinkage rather than selection).

In this paper, we consider a multivariate Bayesian regression model and assign priors that favor sparseness in terms of number of variables (genes) used. We introduce the use of different priors to promote different degrees of sparseness using a unified two-level hierarchical Bayesian model. In our

*To whom correspondence should be addressed.

first model, we assign a zero-mean Gaussian prior to β with an independent prior distribution for the unknown variance of β. This model is related to ARD, although we perform full Bayesian analysis rather than marginal-likelihood maximization. We use a Laplace prior in our second model as it is known to promote sparseness (Williams, 1995), which is equivalent to the Lasso model. Our third model is based on the non-informative Jeffreys prior suggested by Figueiredo (2001). This particular prior does not contain any hyper-parameter by which we can implement variable selection automatically as well as strongly induce sparseness in the model. Importantly, the number of selected genes is decided automatically. Unlike other approaches, which are based on approximations, we will perform full Bayesian analysis exploiting simulation based on Markov chain Monte Carlo (MCMC) methodology (Gelfand and Smith, 1990; Gilks *et al*., 1996) to derive the estimates (as well as the uncertainty distributions) of the unknown parameters.

We apply our methods to a leukemia dataset from Golub *et al*. (1999) and also to a dataset from Hendenfalk *et al*. (2001). The idea is to identify a small number of genes having the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and therapeutics.

MODEL

Suppose that n independent binary random variables (e.g. normal and cancer), $Y_1, \ldots, Y_n$ are observed. $Y_i = 1$ indicates that sample i is cancer or one type of cancer (e.g. ALL, BRCA1) and $Y_i = 0$ indicates that sample i is normal or the other type of cancer (e.g. AML, BRCA2 and sporadic). For each sample, we measure gene expression levels for a set. Let X_{ij} denote the gene expression level of the j-th gene for the i-th sample and we form the data matrix $\mathbf{X}$ as

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

Define the binary regression model as $p_i = P(Y_i = 1) = \Phi(\mathbf{X_i}\beta)$, $i = 1, \ldots, n$, where β is the vector of unknown regression parameters, $\mathbf{X_i}$ is the i-th row vector of the matrix $\mathbf{X}$ and Φ is the standard normal cumulative density function linking the probability p_i with the linear structure $\mathbf{X_i}\beta$. This is known as probit model.

Albert and Chib (1993) introduce n independent latent variables $\mathbf{Z} = (Z_i, \ldots, Z_n)$ into the problem, where $Z_i \sim N(\mathbf{X_i}\beta, 1)$ and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \leq 0$. This approach connects the probit binary regression model for Y_i to a normal linear regression model for the latent variable Z_i.

We consider different priors for β in a two-level hierarchical Bayesian model. This model involves a zero-mean Gaussian prior for β with unknown variances. Then, we assign choices of priors for the variances assuming that they are independent. So, the prior distribution of β is

$$\beta|\mathbf{\Lambda} \sim N(\mathbf{0}, \mathbf{\Lambda}),$$

where $\mathbf{0} = (0, \ldots, 0)'$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and λ_i is the variance of β_i. We assign three different choices of prior distributions for Λ which develop three different models inducing different degrees of sparsity to select the number of genes used.

Prior distribution for the $\mathbf{\Lambda}$

For Model I, we assign a conjugate Inverse Gamma prior for each λ_i in $\mathbf{\Lambda}$ as $\mathrm{IG}(a/2, 2/b)$. Here a random variable X is said to follow Inverse Gamma distribution if $\mathrm{IG}\left(\frac{a}{2}, \frac{2}{b}\right) \sim \left(\frac{1}{X}\right)^{(a/2)+1} \exp\left(-\frac{b}{2X}\right)$. Note that we have two hyper-parameters, a and b, to be adjusted. Usually we adjust a and b in such a way that the variance of λ is very large. This model is equivalent to ARD model of Li *et al*. (2002). Assuming independence among λ_is, the prior distribution of $\mathbf{\Lambda}$ is given by

$$\mathbf{\Lambda} \sim \prod_{i=1}^{p} \mathrm{IG}\left(\frac{a}{2}, \frac{2}{b}\right).$$

In Model II, we assign a Laplace prior for β to promote sparseness (so that irrelevant parameters are set exactly to zero). We can express the Laplace prior distribution as a scale mixture of Normal priors, which is equivalent to a two-level hierarchical Bayesian model. The Laplace prior can be expressed as a zero-mean Gaussian prior with an independent exponentially distributed variance:

$$\pi(\beta_i|\gamma) = \int_0^\infty \pi(\beta_i|\lambda_i)\pi(\lambda_i|\gamma)d\lambda_i \sim \mathrm{Laplace}\left(0, \frac{1}{\sqrt{\gamma}}\right).$$

We assign an exponential distribution for the prior distribution of λ_i, which is equivalent to assigning a Laplace prior for β. Here, a random variable X is said to follow exponential distribution with parameter γ denoted as $\mathrm{expon}(\gamma)$ and expressed as $\mathrm{expon}(\gamma) = \frac{\gamma}{2}\exp\left(-\frac{\gamma x}{2}\right)$.

The prior distribution of $\mathbf{\Lambda}$ (again with the assumption of independence among λ_i) is given by

$$\mathbf{\Lambda} \sim \prod_{i=1}^{p} \mathrm{expon}(\gamma).$$

Here again we need to fix the hyper-parameter γ in such a way that the variance of λ is high. This is similar to the Lasso model but has added flexibility due to the choices of multiple λs as against one choice in the Lasso method.

In Model III, we attempt to avoid the problem of fixing the hyper-parameters by letting the prior distribution of $\mathbf{\Lambda}$ be a

non-informative Jeffreys prior as

$$\mathbf{\Lambda} \sim |\mathbf{I}(\mathbf{\Lambda})|^{1/2} = \prod_{i=1}^{p} \frac{1}{\lambda_i}.$$

As already shown in Figueiredo (2001) and in our experimental results, this prior strongly induces sparseness and yields good performance.

COMPUTATION

The posterior distribution is not available in explicit form so we use the MCMC method (Gilks *et al*., 1996), specifically Gibbs sampling (Gelfand and Smith, 1990) to simulate the parameters from the posterior distribution.

The full conditional distribution of $\mathbf{Z}$ has a truncated normal distribution. The random variables $Z_1, \ldots, Z_n$ are independent with

$$Z_i|\beta, Y_i = 1 \propto N(\mathbf{X_i}\beta, 1) \text{ truncated at the left by 0}$$

$$Z_i|\beta, Y_i = 0 \propto N(\mathbf{X_i}\beta, 1) \text{ truncated at the right by 0.}$$

We generate random numbers Z_i using Robert's (1995) optimal exponential accept-reject algorithm.

In the two-level hierarchical Bayesian model with zero-mean Gaussian priors and independently distributed variances for β, the full conditional distribution of β is as follows.

$$\pi(\beta|Z, Y, \mathbf{\Lambda}) \propto N(\Sigma \mathbf{X}' Z, \Sigma),$$

where, $\Sigma = (\mathbf{X}'\mathbf{X} + \mathbf{\Lambda}^{-1})^{-1}$. We have used the Woodbury–Sherman–Morrison matrix identity to reduce the dimension of the matrix, from p to n. This makes the computation much faster because we have cDNA data which has a high dimensionality corresponding to the small sample ($n << p$).

$$\Sigma = \mathbf{\Lambda} - \mathbf{\Lambda}\mathbf{X}'(\mathbf{X}\mathbf{\Lambda}\mathbf{X}' + \mathbf{I})^{-1}\mathbf{X}\mathbf{\Lambda}.$$

The full conditional distribution of $\mathbf{\Lambda}$ for the Inverse Gamma prior (Model I) is the following:

$$\pi(\mathbf{\Lambda}|Z, Y, \beta) \propto \prod_{i=1}^{p} \text{IG}\left(\frac{a+1}{2}, \frac{2}{b+\beta_i^2}\right).$$

The full conditional distribution of $\mathbf{\Lambda}$ for the exponential prior (Model II) is the following:

$$\pi(\mathbf{\Lambda}^{-1}|Z, Y, \beta) \propto \prod_{i=1}^{p} \text{InvGauss}\left(\frac{\sqrt{\gamma}}{\beta_i}, \gamma\right),$$

where InvGauss denotes the inverse Gaussian distribution. The inverse Gaussian distribution for a random variable X is expressed as

$$\text{InvGauss}(\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda}{2\mu^2}\frac{(x-\mu)^2}{X}\right), \quad X \geq 0.$$

We use the algorithm of Michael *et al*. (1976) to generate the random number from the inverse Gaussian distribution.

The full conditional distribution of $\mathbf{\Lambda}$ with the Jeffreys prior (Model III) is the following:

$$\pi(\mathbf{\Lambda}^{-1}|\mathbf{Z}, Y, \beta) \propto \prod_{i=1}^{p} G\left(\frac{1}{2}, \frac{2}{\beta_i^2}\right),$$

where G is the Gamma distribution.

In practice, many of the λ_i approach zero, implying those genes can be pruned from the model. During MCMC iteration we delete genes using the criterion $\lambda_i < 10^{-12}$ as in Li *et al*. (2002). Also we re-introduce a gene which has been eliminated if it has large enough variance (10 or more). However, we found little change in performance on varying this re-introduction bound.

Finally, we obtain the predictive classification of a new observation Y_{new}, conditioning on the gene expression level X using the Monte-Carlo estimate:

$$\widehat{P}(Y_{\text{new}} = 1|X) = \frac{1}{m}\sum_{t=1}^{m} p(Y_{\text{new}} = 1|X, \beta^t, Z^t, \Lambda^t), \quad (1)$$

where $\beta^t, Z^t,$ and Λ^t are the MCMC samples from the posterior distribution.

APPLICATION OF GENE SELECTION

Leukemia dataset

We apply our method to the Leukemia dataset that has been extensively studied by Golub *et al*. (1999). The authors gathered bone marrow or peripheral blood samples from 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). The data are split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. The gene expression levels for 7129 human genes are produced. Golub *et al*. (1999) investigated the use of a weighted voting scheme on the training samples and correctly classified 36 of the 38 training samples and also correctly classified 29 of the 34 test sample, failing to predict correctly on 5. Using Golub's training data, we identify the 500 most significant genes by using two sample t-test statistics. We start with the 500 genes out of 7129, which include all the significant genes identified by Lee *et al*. (2003) and Li *et al*. (2002). We run the MCMC sampler (in our case, Gibbs sampling with 50 000 iterations and 20 000 burn-in). The priors are as follows. We assume $E(\lambda_i) = 10$ and $\text{var}(\lambda_i) = 100$ a priori for Models I and II and fix the hyper-parameters in that way.

We obtain samples from the marginal posterior distribution and obtain the estimates for β_is and λ_is. We plot the absolute values of β_i in Figure 1. The sparseness of β_i has been seen significantly in Model III and we can also see that absolute values of β_i in Model I are usually bigger than those in Model II.

We select genes using the posterior variance of β while keeping in mind that variables with smaller variance will have

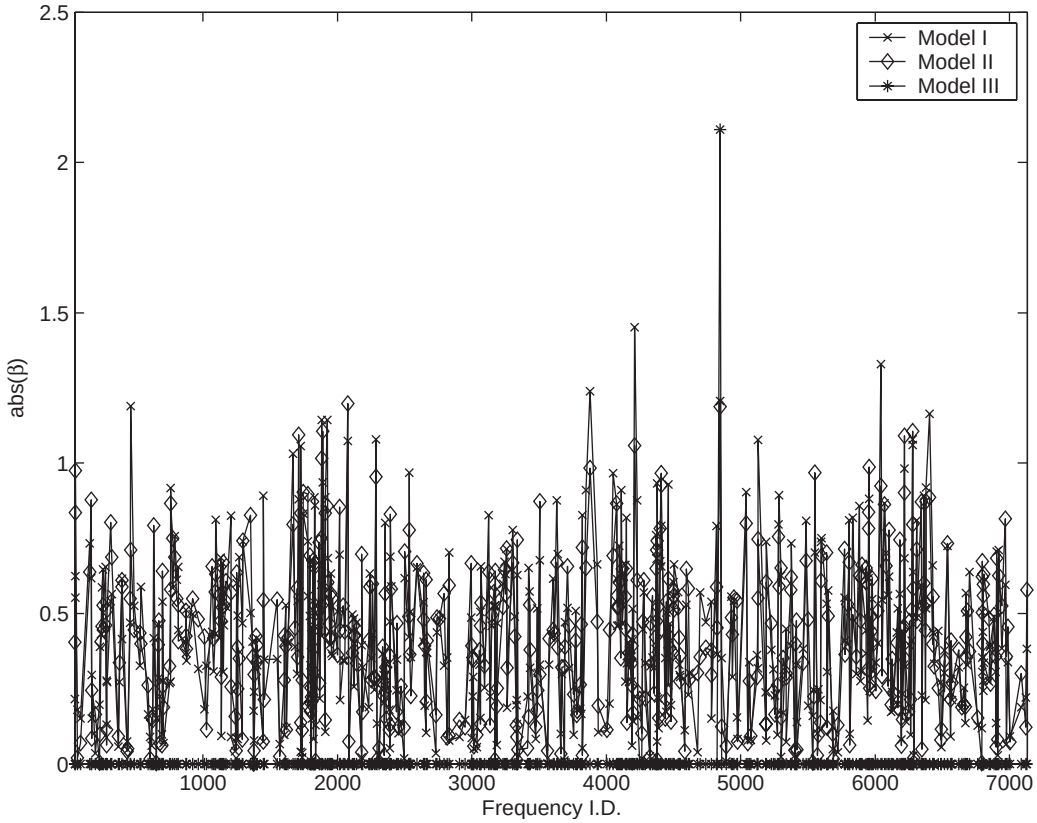

Fig. 1. Leukemia data: absolute value of β_i for three models. y-axis shows the absolute value of β_i and the x-axis shows the Frequency ID.

no effect and should be excluded from the model. Figure 2 shows the variance of β_i for each model. We can identify the genes having significantly larger variances than the others. Both Models I and II contain 20 genes which have significantly larger variance than the others. For Models I and II, we use these genes to perform prediction on the test data. The results are in Table 1 (we not only predict the correct classification but also the probability related to it). There are two misclassifications (5th and 6th) by both Models I and II. The top four selected genes are common to both of them: Zyxin (which encodes a LIM domain protein localized at focal contacts in adherent erythroleukemia cells; Maccalma *et al*., 1996); cell division control related protein (hCDCrel-1) mRNA (which is a partner gene of MLL in some Leukemias; Osaka *et al*., 1999); HoxA9 mRNA (which collaborates with other genes to produce highly aggressive acute leukemia disease; Thorsteinsdottr *et al*., 1999) and MacMarcks (this gene transcription is stimulated rapidly by tumor necrosis factor-α in human promyelocytic leukemia cells (Harlan *et al*., 1991).

In Model III, only Zyxin is selected due to significantly larger variance than others. Zyxin has the third and second rank according to Models I and II, respectively. The selected Zyxin is also one of leading genes in Lee *et al*. (2003) and Golub *et al*. (1999). Our prediction result based on Model III is in Table 1, which shows that there are three misclassifications using only one gene. Golub *et al*. (1999) used 50 genes to predict and had five misclassifications on test data. These results appear to improve predictions done by Golub *et al*. (1999), having fewer misclassifications while also using fewer genes.

In these small data and high-dimensional problems, several models can fit the data well, each using a distinct set of genes. To investigate the issue, Li *et al*. (2002) randomly partition the data into two disjoint subsets of equal size and fit the model on both sets. After training they match the common number of genes to both models. These data are heterogeneous as all 38 training samples were obtained from adult bone marrow; some test samples came from peripheral blood or pediatric patients. This type of random partitioning and resampling of the data will make the data more homogeneous (Smith *et al*., 2002). Following this idea, we make new training and test datasets by randomly splitting the 72 samples in half ($36 + 36$ samples). We perform 50 re-samplings and select the top 20 genes. The top 20 selected genes for all the three models were in common at least 24% of times in the re-sampling results. The top four genes for Models I and II were in common 50–70% of times. For Model III, we found Zyxin was in common 80% of times.

These data are not very homogeneous as observations were taken from different cells and to control the variability we

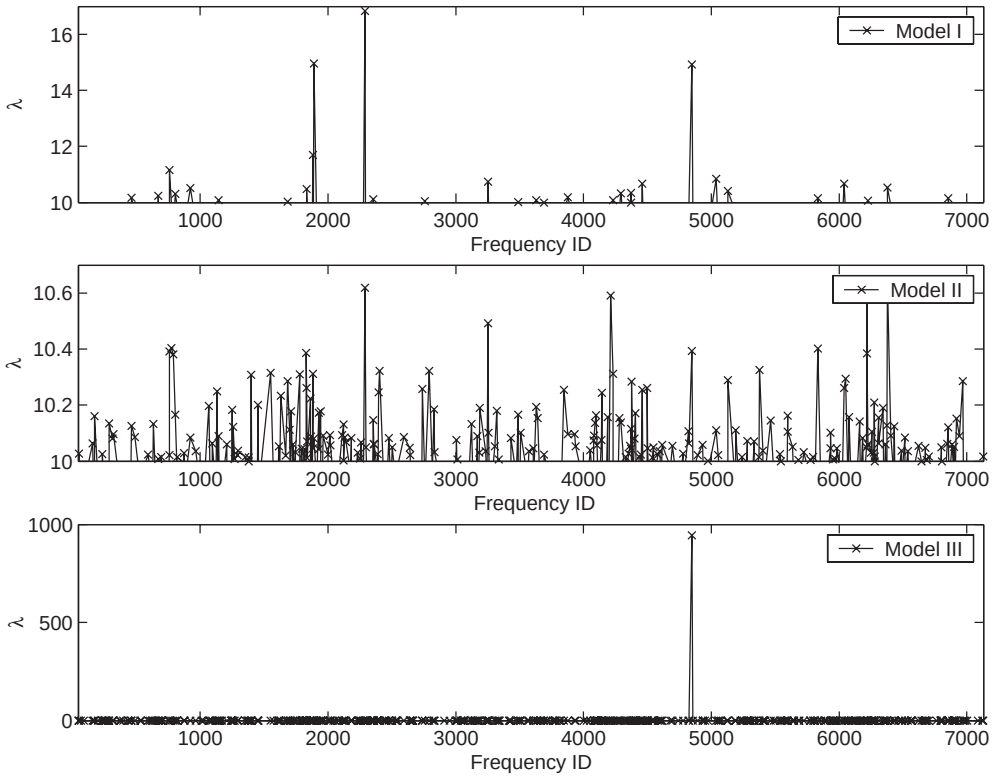

Fig. 2. Leukemia data: the variance of β_i for three models. y-axis shows the variance of β_i and the x-axis shows the Frequency ID.

Table 1. Leukemia data: the prediction of the test data

Y	Normal $P(Y\|X_{\text{test}})$	Laplace $P(Y\|X_{\text{test}})$	Jeffrey $P(Y\|X_{\text{test}})$	Y	Normal $P(Y\|X_{\text{test}})$	Laplace $P(Y\|X_{\text{test}})$	Jeffrey $P(Y\|X_{\text{test}})$
1	1	1	1	1	1	1	1
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	0	0	1	1	1	1	1
1	0	0	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	0	0	0	0
0	0	0	0	1	0.939	0.999	0
0	0	0	0	1	1	1	0
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
1	1	1	0	1	1	1	1

re-analyze on a subcategory of the data. For example, ALL cells can be either T-cells or B-cells and we apply this method to determine genes which are likely to be differentially expressed between ALL T-cells and ALL B-cells (Yeoh *et al*., 2002). In this manner, we controlled the heterogeneity of the sample type as much as possible by focusing on the B-cells and the T-cells experiments within the ALL group. This gives two reasonably homogeneous sample types, for which we still have many observations. We use 38 samples as training dataset and use the 9 samples as the test dataset, which is the same procedure as that of Grant *et al*. (2002). The top four selected genes in Model I are ID 6855, 5542, 1882 and 1962.

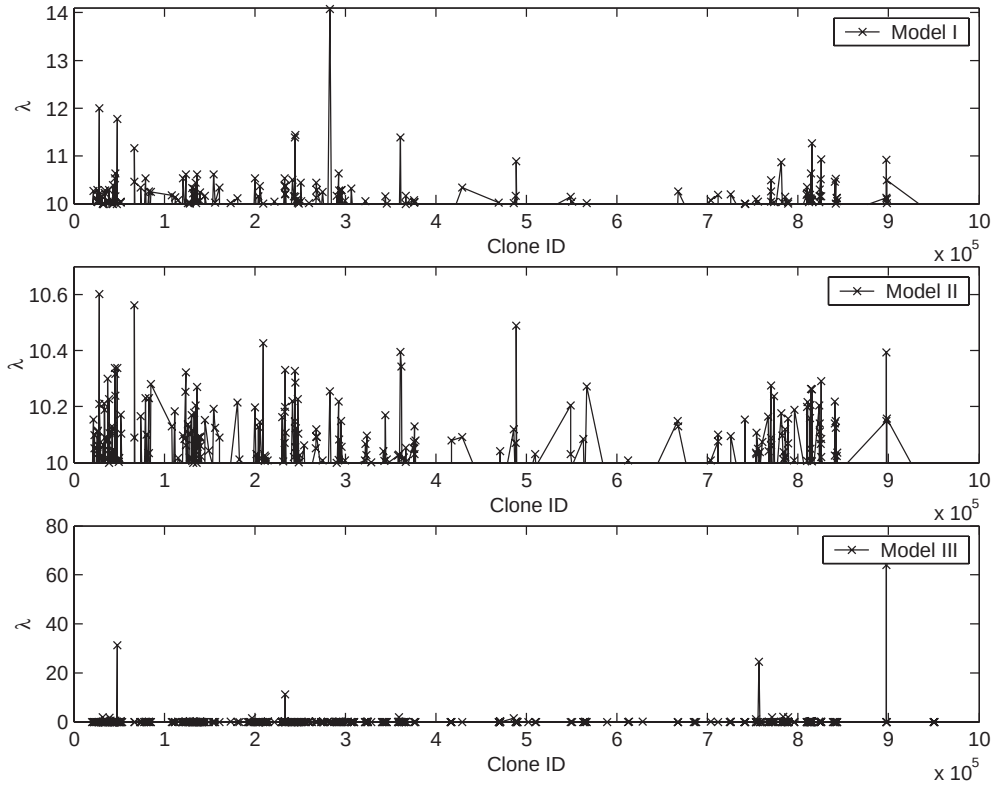

Fig. 3. Breast cancer data: the variance of β_i for three models. y-axis shows the variance of β_i and the x-axis shows the Clone ID.

The top selected gene, ID 6855 is TCF-3 transcription factors (E2A immunoglobulin enhancer binding factors E12/E4. Heterodimers between TCF3 and tissue-specific basic helix–loop–helix (bHLH) proteins play major roles in determining tissue-specific cell fate during embryogenesis, such as muscle or early B-cell differentiation (Kamps *et al*., 1990). They are involved in a form of pre-B-cell acute lymphoblastic leukemia (B-ALL) through a chromosomal translocation which involves PBX1 and TCF3. T-cell Antigen CD7 precursor (ID 5542), is one of two common selected genes by Dudoit *et al*. (2000) and Grant *et al*. (2002). The top four genes selected by Model II are ID 6967, 1882, 6855 and 4342. The top selected gene, ID 6967 is SELL Leukocyte adhesion protein beta subunit (ITGB2). The ITGB2 protein product is the integrin-β chain β2. Integrins are integral cell-surface proteins composed of an α-chain and a β-chain. A given chain may combine with multiple partners resulting in different integrins. For example, β2 combines with the αL chain to form the integrin LFA-1, and combines with the αM chain to form the integrin Mac-1. Integrins are known to participate in cell adhesion as well as cell-surface-mediated signaling. The gene TCF7 transcription factor 7 (T-cell specific) (clone ID 4342) is selected by the Model III. This gene is one of the two common selected genes by Dudoit *et al*. (2000) and Grant *et al*. (2002). The Tcf7 gene encodes a transcription factor that is a member of the high mobility group protein family. Expression of Tcf7 is specific to T-cells, and the gene product was originally designated TCF-1, as a T-cell-specific transcription factor. A closely related factor, LEF-1 (lymphocyte transcription factor) is expressed in both T- and B-cell lineages. Both TCF-1 and LEF-1 arise from the same gene, Tcf7, by alternative splicing and the use of dual promoters (Kingsmore *et al*., 1995). The detailed results of the analyses are in the Supplementary website.

Hereditary breast cancer dataset

As a second study we also apply our method to a breast cancer dataset (Hendenfalk *et al*., 2001) from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2 or from patients not expected to carry a hereditary predisposing mutation. Pathological and genetic differences appear to imply different but overlapping functions for BRCA1 and BRCA2. Hendenfalk *et al*. (2001) examined 22 breast tumor samples from 21 breast cancer patients. A total of 15 women had hereditary cancer, 7 having tumors with BRCA1 and 8 having tumors with BRCA2. A total of 3226 genes were used for each breast tumor sample. We use our method to classify BRCA1 versus the others (BRCA2 and sporadic).

We used initial two-sample t-test statistics to identify the 500 most significant genes and run the MCMC sampler as in the previous example. We chose the same hyper-parameters as in the previous example. The variances of 500 genes are

plotted in Figure 3. Some of the leading genes selected by these approaches appear among the 10 strongest genes in the list in Kim *et al*. (2002) and Lee *et al*. (2003). For both Models I and II, we selected 25 genes which have significantly larger variances than others. The leading gene (by both the approaches) is keartin8 (KRT8), a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry, and cytokeratin8 abundance has been shown to correlate well with node-positive disease (Brotherick *et al*., 1998). Another top selected gene is tumor-associated antigen L-6 (TM4SF1), a member of a family of integral membrane proteins, several of which are also overexpressed in tumors (Marken *et al*., 1992). Antigen L-6 is frequently overexpressed in carcinomas, and antibody binding to L-6 on tumors in nude mouse models inhibits their outgrowth (Hellstrom *et al*., 1986).

In Model III, only four genes appeared to be the selected ones with significantly high variance. Keratin 8 and TM4SF1 are the top leading genes in Kim *et al*. (2002) and Lee *et al*. (2003) as well as in our previous two models. The other two genes are TOB1 and CTP syntheses which also appeared in all the previously mentioned lists. The gene TOB1 interacts with the oncogene receptor ERBB2, and is found to be more highly expressed in BRCA2 and sporadic cancers, which are likewise more likely to harbor ERBB3 gene amplifications. TOB1 has an anti-proliferative activity that is apparently antagonized by ERBB2 (Matsuda *et al*., 1996).

We have checked the sensitivity (stability) of our analysis by adding a Gaussian noise to the expression values as in Lee *et al*. (2003). We re-analyzed the data contaminated by Gaussian noise to obtain the newly selected genes and have reproduced the results in Supplementary website. The table shows that the analysis is quite stable, as it is selecting almost similar genes with a different noise level over the expression values.

We also check the model adequacy by leave-one-out cross-validation (CV). We exclude a single data point and predict the $P(Y = 1|X)$ for that data point using Equation (1). For Models I and II, we use the 25 selected genes and for Model III the 3 selected genes to perform the CV. We compare the result of this CV with the observed response. Our CV results are reported in the Supplementary website. There are no misclassifications by Models I and II and two misclassifications (17th and 18th sample) by Model III. We compare our CV results with other popular classification algorithms in Table 2. All other methods have used 51 genes. It is clear from the results that our methods improve the classification accuracy, having fewer misclassifications while also using fewer genes.

Table 2. Feature selection for the breast cancer data: 51 features used by Hendenfalk *et al*. (2001)

	Model	Cross-validation error[a]
1	Feed-forward neural networks (three hidden neurons, one hidden layer)	1.5 (average error)
2	Gaussian kernel	1
3	Epanechnikov kernel	1
4	Moving Window kernel	2
5	Probabilistic neural network ($r = 0.01$)	3
6	kNN ($k = 1$)	4
7	SVM linear	4
8	Perceptron	5
9	SVM nonlinear	6

[a]Number of misclassified samples.

DISCUSSION

We propose two-level hierarchical Bayesian models for variable selection which assume priors favoring sparseness in parameters. We employ latent variables to specialize the model to a regression model. We use simulation-based MCMC methodology to derive the estimates of the unknown parameters. All the three models provide good performance in terms of gene selection but Model III based on Jeffreys prior is preferable as there is no need to specify the hyperparameter or any type of threshold values. Simpler methods based on scores such as Fisher score or correlation coefficients can be used for gene selection but usually they would select a much larger number of genes and due to small sample size the method may produce instability in the classification process. Due to Bayesian setup, we have a coherent way to predict (assign) new samples to particular categories. Rather than hard rules (in or out) of assignment, we can evaluate the probability (chance) that the new sample will be in one of the categories which is more helpful for decision making. Also use of smaller number of important genes simplifies the experimental procedure.

Our gene selection method is based on the posterior mean of λ. We use informal, exploratory plots to find the genes with significantly large value of λ. A formal choice of cut-off value to select significant λ based on posterior or predictive criteria will be a topic of future research.

All through our analysis, we assume data are independent and consider only binary classifiers. Future research will consider the gene with interaction situations and extend the analysis to multi-category models.

ACKNOWLEDGEMENTS

The research was partially supported by the National Science Foundation grant DMS-020321, National Cancer Institute Grant CA-57030.

REFERENCES

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

Brotherick,I., Robson,C.N., Browell,D.A., Shenfine,J., White,M.D., Cunliffe,W.J., Shenton,B.K., Egan,M., Webb,L.A., Lunt,L.G., Young,J.R. and Higgs,M.J. (1998) Cytokeratin expression in

breast cancer: phenotypic changes associated with disease progression. *Cytometry*, **32**, 301–308.

Campbell,C. (2002) Kernel methods: a survey of current techniques. *Neurocomputing*, **48**, 63–84.

Devore,J. and Peck,R. (1997) *Statistics: The Exploration and Analysis of Data*, 3rd edn. Duxbury Press, Pacific Grove, CA.

Dudoit,S., Yang,Y.H., Callow,M. and Speed,T. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *UC Berkley Technical Report No. 578*, University of California, Berkley, CA.

Figueiredo,M.A.T. (2001) Adaptive sparseness using Jeffreys prior. *Proc. Adv. Neural Inform. Process Syst.*, 14, In Dietteriech,T., Becker,S. and Gharamani,Z. (eds), 697–704.

Gelfand,A. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **88**, 881–889.

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

Golub,T.R., Slonim,D., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J., Coller,H., Loh,M., Downing,J., Caligiuri,M., Bloomfield,C. and Lender,E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grant,G., Manduchi,E. and Stoeckert,C. (2002) Using non-parametric methods in the context of multiple testing to identify differentially expressed genes. *Conference on Critical Assessment of Microarray Data Analysis (CAMDA'00)*. In Lin,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer Academic Publishers, Boston, MA, pp. 37–55.

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

Harlan,D.M., Graff,J.M., Stumpo,D.J. and Blackshear,P.J. (1991) The human myristoylated alanine-rich C kinae subtrate 9MARKCKS gene (MACS). Analysis of its gene product, promoter, and chromosal localization. *J. Biol. Chem.*, **266**, 14399–14405.

Hellstrom,I., Beaumier,P.L. and Hellstrom,K.E. (1986) Antitumor effects of 16, an IgG2a antibody that reacts with most human carcinomas. *Proc. Natl Acad. Sci., USA*, **83**, 7059–7063.

Hendenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P. Gusterson,B., Esteller,M., Kallioniemi,O.P., Wilfond,B., Borg,A. and Trent,J. (2001) Gene expression profiles in hereditary breast cancer. *N. Eng. J. Med.*, **344**, 539–548.

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

Kim,S., Dougherty,E.R., Barrera,J., Chen,Y., Bitter,M. and Trent,J.M. (2002) Strong feature sets from small sapmles C. *J. Comput. Biol.*, **7**, 673–679.

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

Lee,K.E., Sha,N., Dougherty,E.R., Vanucci,M. and Mallick,B.K. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

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

Maccalma,T., Otte,J., Hensler,M.E., Grzeschik,K.H. and Beckerle,M.C. (1996) Molecular characterization of human zyxin. *J. Biol. Chem.*, **271**, 31470–31478.

Marken,J.S., Schieven,G.L., Hellstrom,I., Hellstrom,K.E. and Aruffo,A. (1992) Cloning and expression of the tumor associated antigen. *Proc. Natl Acad. Sci., USA*, **89**, 3503–3507.

Matsuda,S., Kawamura-Tsuzuku,J., Ohsugi,M., Yoshida,M., Emi,M., Nakamura,Y., Onda,M., Yoshida,Y., Nishiyama,A. and Yamamoto,T. (1996) Tob, a novel protein that interacts with P185ERBB2, is associated with antiproliferative activity. *Oncogene*, **12**, 705–713.

Michael,J.R., Schucany,W.R. and Haas,R.W. (1976) Generating random variates using transformations with multiple roots. *Am. Stat.*, **30**, 88–90.

Osaka,M., Rowley,J.D. and Zeleznik-Le,N.J. (1999) MSF, a fusion partner gene of MLL, in a theraphy-related acute myeloid leukemia with at(11:17). *Proc. Natl Acad. Sci., USA*, **96**, 6428–6433.

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

Robert,C. (1995) Simulation of truncated normal variables. *Statist. Comput.*, **5**, 121–125.

Smith,A., Satagopan,J., Gonen,M. and Begg,C. (2002) Exploring class prediction for leukemia gene expression data. *Conference on Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'00)*, December 18–19, Duke University, NC.

Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Thorsteinsdottir,U., Krosl,J., Hoang,T., and Sauvageau,G. (1999) The oncoprotein E2A-pbx collaborates with IIoax2 to acutely transform primary bone mallow cells. *Mol. Cell Biol.*, **19**, 6355–6366.

Tipping,M.E. (2000) The relevance vector machine. *Adv. Neural Inform. Process. Syst.*, **12**, 652–658.

Tipping,M.E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Res.*, **1**, 211–244.

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

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

Williams,C. (1998) Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In *Learning and Inference in Graphical Models*, Jordan,M.I. (ed.), Kluwer.

Williams,P. (1995) Bayesian regularization and pruning using a Laplace prior. *Neural Comput.*, **7**, 117–143.

Williams,C. and Barber,D. (1998) Bayesian classification with Gaussian priors. *IEEE Trans. Pattern Anal. Machine Intell.*, **20**, 1342–1351.

Yeoh,E., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.