BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 7 2005, pages 1055–1061
doi:10.1093/bioinformatics/bti092

Gene expression

A novel means of using gene clusters in a two-step empirical Bayes method for predicting classes of samples

Yuan Ji[1,*], Kam-Wah Tsui[2] and KyungMann Kim[3]

[1]Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA, [2]Department of Statistics, The University of Wisconsin-Madison, Madison, WI 53706, USA and [3]Department of Biostatistics and Medical Informatics, The University of Wisconsin-Madison, Madison, WI 53792, USA

Received on August 18, 2004; revised on September 22, 2004; accepted on October 7, 2004
Advance Access publication October 28, 2004

ABSTRACT

Motivation: The classification of samples using gene expression profiles is an important application in areas such as cancer research and environmental health studies. However, the classification is usually based on a small number of samples, and each sample is a long vector of thousands of gene expression levels. An important issue in parametric modeling for so many gene expression levels is the control of the number of nuisance parameters in the model. Large models often lead to intensive or even intractable computation, while small models may be inadequate for complex data.

Methodology: We propose a two-step empirical Bayes classification method as a solution to this issue. At the first step, we use the model-based cluster algorithm with a non-traditional purpose of assigning gene expression levels to form abundance groups. At the second step, by assuming the same variance for all the genes in the same group, we substantially reduce the number of nuisance parameters in our statistical model.

Results: The proposed model is more parsimonious, which leads to efficient computation under an empirical Bayes estimation procedure. We consider two real examples and simulate data using our method. Desired low classification error rates are obtained even when a large number of genes are pre-selected for class prediction.

Availability: Supplemental materials including technical details are available at "http://odin.mdacc.tmc.edu/~yuanj/papers/sup.pdf". An R program for computation is available upon request by email to Yuan Ji (yuanji@mdanderson.org)

Contact: yuanji@mdanderson.org

1 INTRODUCTION AND THE PRELIMINARY STEP

1.1 Introduction

The classification of samples using gene expression profiles has received much attention. Here, samples denote target objects to be classified, and 'control' denotes the reference condition, e.g. placebo, healthy cell or standard treatment. In cancer research, classifying cancer types is a crucial part of forming the diagnosis, but standard classification methods rely on clinical variables that provide limited and unreliable information. Using the gene expression levels to classify cancer types could lead to finer and more reliable classification results and thereby improve treatments for patients. In toxicology and environmental health studies, an initial screening of toxic chemicals is essential in order to reduce total cost of the analysis. Success in classifying toxic chemicals will greatly accelerate the entire process of screening and testing.

Golub *et al*. (1999) and Thomas *et al*. (2001) conducted some early work in this area. The former proposed a voting method for a binary classification problem from a leukemia experiment. The latter introduced a Bayesian multinomial–Dirichlet classification model based on discretized gene expression levels. The problem with the approach by Thomas *et al*. (2001) is that the discretization of continuous gene expression levels results in a loss of information. As more applications appear in the literature, the focus is being directed to the construction of more accurate and efficient classification methods for high-dimensional microarray data. Dudoit *et al*. (2002) provided a comprehensive review of several discriminant methods, such as the classification tree and the linear and quadratic discriminant analysis; however, they did not examine a Bayesian classification method.

The Bayesian classification method is known for its flexibility and accuracy. Keller *et al*. (2000) and West *et al*. (2001) attempted to apply the traditional Bayesian classification method to microarray data. Usually the data contain only a small number of samples (target objects), while a long vector of thousands of gene expression levels, an expression profile, is recorded for each sample. The Bayes rule is trained using all the profiles for the samples with known classes, or the training set. Because of the high dimensionality of these profiles, it is not clear how the traditional Bayesian classification method should be implemented, especially in controlling the number of nuisance parameters in statistical models. For example, if one assigns a distinct variance for each gene expression level, the model will contain thousands of unknown parameters and thus will often become computationally intractable. On the contrary, it is unrealistic to assume that all the gene expression levels share the same variance due to the complex nature of gene expression data. In fact, genes with larger expression levels tend to have larger variabilities in their expression.

We propose a two-step empirical Bayes classification method to address the high dimensionality issue mentioned above. The idea is to reduce the number of nuisance parameters by exploiting the underlying homogeneity in the gene expression profiles. At the preliminary

*To whom correspondence should be addressed.

Table 1. The group structure of microarray gene expression data

Sample	Class label	Group 1			Group 2				Group L		
		1	$\cdots$	n_1	n_1+1	$\cdots$	n_2	$\cdots$	$n_{L-1}+1$	$\cdots$	n_L
1	C_1	x_{11}	$\cdots$	$x_{n_1,1}$	$x_{n_1+1,1}$	$\cdots$	$x_{n_2,1}$	$\cdots$	$x_{n_{L-1}+1,1}$	$\cdots$	$x_{n_L,1}$
.	.	.		.	.		.	.	.		.
.	.	.		.	.		.	.	.		.
.	.	.		.	.		.	.	.		.
m	C_m	x_{1m}	$\cdots$	$x_{n_1,m}$	$x_{n_1+1,m}$	$\cdots$	$x_{n_2,m}$	$\cdots$	$x_{n_{L-1}+1,m}$	$\cdots$	$x_{n_L,m}$

Each row is an expression profile for the sample at that row.

step of our method, we assign gene expression levels under the control conditions to abundance groups, which is realized using the model-based cluster algorithm. Here, the term 'abundance group' refers to the group of gene expression levels coming from the same underlying probability distribution under control conditions. This application of a gene cluster algorithm is quite different from the traditional aim of clustering, where genes are clustered to explore the functional or structural relationship among them. We discuss the details of this step in the next section. At the second step, we integrate the grouping information of genes into a Bayesian model for expression levels under both sample conditions and control conditions. Specifically, for each group of gene expression levels, we use a distinct variance parameter. Since the number of groups is usually substantially smaller than the number of genes, our model is much more parsimonious than the full model with one variance per gene. On the contrary, it is more complex than the naive model with only one variance for all the genes. Moreover, the variance estimates are based on the information from all the genes in the same group. Hence, we are borrowing strength across genes.

We use the Bayes rule to classify samples and estimate parameters using an empirical Bayes procedure. For convenience, we call our method the 'EBC' method to denote empirical Bayes classification.

1.2 The preliminary step

Cluster analysis has been used to reveal structural or functional patterns of genes. For example, Eisen *et al*. (1998) proposed the hierarchical clustering and Tamayo *et al*. (1999) introduced the self-organizing maps. Both procedures are based on non-parametric methods. Model-based methods have been investigated by Fraley and Raftery (1999), McLachlan and Peel (2000) and Yeung *et al*. (2001).

We choose the model-based cluster algorithm by Fraley and Raftery (1999) for our preliminary step with a very different purpose. The model-based cluster algorithm groups gene expression levels according to their underlying probability distributions. As a result, gene expression levels within the same cluster are supposed to have the same mean and variance expression levels. It is this feature that we utilize to reduce the number of parameters in our later statistical models. In what follows, the terms 'cluster' and 'abundance group' are interchangeable.

Details of the model-based cluster algorithm and how to implement it for microarray gene expression data are provided in our supplemental materials. Note that we only use the gene expression levels under control conditions to form the abundance groups. The reason is that different sample conditions will result in the alteration of gene expression levels in different ways. Gene expression levels under control conditions are most closely related to the natural expression behaviors. When gene profiles under control conditions are not available, we recommend using gene profiles under sample conditions that belong to the same class to implement the model-based cluster algorithm. For example, if the data contain 30 samples to be classified to five classes (with possibly more than one sample per class), we may only use the gene profiles under the samples that belong to the same class for the model-based cluster algorithm.

After applying the model-based cluster algorithm, each gene has a group label, which forms a special data structure, as shown in Table 1. We note that there are two kinds of labels: the group labels for genes obtained from the clustering procedure, and the class labels for the samples obtained from external knowledge, such as different classes of cancer. Each row in Table 1 contains a gene profile for a sample; several samples may belong to the same class. We denote the class of each sample t by C_t. Here, the samples could be cancer patients with different types of cancer and hence each t represents a patient and C_t is the type of cancer the patient t has. Quantity K is the total number of classes under consideration—the number of types of cancers, for example. According to the model-based cluster algorithm, the n genes are grouped into L groups; the original genes $\{1,\ldots,n\}$ are rearranged and relabeled as $\{(1,\ldots,n_1),(n_1+1,\ldots,n_2),\ldots,(n_{L-1}+1,\ldots,n_L)\}$ with genes $(n_{l-1}+1,\ldots,n_l)$ in group l, $l=1,\ldots,L$, where $n_0=0$ and $n_L=n$. The number of clusters L is determined by the Bayesian information criterion (BIC). At last, we let $Z_g=l$ represent that gene g belongs to group l.

We introduce the statistical model in Section 2. In Section 3, we evaluate the performance of the proposed method with two examples using real data. We describe simulation studies that further explore the properties of our method in Section 4. Finally, we summarize our findings and conclusions in Section 5.

2 EMPIRICAL BAYES CLASSIFICATION

Our model is built upon the structure described in Table 1. Let X_{gt} denote the expression level of gene g measured under sample t. Let X_{gc} denote the expression level of gene g under the corresponding control conditions. Index c is sample-specific and should actually be written as c_t. For simplicity, we will suppress the subindex t hereafter. In cDNA arrays, X_{gt} and X_{gc} are the fluorescence intensities for each gene. In single-channel microarrays, such as oligonucleotide arrays, X_{gt} and X_{gc} are the final quantifications under the sample conditions and the control conditions, respectively. For gene $g=1,\ldots,n_L$, group $l=1,\ldots,L$, sample $t=1,\ldots,m$, and

class $i = 1, \ldots, K$, we propose the following distributions:

$$h(X_{gt})|C_t = i, Z_g = l \overset{\text{indep.}}{\sim} N(\mu_l + \Delta_{gi}, \sigma_l^2), \quad (1)$$

$$h(X_{gc})|Z_g = l \overset{\text{indep.}}{\sim} N(\mu_l, \sigma_l^2), \quad (2)$$

where μ_l and σ_l^2 are the mean and variance expression levels of genes in group l, and Δ_{gi} is the differential effect of class i on gene g. The function $h(\cdot)$ represents a certain transformation (e.g. Box–Cox or log) after which the transformed gene expression levels satisfy the normality distribution assumption.

According to the model given by Equations (1) and (2), the gene expression levels under the control X_{gc} follow the same distribution if they are in the same group, which is consistent with the model-based cluster algorithm. When genes are measured under sample condition t, their mean expression levels are changed by adding a new sample effect Δ_{gi} to μ_l, the mean expression level under the control conditions. The key parameters are the Δ_{gi} because they determine the pattern in the change of gene expression levels when genes are affected by different classes. This pattern will then allow us to train the classification rule that predicts the classes of new observations.

For simplicity and without loss of generality, let

$$h(\cdot) = \log(\cdot),$$

i.e. the gene expression levels X_{gt} and X_{gc} follow a log-normal distribution, which has been often used in the literature. Denote the difference

$$D_{gt} = \log X_{gt} - \log X_{gc}.$$

Then,

$$D_{gt}|C_t = i, Z_g = l \sim N(\Delta_{gi}, 2\sigma_l^2). \quad (3)$$

The quantity D_{gt} is the difference in gene expression levels on the log scale between the sample and control conditions.

Given a set of training samples $t = 1, \ldots, m$ as shown in Table 1, we compute the quantity D_{gt} for every gene g and sample t. For a sample t, the set of quantities $\{D_{gt}; g = 1, \ldots, n_L\}$ becomes an expression profile for that sample. To predict the class of a new sample $m + 1$ using its corresponding gene expression profiles $\mathbf{D_{m+1}} = \{D_{g,m+1}; g = 1, \ldots, n\}$, we follow the Bayes rule, which assigns the new sample to the class j^* where

$$j^* = \arg\max_j P(C_{m+1} = j|\{D_{gt}\}, \mathbf{D_{m+1}}), \quad j = 1, \ldots, K.$$

The unknown parameters in (3) are the differential effect Δ_{gi} and variance σ_l^2, the prior distributions of which are given by

$$\Delta_{gi}|\sigma_l^2 \overset{\text{indep.}}{\sim} N(\Delta_{gi0}, 2a^{-1}\sigma_l^2) \quad (4)$$

and

$$\sigma_l^{-2} \overset{\text{indep.}}{\sim} \text{Gamma}(\alpha_l, \beta_l), \quad (5)$$

where Gamma(,) denotes the gamma distribution with a density function $f(y|\alpha, \beta) = y^{\alpha-1}e^{-y/\beta}/\Gamma(\alpha)\beta^\alpha$, $y > 0$. The scaling parameter a in Equation (4) is a positive nuisance parameter controlling the vagueness of the prior for Δ_{gi}. We estimate the hyper-parameters Δ_{gi0}, α_l and β_l using the marginal method of moments (MMOM), the details of which are given in the supplemental material.

After some algebraic manipulation and using Bayes' theorem, the posterior classification probability can be shown to have the form

$$P(C_{m+1} = j|\{D_{gt}\} = \prod_{l=1}^{L}\left[\left(\frac{1}{\xi_j}\right)^{\frac{n_l - n_{l-1}}{2}} \times \left(\frac{1}{\eta_{jl}}\right)^{\zeta_l}\right] \quad (6)$$

for $j = 1, \ldots, K$, up to a constant, where

$$\xi_j = 1 + \frac{1}{a + k_j},$$

$$\eta_{jl} = \frac{\sum_{g\in l}(D_{g,m+1} - \hat{\Delta}_{gj0})^2}{2\xi_j} + \frac{1}{2}\sum_{g\in l}\sum_{i=1}^{K}\sum_{t\in i}(D_{gt} - \hat{\Delta}_{gi0})^2 + \frac{1}{\hat{\beta}_l},$$

$$\zeta_l = \tfrac{1}{2}(m+1)(n_l - n_{l-1}) + \hat{\alpha}_l,$$

and

$$\hat{\Delta}_{gi0} = \frac{\sum_{t\in i} D_{gt}}{k_i}.$$

The derivation of this equation and the values of the parameters in the equation are given in the supplemental materials.

The probability given in (6) is the classification probability to be calculated for each class j. The Bayes rule assigns the class j^* that maximizes the right-hand side of Equation (6).

3 EXAMPLES

Throughout the examples and simulation studies, we estimate the number of gene abundance groups using the BIC.

A general caution when applying any classification method is to account for the selection bias (Ambroise and McLachlan, 2002). For microarray gene expression data, the bias occurs when genes are pre-selected using both the training set and the testing set, while the classification error rate is based on the results from the testing set only. In our examples to be presented, we re-emphasize the issue of selection bias so that all researchers, including statisticians, clinicians and biologists, are aware of this commonly made mistake.

3.1 Classification of toxicants with cDNA microarrays

We analyze a challenging data set containing 24 treatments but belonging to five classes, with one class containing only two treatments. The 24 treatments are combinations of different chemicals with multiple time courses, which are the samples to be classified. The expression levels of about 1242 genes are measured under the 24 treatment conditions and corresponding 24 control conditions. (See Thomas *et al.*, 2001 for detailed descriptions of the experiment.)

Our method is applicable using any number of genes. However, to demonstrate the performance of the method when different sets of genes are used, we pre-select genes using a simple score d, described as follows: for each gene, we first compute the average expression levels across all the samples in the same class; we call them class averages. We obtain five class averages per gene for this data set since it has five classes. The d-score is the average absolute difference of all distinct pairs of these class averages. The d-score is a crude

Table 2. Results for classifying 24 treatments in the cDNA microarray data

	Number of misclassifications							
n	30	40	100	200	300	500	1000	1242
Unbiased (rate)	5 (21%)	2 (8%)	2 (8%)	2 (8%)	3 (12%)	3 (12%)	3 (12%)	3 (12%)
Biased (rate)	5 (21%)	1 (4%)	1 (4%)	1 (4%)	1 (4%)	0 (0%)	2 (8%)	3 (12%)

The number of misclassifications is obtained using the LOOCV procedure.

Table 3. Summary of the two-class problem for the leukemia data

	Number of misclassifications							
n	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
Two-class C.V. (rate)	1 (3%)	2 (5%)	1 (3%)	1 (3%)	0 (0%)	2 (5%)	1 (3%)	1 (3%)
Two-class T.S. (rate)	5 (15%)	3 (9%)	3 (9%)	2 (6%)	1 (3%)	3 (9%)	2 (6%)	1 (3%)

'C.V.' represents the number of misclassifications (out of 38 samples) for the training set using the LOOCV procedure. 'T.S.' represents the number of misclassifications (out of 34 samples) for the test set.

Table 4. Summary of the three-class problem for the leukemia data

	Number of misclassifications							
n	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
Three-class C.V. (rate)	0 (0%)	0 (0%)	1 (3%)	2 (6%)	1 (3%)	2 (6%)	2 (6%)	2 (6%)
Three-class T.S. (rate)	4 (12%)	3 (9%)	3 (9%)	1 (3%)	1 (3%)	2 (6%)	2 (6%)	1 (3%)

'C.V.' represents the number of misclassifications (out of 38 samples) for the training set using the LOOCV procedure. 'T.S.' represents the number of misclassifications (out of 34 samples) for the test set.

measure of how much each gene discriminates between different classes without adjusting for any variabilities. A more elaborate procedure such as the one described in Dudoit *et al*. (2002) for gene selection could be used, but since we want to investigate the performance of our method even when noisy genes are pre-selected, we will not use such a gene selection procedure. We pre-select n genes using the d-score and count the number of misclassified treatments using the leave-one-out cross-validation (LOOCV) procedure. The selection bias is adjusted in such a way that the pre-selection of genes is performed either using the training set only, or it is performed at each iteration of the cross-validation. We summarize the results in Table 2.

The second and third rows in Table 2 contain the numbers of misclassifications using different numbers of genes, with and without correction for the selection bias, respectively. Except for the case where all 1242 genes are selected, the error rate with the selection bias is never smaller than that with the selection bias corrected. When all 1242 genes are selected, the error rate is the same with or without correcting the selection bias, because there cannot be any bias when all the genes are selected.

3.2 Leukemia data from oligonucleotide microarrays

The leukemia data set was first reported in Golub *et al*. (1999), and has been analyzed extensively by a number of investigators. About 7200 genes have been measured. Initially, samples are classified into two types of leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Later, two subclasses in ALL are identified as T-cell and B-cell, and subsequently, it becomes a three-class classification problem. The training set contains 38 samples, while another independent test set contains 34 samples.

Here we consider both the two-class and three-class problems.

We assess the prediction for the samples in the test set as well as the prediction for the samples in the training set using the LOOCV procedure. We pre-select n genes according to their d-scores, with $n = 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}$ and 2^{11}. The results are summarized in Table 3 for the two-class problem and in Table 4 for the three-class problem.

The error rates for predicting samples in the test set are usually higher than the ones from the LOOCV using the training set only. When the number of genes pre-selected is small (e.g. 2^5), the error rate for the samples in the test set is around 12–15%. However, when the number of pre-selected genes increases, the error rate goes down quickly. Because the d-score is a very crude measure for pre-selecting genes, when the number of pre-selected genes is small, there is not enough information to discriminate the samples. When more genes are pre-selected, we are more likely to include discriminant genes and therefore do a better job of classifying them. We also see this in Table 2 for the cDNA arrays: when $n = 30$, the error

rate is about 21%, and it drops to 8% once we pre-select $n = 40$ genes.

The EBC method produces error rates that are satisfactory compared to the ones in the current literature. The results seem to support the claims by Ambroise and McLachlan (2002), that when accounting for selection bias, the error rates for the leukemia data increase. We should point out that without adjusting for selection bias, our classification results contain no more than one misclassification in all cases considered.

The numbers of abundance groups we obtain from the model-based cluster algorithm range from 2 to 41; our model contains at most 41 variance parameters, compared to about 7200 in the model with one distinct variance per gene.

4 SIMULATION STUDIES

We conduct simulation studies to investigate the effects of model size, i.e. the number of nuisance parameters in the model, to the classification precision.

We consider a classification problem for m samples belonging to three classes with $m/3$ samples in each class. For simplicity, we directly simulate the differences D_{gt} based on the normal distribution given by Equation (3), and call the simulated values the gene expression levels in what follows. We simulate m gene expression levels for each of $(9 + B)$ genes. Among these $(9 + B)$ genes, B have expression levels simulated from the standard normal distribution $N(0, 1)$ for all the m samples. Since their expression levels do not discriminate between classes, the B genes are regarded as noisy genes in the classification procedure. For each sample t, three of the remaining nine genes have expression levels simulated from $N(C_t \times \Delta_1, \sigma_1^2)$, another three from $N(C_t \times \Delta_2, \sigma_2^2)$, and the remaining three from $N(C_t \times \Delta_3, \sigma_3^2)$, where $C_t \in \{1, 2, 3\}$ denotes the class of sample t. The quantities Δ_1, Δ_2 and Δ_3 are simulated from the normal priors $N(1, 0.5^2)$, $N(3, 0.5^2)$ and $N(-3, 0.5^2)$, respectively. Due to the differences in the means of simulated gene expression levels, the resulting gene expression profiles will discriminate between classes. The precisions σ_1^{-2}, σ_2^{-2} and σ_3^{-2} are simulated from gamma priors Gamma$(2.5, 4/3)$, Gamma$(3.5, 4/3)$ and Gamma$(3.5, 4/3)$, respectively.

With this simulation scheme, the nine informative genes not only predict the classes of samples but also form three abundance groups, with each set of three genes generated using the same distributions forming one group. Therefore, we have a total of four groups with the B non-informative genes forming the fourth group.

We compare three different methods. The first one is our EBC method. The second one is called the EBTC method, in which we replace the model-based cluster algorithm in the preliminary step by using the true simulated group structure (with the four groups), and carry out the same second step of empirical Bayes classification as in the EBC method. The third is called the EB1C method, in which we assume that all the gene expression levels are in one abundance group, and carry out the same second step of empirical Bayes classification as in the EBC method. In EB1C, all the gene expression levels share the same variance.

We simulate the expression levels of the 2009 ($B = 2000$) genes for $m = 30$ samples, respectively; and we repeat the whole procedure 500 runs. Among the three classification methods, the EBTC is expected to produce the smallest number of classification errors since it uses the true grouping information of the 2009 genes.

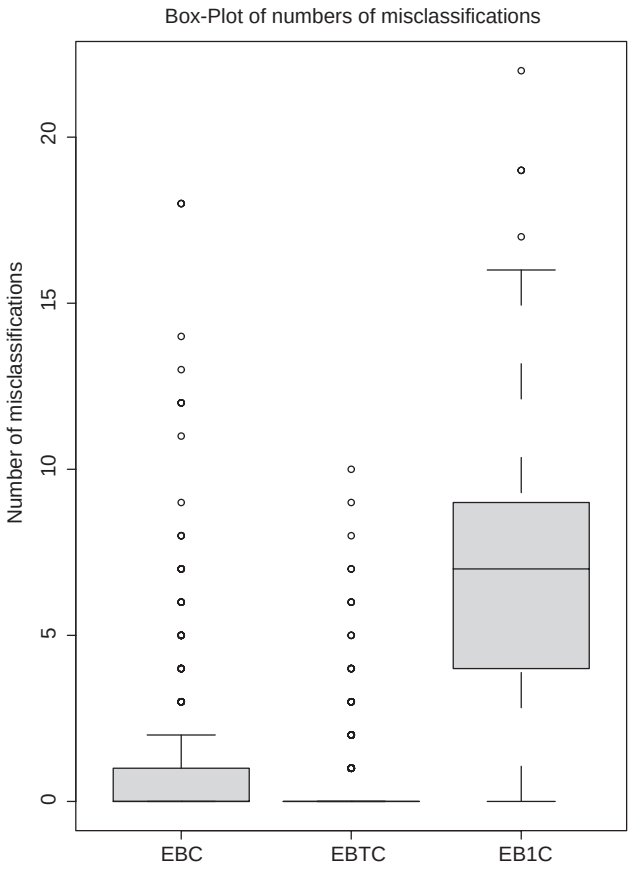

Fig. 1. Box-plots of the numbers of misclassifications (out of $m = 30$ samples) for the three methods when gene expression levels are generated using normal distributions.

The simulation results are summarized in Figure 1. In most runs, the number of misclassifications from the EBC and the EBTC methods are much smaller than the one from the EB1C method. The 95% percentile for the EBC method is only 2, which means that 95% of the time the number of misclassifications is at most 2. For the EB1C method, the 95% percentile is 16. EBTC performs the best as expected, yielding zero misclassification most of the time. Looking closely at the grouping results from the model-based cluster algorithm, we find that the estimated group labels do not all match the true group labels. This seems to suggest that even when the estimated group structure is somewhat different from the true one, the EBC method can still classify samples correctly.

Next, we investigate the performance of our EBC method when the normality assumption is violated. We use the notation gamma(μ, σ^2) to denote a gamma distribution with mean μ and variance σ^2. Specifically, we simulate expression levels of B non-informative genes from gamma$(5, 5)$ for all m samples. The expression levels of the nine informative genes are generated as follows: for sample t in class j, simulate $D_{1t}, D_{4t}, D_{7t} \sim$ gamma(μ_{1j}, σ_1^2), $D_{2t}, D_{5t}, D_{8t} \sim$ gamma(μ_{2j}, σ_2^2) and $D_{3t}, D_{6t}, D_{9t} \sim$ gamma(μ_{3j}, σ_3^2), where the means and variances are simulated from another set of gamma distributions with $\mu_{1j} \sim$ gamma$(j \times 0.5, 1)$, $\mu_{2j} \sim$ gamma$(j \times 5, 1)$ and $\mu_{3j} \sim$ gamma$(j \times 5 + 15, 1)$; $\sigma_1^2 \sim$ gamma$(1, 1)$ and $\sigma_2^2, \sigma_3^2 \sim$ gamma$(2, 1)$. Indices $\{1, 4, 7\}$, $\{2, 5, 8\}$ and $\{3, 6, 9\}$ denote three

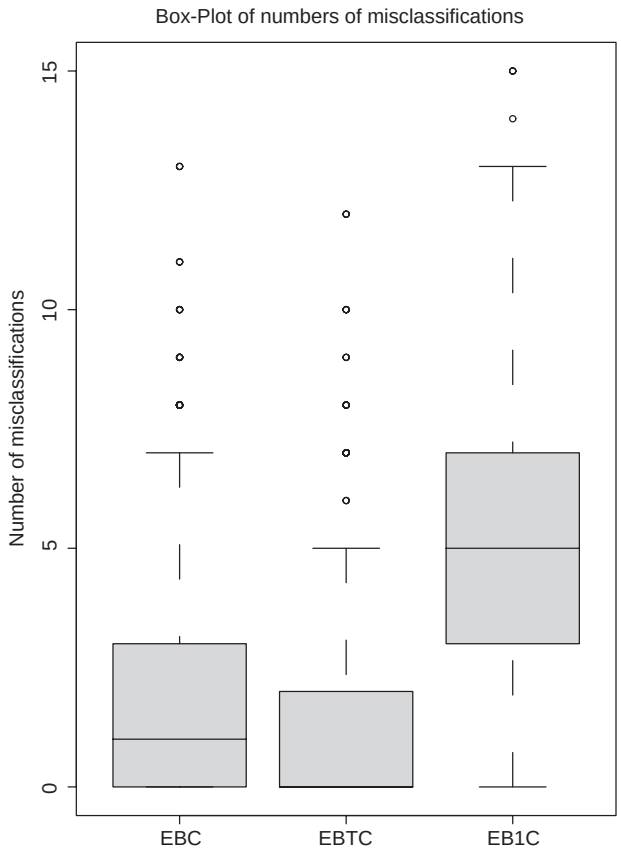

Fig. 2. Box-plots of the numbers of misclassifications (out of $m = 30$ samples) for the three methods when gene expression levels are generated using gamma distributions.

groups of discriminant genes. Because of the way we simulate their expression levels as mentioned above, they discriminate between the classes.

We use $B = 2000$ non-informative genes and $m = 30$ samples. We repeat 500 simulation runs. Figure 2 presents three box-plots of the numbers of misclassifications from the three methods, EBTC, EBC and EB1C. Again, we observe a trend

$$\text{EBTC} < \text{EBC} < \text{EB1C}$$

that describes the order of the mean and variance of the three boxes. The EBC method performs better than the EB1C method, and the EBTC is the best.

The number of groups predicted by the model-based cluster algorithm at each simulation does not always equal 4, but the resulting classification precision does not seem to be affected much by the grouping results. This finding is consistent with the simulation study under the normality assumption.

5 DISCUSSION

Due to the high dimensionality of microarray gene expression data, it is not clear how the traditional Bayesian classification method should be directly implemented. Specifically, since the number of nuisance parameters in Bayesian models significantly impact the final classification results, a large model with many parameters often becomes computationally intractable, and an oversimplified model is not adequate to capture the relationship between genes.

We have developed an empirical Bayes method, EBC, for classifying samples based on gene expression profiles. We use a model-based cluster algorithm as a tool in the preliminary step to group the gene expression levels. As a result, we reduce the number of nuisance parameters in the Bayesian model given by Equations (1)–(3). Specifically, the number of variance parameters equals the number of groups, which is substantially smaller than the number of genes. In contrast, for a model with a distinct variance parameter for each gene, we would have thousands of variances to estimate. We obtain the Bayesian classification rule in a closed form; the time for computing the rule is within minutes with a regular personal computer. Our simulations present inferior classification results when using an oversimplified Bayesian model, where all the genes share the same variance.

We choose a model-based algorithm for grouping genes because it allows us to determine the number of groups objectively using the BIC and because it nicely fits our Bayesian model. The mixture of normal distributions used in the model-based cluster algorithm imply that the gene expression levels within the same group follow the same normal distribution, which is consistent with our model given by Equation (2). In general, we can replace the model-based cluster algorithm with another sensible cluster method, and proceed to use the Bayesian models given by Equations (1)–(3). For example, we obtained very similar results by using the K-means cluster algorithm with the same number of abundance groups as the one suggested by the BIC from the model-based cluster algorithm. However, without knowing the number of groups in advance, the K-means cluster algorithm is hard to implement. Moreover, the proposed two-step method is fully model-based, as our abundance groups are obtained according to a mixture of normal distributions. This is a distinction of our method in comparison to others that use an algorithm-based method. Statistically, since the genes within the same clusters share the same variance parameters in our model, the expression levels become dependent. Accounting for dependency, usually a desired function, is taken into consideration by our model.

Our method is motivated by the biological problem itself, and our model is developed naturally according to the information contained in the data. Other methods such as the support vector machine could also yield nice results, but biological motivation for such methods is difficult and hence may be hard for biologists to understand.

Finally, we discuss the effect of the scaling parameter a in the prior distribution given by Equation (4). In the moment estimation procedure described in Section 3 in the supplemental materials, we used $a = 2$. However, the value of a does not affect the results of the classification much. In the classification rule given by Equation (6), the scaling parameter a only appears in $\xi_j = 1 + 1/(a + k_j)$. When a increases from 0 to ∞, ξ_j decreases from $1 + 1/k_j$ to 1 with only a change of $1/k_j$, where $k_j \geq 1$ for all j. Hence, the value of a does not play a deciding role in the classification probability. However, using a small value of a is preferred because it implies a vague prior for Δ_{gi}.

ACKNOWLEDGEMENTS

The authors thank Chris Bradfield and the Bradfield Laboratory for providing a microarray gene expression data set which is used in one of the examples in this paper. We also thank anonymous

referees for providing constructive comments. This research was partially supported by the NCI grant CA52733, the Merck Foundation fellowship, and the University of Texas SPORE in Prostate Cancer grant CA90270.

REFERENCES

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **90**, 6562–6566.

Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–38.

Dudoit,S., Fridlyand,J. and Speed,T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Assoc.*, **97**, 77–87.

Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Fraley,C. and Raftery,A. (1999) MCLUST: software for model-based cluster analysis. *J. Classif.*, **16**, 297–306.

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

Keller,A., Schummer,M., Hood,L. and Ruzzo,W. (2000) Bayesian classification of DNA array expression data. *Technical Report UW-CSE-2000-08-01*, Department of computer science and engineering, University of Washington, Washington, DC.

McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. John Wiley & Sons, New York.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewar,S., Dmitrovsky,E., Lander,E. and Golub,T. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci.*, **96**, 2907–2912.

Thomas,R., Rank,D., Penn,S., Zastrow,G., Hayes,K., Pande,K., Glover,E., Silander,T., Craven,M., Reddy,J., Jovanovich,S. and Bradfield,C. (2001) Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol. Pharmacol.*, **60**, 1189–1194.

West,M., Blanchette,C., Holly,D., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.,Jr, Marks,J. and Nevins,J. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11 462–11 467.

Yeung,K., Fraley,C., Murua,A., Raftery,A. and Ruzzo,W. (2001) Model based clustering and data transformations for gene expression data. *Biostatistics*, **17**, 977–987.