# Bayesian Models for Gene Expression With DNA Microarray Data

Joseph G. IBRAHIM, Ming-Hui CHEN, and Robert J. GRAY

Two of the critical issues that arise when examining DNA microarray data are (1) determination of which genes best discriminate among the different types of tissue, and (2) characterization of expression patterns in tumor tissues. For (1), there are many genes that characterize DNA expression, and it is of critical importance to try and identify a small set of genes that best discriminate between normal and tumor tissues. For (2), it is critical to be able to characterize the DNA expression of the normal and tumor tissue samples and develop suitable models that explain patterns of DNA expression for these types of tissues. Toward this goal, we propose a novel Bayesian model for analyzing DNA microarray data and propose a model selection methodology for identifying subsets of genes that show different expression levels between normal and cancer tissues. In addition, we propose a novel class of hierarchical priors for the parameters that allow us to borrow strength across genes for making inference. The properties of the priors are examined in detail. We introduce a Bayesian model selection criterion for assessing the various models, and develop Markov chain Monte Carlo algorithms for sampling from the posterior distributions of the parameters and for computing the criterion. We present a detailed case study in endometrial cancer to demonstrate our proposed methodology.

KEY WORDS: Bayesian inference; DNA; Gene expression; Hierarchical prior; L measure; Microarray.

## 1. INTRODUCTION

DNA microarrays and other high-throughput methods for analyzing complex nucleic acid sequences now make it possible to rapidly, efficiently, and accurately measure the levels of virtually all of the genes expressed in a biological sample. Two of the critical issues that arise when examining DNA microarray data are (1) determination of which genes best discriminate between the different types of tissue and (2) characterization of expression patterns in tumor tissues. For (1), because there are so many genes in the human genome, it is of critical importance to possibly identify a small subset of genes that best discriminate between normal and tumor tissues. For (2), it is critical to be able to characterize the DNA expression of the normal and tumor tissue samples and to develop suitable models that explain patterns of DNA expression for these types of tissues.

In DNA microarray data, it is common to have thousands of genes for a single individual and relatively few individuals in the dataset. Developing models for such data structures can be complicated, and computational methods are generally quite intensive. Literature on discrimination and cluster analysis methods in DNA gene expression includes work by Eisen, et al. (1998), Golub et al. (1999), Spellman et al. (1998), Tamayo et al. (1999), Tibshirani et al. (2000), and Hastie et al. (2000). Most of these methods use various types of clustering algorithms, such as self-organizing maps, k-means clustering, and hierarchical clustering, to discriminate and characterize patterns of gene expression. The aim of these methods is to identify clusters of genes that show similar expression patterns. However, none of these methods develop statistical methods for modeling the gene expression levels via a suitable parametric model, and thus they do not address specific questions regarding formal assessment of the differences between the means of the tissue types for each gene or assessment of the fit of a specific model to the data. Some work has been done on using parametric models to characterize patterns of gene expression by Newton, et al. (2001), Chen, et al. (1997), and West, et al. (2000). Related work also includes that of Efron, et al. (2001), who used a nonparametric empirical Bayes procedure for gene profiling. However, these methods are quite different than what we propose here and are discussed in further detail in Section 5.

In this article we develop a novel class of parametric statistical models for analyzing DNA microarray data. It is our goal to build a suitable parametric model that allows us to compare between normal and tumor tissues and to characterize the genes that best distinguish between tissue types. In addition, we wish to develop model assessment techniques so that we can assess the fit of a class of competing models. The Bayesian paradigm is very well suited for examining DNA microarrays, because computation of the posterior distribution is all that is required for making the desired inferences, such as the computation of quantiles, standard deviations, credible sets, and predictions. Because of the data structure of DNA microarrays, frequentist inference using parametric models does not appear feasible, and computing variances and other quantities based on asymptotic theory does not appear tenable. Thus the Bayesian paradigm appears to be better suited for these types of problems. The model that we propose here for modeling gene expression is novel and has several attractive properties. In Section 2 we describe the basic data structure for our DNA microarray. In Section 3 we propose a parametric model for gene expression data and derive some of its properties. In Section 4 we propose a class of hierarchical priors for the parameters that allow us to borrow strength across genes for

Joseph G. Ibrahim is Associate Professor, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02115 (E-mail: *ibrahim@jimmy.harvard.edu*). Ming-Hui Chen is Associate Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 (E-mail: *mhchen@stat.uconn.edu*). Robert J. Gray is Senior Lecturer, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02115 (E-mail: *gray@jimmy.harvard.edu*). The authors thank the editor, associate editor, and two referees for several suggestions that greatly improved the article. The authors also thank George Mutter for his invaluable discussions and for giving permission to use the data. Ibrahim's research was partially supported by National Institutes of Health (NIH) grants CA 70101 and CA 74015. Chen's research was partially supported by National Science Foundation (NSF) grant DMS-9702172 and NIH grants CA 70101 and CA 74015. Gray's research was partially supported by NIH grant CA 57253.

making inferences. The properties of the priors are examined in detail. In Section 5 we propose a *gene selection* algorithm for identifying subsets of genes which may be different across the groups, and in Section 6 we introduce a Bayesian model selection criterion for formally assessing the various models derived from the gene selection algorithm. Markov chain Monte Carlo (MCMC) algorithms for sampling from the posterior distributions of the parameters and for computing the criterion are also developed. We demonstrate our methodology with a detailed data analysis of a dataset in endometrial cancer in Section 7, and give a brief discussion in Section 8.

## 2. THE DATA STRUCTURE

The endometrial data analyzed here were obtained from the Affymetrix Hu6800 DNA array. This array contains more than 7,000 probe sets, which are now thought to represent about 5,600 distinct genes. Each probe set consists of a number (generally 16–20) of perfect match and mismatch probe pairs. The perfect match probe is a single-stranded DNA oligonucleotide complimentary to a specific DNA sequence, approximately 25 bases long. The mismatch probe has a single base mismatch at the central (i.e., 13th) position. Using a number of different probe pairs obtained from the same gene should give greater specificity than is possible with a single probe.

The data analyzed here comprise interpreted relative expression levels given by the proprietary Affymetrix GeneChip software. First, an average hybridization intensity is calculated for each probe from the raw intensity levels read by the scanner. Then the software calculates local background levels across the array; these are subtracted from the intensities for the individual probes. A filtering algorithm is used to remove probe pairs that give overly atypical results, and an average difference value is computed as the average of the differences in the intensity of the binding to the perfect match minus the mismatch probes in the remaining pairs. The algorithm also provides an absolute call of present (P), marginal (M), or absent (A), based in part on a local calculation of noise and the proportion of pairs in which there is greater binding to the mismatch than the perfect match. The data we start with consist of the average difference values, which represent relative levels of expression of the corresponding gene in the sample, and the absolute calls.

In addition to the average difference metric, the Affymetrix software computes three other metrics, which are used in making the absolute call. The three other metrics are the positive fraction, the positive–negative ratio, and the log-average ratio. The positive fraction and the positive–negative ratio are calculated from the number of positive and negative probe pairs. The log-average ratio is derived from the ratio of the perfect match probe cell intensity to that of the control mismatch. The absolute call is determined by examining the three analysis metrics: the positive fraction, the positive–negative ratio, and the log-average ratio.

Nonbiological factors can contribute to the variability of the data in many biological assays. In DNA microarrays, variations in the amount and quality of target hybridized to the array, the amount of stain applied, or other experimental variables may contribute to an overall variability in hybridization intensities. To reliably compare data from multiple probe arrays, differences of nonbiological origin must be minimized through a process known as *normalization*. In the dataset that we examine, the data were normalized as follows. For each sample, the raw average difference values, regardless of the call or the sign of the value, were rescaled to sum to 3,000,000 over all genes. Then average difference values below 20 were set to 20, and the average difference for any probe set with a call of A or M was also set to 20, regardless of the value. This was done in part because the technology was not thought to accurately discriminate among low levels of expression. The value of 20 is arbitrary, but the results are not restrictive to the exact cutoff used. Thus our data analysis uses the four metrics mentioned earlier. The average difference is used as the measure of expression level, and the absolute call is used to determine which values are set to 20.

In addition, there are two primary reasons for truncating the values. First, the Affymetrix technology is thought to produce poor discrimination at low levels of expression (which was mentioned by, e.g. Tamayo et al. 1999). Second, we want to draw inference on expression of identified genes and expressed sequence tags. Differences at low levels (especially at negative values) quite possible result from differences in binding to the mismatch probes, rather than binding to the perfect match probes. What will bind to the mismatch probes is generally not known, so the differences at low or negative expression levels cannot be reliably attributed to the differences in expression of the target genes. If we were only interested in using the array as a predictor and not in inferences on genes, then there probably would be additional information in the low values that we are not using. The threshold of 20 is somewhat arbitrary but has been used elsewhere for this chip (Tamayo et al. 1999), and the inferences are not very sensitive to the exact threshold value (e.g., 10, 30, etc.). The Affymetrix call is a measure of both whether the expression level indicates that the gene is present and of the reliability of the calculated expression level as a measure of the expression of the target gene. In the same spirit of drawing inference on expression of identified targets, we decided to also set calls other than present to the threshold value.

For our choice of normalization, we note that GeneChip applies an algorithm and produces an expression level. We believe that analyzing those levels is valid. However, there is obvious sample-to-sample variation in overall expression levels, and to try to eliminate some of this variation we rescaled to a common overall level. Here we do not consider more elaborate schemes, such as nonparametric regression calibration, because they also involve arbitrary choices and their motivation is less clear than for global rescaling. We use the value 3,000,000 in our dataset because the rescaled sum of expression levels needs to take a common value and we needed to choose some number for that common level. The value 3,000,000 is somewhat arbitrary, but the justification for this choice is based on the fact that it is near the center of the sums of the unnormalized values, which ranged from 1,323,548 to 5,300,324, and affects the inferences only through its relationship with the lower threshold of 20.

The study that we examine here involves patients with endometrial cancer. Identifying gene expression patterns between normal and cancer tissue for this disease is critical,

because it can shed light on the genes that play a major role for endometrial cancer. This critical information can be used to determine heritability of the disease as well as target appropriate therapies for treating this disease.

The data for this specific study comprise 14 individuals (samples). Of these 14 individuals, 4 are normal tissues and 10 are endometrial cancer tissues. Because values below 20 and values with a call of A or M were set to 20, nearly two-thirds of the expression levels were set to 20. Probe sets with resulting value above 20 for fewer than three genes were then excluded from the analysis, leaving data from 3214 genes for each individual. Thus the dataset consists of a $3214 \times 14$ contingency table of gene expression levels, where genes are represented by rows (3214 rows) and the individuals are represented by the columns (14 columns), resulting in a total of $3214 \times 14 = 44{,}996$ gene expression levels. Table 1 gives the number of genes (rows) with expression levels set equal to 20. For example, in the first row of Table 1, 805 genes (rows) in the contingency table had all expression levels not equal to 20 over all 14 individuals. Similarly, the second row of Table 1 reveals that 228 genes (rows) in the contingency table had exactly 1 gene expression level set equal to 20 over all 14 individuals, and so forth.

## 3. THE GENERAL MODEL

Our goal is to develop a statistical model that allows us to characterize the gene behavior in normal and cancer tissues. Toward this goal, we develop the general model as follows. Suppose that we have two tissue types (normal and tumor) with $n_j$ individuals available for each tissue type, and suppose that $G$ genes are measured for each individual. Let $x$ denote the expression level for a given gene, which is the normalized average difference as discussed previously. Because so many of the genes are not expressed, (i.e., take the value 20 for our dataset), the random variable $x$ can be viewed as a mixture random variable with a discrete and continuous component. The discrete component is a point mass at $c_0$ ($c_0 = 20$ for the endometrical cancer data), and the continuous component $y$ is the expression level of the gene. Thus $c_0$ is the threshold value for which a gene is considered as not expressed. Therefore, if $x = c_0$, then the gene is said to be not expressed, that is, absent. If $x = c_0 + y$, then the gene is said to be positively expressed, that is, present. Let $p = P(x = c_0)$, and thus the mixture random variable $x$ can be written as

$$x = \begin{cases} c_0 & \text{with probability } p \\ c_0 + y & \text{with probability } 1 - p. \end{cases} \tag{1}$$

We take $y$ to have a lognormal distribution. This general parametric model appears to be quite natural for modeling gene expression data. Let $j = 1, 2$ index the tissue type (normal vs. tumor) and let $x_{jig}$ denote the gene expression mixture random variable for the $j$th tissue type for the $i$th individual, $i = 1, 2, \ldots, n_j$, and the $g$th gene, $g = 1, \ldots, G$. Similarly, $y_{jig}$ denotes the continuous component of the gene expression level for the $j$th tissue type for the $i$th individual and the $g$th gene. Conditional on the parameters $(\mu_{jg}, \sigma^2_{jg})$, we assume that $y_{111}, \ldots, y_{2,n_2,G}$ are independent each with lognormal density

$$p(y_{jig}|\mu_{jg}, \sigma^2_{jg}) = (2\pi)^{-1/2} y_{jig}^{-1} \sigma_{jg}^{-1} \times \exp\left\{-\frac{1}{2\sigma^2_{jg}}(\log(y_{jig}) - \mu_{jg})^2\right\}. \tag{2}$$

In Section 4 we discuss how the priors incorporate possible dependence among the genes. Now let $\delta_{jig} = 1$ if $x_{jig} = c_0$ and 0 otherwise. Also, let $p_{jg} = P(x_{jig} = c_0) \equiv P(\delta_{jig} = 1)$, and $1 - p_{jg} = P(x_{jig} = c_0 + y_{jig})$, $\boldsymbol{\delta} = (\delta_{111}, \ldots, \delta_{2,n_2,G})$, $\boldsymbol{\mu} = (\mu_{11}, \mu_{21} \cdots \mu_{1G}, \mu_{2G})$, $\boldsymbol{\sigma}^2 = (\sigma^2_{11}, \sigma^2_{21}, \ldots, \sigma^2_{1G}, \sigma^2_{2G})$, and $\mathbf{p} = (p_{11}, p_{21}, \ldots, p_{1G}, p_{2G})$, $i = 1, \ldots, n_j$, $j = 1, 2$, $g = 1, \ldots, G$. Letting $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{p})$, the likelihood function for $\boldsymbol{\theta}$ based on the data $D = (x_{111}, \ldots, x_{2,n_2,G}, \boldsymbol{\delta})$ is thus given by

$$L(\boldsymbol{\theta}|D) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \prod_{g=1}^{G} p_{jg}^{\delta_{jig}} (1 - p_{jg})^{1-\delta_{jig}} \, p(y_{jig}|\mu_{jg}, \sigma^2_{jg})^{1-\delta_{jig}}, \tag{3}$$

where $p(y_{jig}|\mu_{jg}, \sigma^2_{jg})$ is as given in (2).

A fundamental question of interest is which genes best discriminate between the normal and tumor tissues. That is, it is of interest to characterize the behavior of the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, P)$ with respect to the normal and tumor tissues. Toward this goal, let

$$\psi_{jg} = E_{\boldsymbol{\delta}, \mathbf{y}}[c_0 \delta_{jig} + (1 - \delta_{jig})(c_0 + y_{jig})|p_{jg}, \mu_{jg}, \sigma^2_{jg}], \tag{4}$$

where $\mathbf{y} = (y_{111}, \ldots, y_{2,n_2,G})$ and the expectation is with respect to the joint distribution of $(\boldsymbol{\delta}, \mathbf{y})$. Thus we have

$$\begin{aligned} \psi_{jg} &= c_0 p_{jg} + (1 - p_{jg})(c_0 + E(y_{jig}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)) \\ &= c_0 p_{jg} + (1 - p_{jg})\left(c_0 + \exp\left\{\mu_{jg} + \frac{\sigma^2_{jg}}{2}\right\}\right). \end{aligned}$$

Comparing the gene expression level means between the normal and cancer tissues, we compute the posterior distribution of

$$\xi_g = \psi_{2g}/\psi_{1g} \tag{5}$$

for each $g = 1, \ldots, G$. From these $G$ posterior distributions, we can compute various posterior summaries, including the posterior mean, standard deviation, quantiles, and probabilities such as $P(\xi_g > 1|D)$, $g = 1, \ldots, G$.

*Table 1. Summary of Endometrial Cancer Data*

| Number of gene expression levels set equal to 20 | Frequency | Percent | Cumulative frequency | Cumulative percent |
|---|---|---|---|---|
| 0 | 805 | 25.0 | 805 | 25.0 |
| 1 | 228 | 7.1 | 1033 | 32.1 |
| 2 | 209 | 6.5 | 1242 | 38.6 |
| 3 | 181 | 5.6 | 1423 | 44.3 |
| 4 | 192 | 6.0 | 1615 | 50.2 |
| 5 | 159 | 4.9 | 1774 | 55.2 |
| 6 | 208 | 6.5 | 1982 | 61.7 |
| 7 | 212 | 6.6 | 2194 | 68.3 |
| 8 | 214 | 6.7 | 2408 | 74.9 |
| 9 | 223 | 6.9 | 2631 | 81.9 |
| 10 | 280 | 8.7 | 2911 | 90.6 |
| 11 | 303 | 9.4 | 3214 | 100.0 |

## 4. PRIORS

Prior elicitation plays a critical role in this model. We wish to specify a class of priors that creates a correlation between the genes for a given individual. We do this by specifying a hierarchical prior as follows. First, for $\mu_{jg}$ we take

$$\mu_{jg}|\sigma^2_{jg}, \mu_{j0} \sim N(\mu_{j0}, \tau_0\sigma^2_{jg}/\bar{n}_j), \tag{6}$$

where

$$\bar{n}_j = \frac{1}{G}\sum_{g=1}^{G}\left(n_j - \sum_{i=1}^{n_j}\delta_{jig}\right), \tag{7}$$

$\tau_0 > 0$ is a specified scalar, and $\mu_{jg}|\sigma^2_{jg}, \mu_{j0}$ are independent for all $(j, g)$. We note that a solid feature of the prior construction in (7) is that it allows us to incorporate genes for which $x_{jig} = c_0$ for all $i = 1, \ldots, n_j$ for a given $(j, g)$. That is, it allows us to include genes that are unexpressed for all of the individuals within a given tissue type, therefore resulting in a more flexible and general prior. We also take independent inverse gamma (IG) priors for $\sigma^2_{jg}$ and thus take

$$\sigma^2_{jg} \sim \text{IG}(a_{j0}, b_{j0}), \tag{8}$$

where $(a_{j0}, b_{j0})$ are hyperparameters for $j = 1, 2$. The prior specification scheme is quite general in that it allows for different hyperparameters for the normal and tumor groups. The genes for a given individual are correlated, and thus we need to build in a prior correlation between the genes for a given individual and tissue type. To do this, we specify a prior on the values of $\mu_{j0}$ in (6). We take

$$\mu_{j0} \sim N(m_{j0}, v^2_{j0}), \qquad j = 1, 2. \tag{9}$$

We note here that in our model formulation, we allow $m_{j0}$ to depend only on $j$. A more general formulation would be to let $m_{j0}$ also depend on $g$, leading to a mean of $m_{jg0}$. However, doing this would make it difficult to borrow strength. Our assumptions here may be biologically plausible if the genes being examined are close to each other on the chromosome. Also, our model assumptions may be tenable if the genes have similar chemical and biological behaviors. Both of these are common scenarios in DNA microarray data. In any case, our model may be viewed as an approximation to the more complex model that allows $m_{j0}$ to depend on $g$.

The hyperprior on the values of $\mu_{j0}$ in (9) induces a priori correlation between all of the values of $\mu_{jg}$. After some algebra, it can be shown that unconditional on $\mu_{j0}$,

$$(\mu_{jg}, \mu_{jg'})' \sim N_2(\boldsymbol{\mu}^*, \Sigma^*),$$

where $\boldsymbol{\mu}^* = (m_{j0}, m_{j0})'$ and

$$\Sigma^* = \begin{pmatrix} \frac{\tau_0\sigma^2_{jg}}{\bar{n}_j} + v^2_{j0} & v^2_{j0} \\ v^2_{j0} & \frac{\tau_0\sigma^2_{jg'}}{\bar{n}_j} + v^2_{j0} \end{pmatrix}.$$

From this derivation, we see that

$$\text{corr}(\mu_{jg}, \mu_{jg'}|\sigma^2_{jg}, \sigma^2_{jg'}, v_{j0}) = \frac{v^2_{j0}}{\{[\tau_0\sigma^2_{jg}/\bar{n}_j + v^2_{j0}][\tau_0\sigma^2_{jg'}/\bar{n}_j + v^2_{j0}]\}^{1/2}}.$$

This correlation structure has interesting features. We see that as $v^2_{j0} \to \infty$ or $\bar{n}_j \to \infty$, $\text{corr}(\mu_{jg}, \mu_{jg'}|\sigma^2_{jg}, \sigma^2_{jg'}, v_{j0}) \to 1$. This implies that more strength is borrowed as $v^2_{j0}$ is increased or $\bar{n}_j$ is increased. For the prior for $\sigma^2_{jg}$ in (8), we take $a_{j0}$ fixed and $b_{j0}$ random for our hierarchical prior. Specifically, we take a gamma prior for $b_{j0}$, that is,

$$b_{j0} \sim \text{gamma}(q_{j0}, t_{j0}), \tag{10}$$

where $(q_{j0}, t_{j0})$ are specified hyperparameters. This prior specification gives us more flexibility and control in borrowing strength across genes as compared to taking $(a_{j0}, b_{j0})$ as fixed hyperparameters. Specifically, the prior on $b_{j0}$ in (10) allows a great deal of control in inducing prior correlation between the genes.

For the values of $p_{jg}$, we specify the prior as follows. We first let

$$e_{jg} = \text{logit}(p_{jg}) = \log\left(\frac{p_{jg}}{1 - p_{jg}}\right) \tag{11}$$

and then specify a normal prior on the values of $e_{jg}$, thus inducing a prior on the values of $p_{jg}$. Thus we take

$$e_{jg} \sim N(u_{j0}, k_{j0}w^2_{j0}), \qquad j = 1, 2. \tag{12}$$

We choose the prior in this way because it is computationally more stable than a direct prior specification on the values of $p_{jg}$. Finally, for the prior for $e_{jg}$ in (12), we take

$$u_{j0} \sim N(\hat{u}_{j0}, h_{j0}w^2_{j0}), \qquad j = 1, 2. \tag{13}$$

In (12) and (13), $\mathbf{k}_0 = (k_{10}, k_{20})$, $\mathbf{h}_0 = (h_{10}, h_{20})$, and $w^2_{j0}$, $j = 1, 2$, are the specified hyperparameters.

In general, if we do not have sufficient historical data or expert opinion for specifying $(m_{j0}, v^2_{j0})$ and $(\hat{u}_{j0}, w^2_{j0})$, then it is desirable to have a set of guide values on which sensitivity analyses can be based. Toward this goal, we propose some potential data-based guide values for these hyperparameters. Because $m_{j0}$ represents a mean gene expression level on the natural logarithm scale, a guide value for $m_{j0}$ is

$$m_{j0} = \frac{1}{N_j}\sum_{i=1}^{n_j}\sum_{g=1}^{G}(1 - \delta_{jig})\log(y_{jig}), \tag{14}$$

where $N_j = \sum_{i=1}^{n_j}\sum_{g=1}^{G}(1 - \delta_{jig})$. Thus a reasonable guide value for $m_{j0}$ is the gene sample mean on the natural logarithm scale over all individuals for tissue type $j$. For $v^2_{j0}$, we take $v^2_{j0} = \eta_{j0}\text{MSG}_j$, where $\text{MSG}_j = 1/(G-1)\sum_{g=1}^{G} n_{jg}(m_{jg0} - m_{j0})^2$, $m_{jg0} = (\sum_{i=1}^{n_j}(1 - \delta_{jig})\log(y_{jig})/\sum_{i=1}^{n_j}(1 - \delta_{jig}))$, and $n_{jg} = n_j - \sum_{i=1}^{n_j}\delta_{jig}$ for $j = 1, 2$, and $\boldsymbol{\eta}_0 = (\eta_{10}, \eta_{20})$ is a vector of chosen scalers. Thus $\text{MSG}_j$ is the mean square of error between genes on the natural logarithm scale. A guide value for $t_{j0}$ is $t^{-1}_{j0} = d_{j0}\text{MSE}_j$, where $\text{MSE}_j = 1/(N_j - G)\sum_{g=1}^{G}\sum_{i=1}^{n_j}(1 - \delta_{jig})(\log(y_{jig}) - m_{jg0})^2$ for $j = 1, 2$ and $\mathbf{d}_0 = (d_{10}, d_{20})$ is a vector of chosen scalars. We see that $\text{MSE}_j$ is just the mean squared error for the gene expression levels on the natural logarithm scale for tissue type $j$. Thus a reasonable guide value for $t^1_{j0}$ is a scalar multiple of the MSE for the $j$th tumor type.

For $\hat{u}_{j0}$ in (13), we propose a guide value of $\hat{u}_{j0} = \log((1/G)\sum_{g=1}^{G}\hat{p}_{jg}/1-(1/G)\sum_{g=1}^{G}\hat{p}_{jg})$, where $\hat{p}_{jg}$ is the sample proportion of gene expression values equal to $c_0$ over all of the individuals for the $j$th tumor type. This guide value for $u_{j0}$ seems quite suitable based on the definition of $e_{jg}$ in (11). Finally, for $w_{j0}^2$, we take a guide value of the form $w_{j0}^2 = \{((1/G)\sum_{g=1}^{G}\hat{p}_{jg})(1-(1/G)\sum_{g=1}^{G}\hat{p}_{jg})\}^{-1}$. Thus we see that the guide value for $w_{j0}^2$ is just the frequentist variance of $(1/G)\sum_{g=1}^{G}\hat{p}_{jg}$.

Finally, we note that at first sight our model appears to be highly parameterized. However, we note that the parameters of primary interest in our model are $\mu_{j0}, v_{j0}, m_{j0}, a_{j0}$, and $b_{j0}$, and the remaining parameters are nuisance parameters. Thus we have a total of 10 primary parameters of interest. In this sense our model is not highly parameterized and is in fact, not overparameterized.

## 5. GENE SELECTION

The fundamental problem is to identify which genes are different between the two tissue types. In this article we develop the following procedure for determining this:

1. Compute the posterior distributions of all the values of $\xi_g$, $g = 1, \ldots, G$, and for each $\xi_g$ compute
$$\gamma_g = P(\xi_g > 1/D).$$
2. Select a "threshold" value, denoted by $\gamma_0$, for $\gamma_g$, for determining which genes are different. Possible values of $\gamma_0$ might be $\gamma_0 = .7$, $\gamma_0 = .8$, $\gamma_0 = .9$, and $\gamma_0 = .95$. Thus if, for example, $\gamma_0 = .7$, then gene $g$ is declared to be different for the two tissue types if $P(\xi_g > 1|D) \geq .70$ or $P(\xi_g > 1|D) \leq .30$. We use this technique as a device for selecting subset models, which are then compared to the overall "full" model.
3. Once a set of genes is declared different using the criterion in step 2, then set the mean parameters for the tissue types to be unequal for that gene in the model. That is, if $P(\xi_g > 1|D) \geq \gamma_0$ or $P(\xi_g > 1|D) \leq 1-\gamma_0$, then set $\mu_{1g} \neq \mu_{2g}$. Otherwise, set $\mu_{1g} = \mu_{2g} \equiv \mu_g$, where $\mu_g$ is treated as unknown. Thus this process creates a submodel from the full model for each chosen $\gamma_0$. Different values of $\gamma_0$ will lead to different submodels. We note that each $\mu_g$ gets assigned a prior that is elicited in a similar fashion as was done for $\mu_{jg}$.
4. Create several submodels using several values of $\gamma_0$ and using the algorithm in step 3. We then evaluate all the submodels using a Bayesian criterion. The criterion that we use to evaluate the models is called the *L measure*. We briefly describe the L measure in Section 6.
5. The model with the smallest L measure is deemed the best-fitting model.

As mentioned in Section 1, three other articles use parametric model-based approaches for identifying differential gene expression. Chen et al. (1997) considered a frequentist model-based hypothesis testing approach for characterizing differential gene expression. They first assumed that the gene expression levels are independently normally distributed random variables in the normal and tumor tissue groups, and then derive the approximate distribution of the ratio $T_k = R_k/G_k$ assuming a constant coefficient of variation. Here $T_k$ is the ratio of the red $(R_k)$ to green $(G_k)$ intensity for gene $k$. Once the distribution of $T_k$ is derived, a test statistic based on $T_k$ is calculated for testing the null hypothesis of equality of mean expression levels, and approximate confidence intervals of the test statistic are derived. Newton et al. (2001) took a different approach than Chen et al. (1997) in that they first modeled $R_k$ and $G_k$ as independent gamma random variates with constant shape parameter and a scale parameter that depends on the gene index $k$. They then derived the density of the ratio $T_k$, which is indexed by the parameter $\rho = \mu_R/\mu_G$, where $\mu_R = E(R_k)$ and $\mu_G = E(G_k)$. The hierarchical model is completed by specifying a prior for the gamma scale parameters and carrying out the inference based on the posterior distribution of $\rho$. The hyperparameters in the hierarchical model are estimated from the data using the marginal likelihood of the data.

West et al. (2000) considered an approach based on probit and linear regression for characterizing differential gene expression. They started out with a probit model for the data for which there is a binary clinical outcome response variable, such as estrogen receptor (ER) status in breast cancer, and the covariates consist of the gene expression levels. The binary response variable is then transformed into a continuous variable via latent variables, and a linear regression model is used to develop the methodology. Because the linear regression model is overparameterized, West et al. (2000) described techniques based on the singular value decomposition to make the model identifiable and considered a class of generalized singular $g$ priors (Zellner 1986) for the regression coefficients. Then inference is based on the posterior distribution of the linearly transformed regression coefficients. Efron et al. (2001) considered a nonparametric approach to gene profiling. Their problem involved characterizing gene expression for a group of patients receiving treatment versus those not receiving treatment. Unlike in the other three approaches discussed above, here Efron et al. (2001) did not use a fully parametric model, but rather considered an empirical Bayes approach to obtain a nonparametric estimate of the probability that a particular gene was affected by the treatment. All of these four approaches are quite different than the model (3) proposed in this article. First, we model the gene expression level as a mixture random variable, taking into account the absolute call in the Affymetrix chip in the modeling scheme. Furthermore, we use a lognormal model for the continuous part of the mixture unlike the aforementioned approaches. Third, we consider novel classes of priors that induce correlation structures between genes and use a gene selection algorithm along with the L measure statistic for evaluating models.

## 6. THE L MEASURE FOR MODEL ASSESSMENT

To evaluate the possible models that arise from our gene selection procedure described in Section 5, we consider a useful Bayesian criterion, the *L measure*, originally introduced by Ibrahim and Laud (1994) and Laud and Ibrahim (1995). This criterion is constructed from the posterior predictive distribution of the data and can be written as a sum of two components, one involving the means of the posterior predictive

distribution and the other involving the variances. It can be viewed as a Bayesian goodness-of-fit statistic that measures the performance of a model by a combination of how close its predictions are to the observed data and the variability of the predictions. The L measure can be used as a general model assessment tool for comparing models and assessing goodness of fit for a particular model, and thus in this sense this criterion is potentially quite versatile.

The construction of the L measure relies on the notion of an imaginary replicate experiment. Let $\mathbf{z} = (z_{111}, \ldots, z_{2,n_2,G})$ denote future values of a replicate experiment. That is, $\mathbf{z}$ is a future response vector with the same sampling density (likelihood) as $\mathbf{x}|\boldsymbol{\theta}$, where $\mathbf{x}$ is defined by (1) and the corresponding likelihood (3). Thus $z_{jig}$ is of the form

$$z_{jig} = \begin{cases} c_0 & \text{with probability } p \\ c_0 + r_{jig} & \text{with probability } 1-p, \end{cases} \tag{15}$$

where $r_{jig}$ has the lognormal density in (2). For convenience of notation, denote the sampling density of $\mathbf{x}$ by $p(\mathbf{x}|\boldsymbol{\theta})$ and that of $\mathbf{z}$ by $p(\mathbf{z}|\boldsymbol{\theta})$. Here we note that $\mathbf{x}$ and $\mathbf{z}$ may represent a transformation of the original data, such as logarithms. In the analyses of Section 7, we take the logarithms of $\mathbf{x}$. The idea of using a future response vector $\mathbf{z}$ in developing a criterion for assessing a model or comparing several models has been well motivated in the literature by Geisser (1993) and the many references therein, along with Ibrahim and Laud (1994), Laud and Ibrahim (1995), Gelfand and Ghosh (1998), and Ibrahim, Chen, and Sinha (2001). The imagined replicate experiment makes $\mathbf{x}$ and $\mathbf{z}$ directly comparable and exchangeable a priori. It seems clear that good models should make predictions close to what has been observed for an identical experiment. With this notion in mind, Ibrahim and Laud (1994) defined their statistic as the expected squared Euclidean distance between $\mathbf{x}$ and $\mathbf{z}$, given by the statistic

$$L_{\text{IL}} = E[(\mathbf{z}-\mathbf{x})'(\mathbf{z}-\mathbf{x})], \tag{16}$$

where the expectation is taken with respect to the posterior predictive distribution of $\mathbf{z}|D$, given by

$$p(\mathbf{z}|D) = \int p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}, \tag{17}$$

and $p(\boldsymbol{\theta}|D)$ denotes the posterior distribution of $\boldsymbol{\theta}$. Straightforward algebra shows that $L_{\text{IL}}$ can be written as

$$L_{\text{IL}} = \sum_{g=1}^{G}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\{\text{var}(z_{jig}|D) + (E(z_{jig}|D) - x_{jig})^2\}, \tag{18}$$

and thus $L_{\text{IL}}$ can be written as a sum of two terms, one term involving the predictive variances and the other like a bias term involving the squared difference between the predictive means and the observed data.

A more general version of (17) has been developed by Ibrahim, Chen, and Sinha (2001) and takes the form

$$L = \sum_{g=1}^{G}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\text{var}(z_{jig}|D) + \nu\sum_{g=1}^{G}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(E(z_{jig}|D) - x_{jig})^2, \tag{19}$$

where $0 < \nu < 1$. The quantity $\nu$ plays a major role in (19), it can be interpreted as a weight term in the squared bias component of (19). Ibrahim and Laud (1994) used $\nu = 1$ and thus gave equal weight to the squared bias and variance components. Allowing $\nu$ to vary between 0 and 1 gives us a great deal of flexibility in the trade-off between bias and variance, and thus results in values of $\nu$ that are more desirable than the others. This suggests the question of whether certain values of $\nu$ are "optimal" in some sense for model selection purposes. Ibrahim et al. (2001) addressed this optimality issue for the linear model and theoretically showed that certain values of $\nu$ yield more highly desirable properties of the L measure than other values of $\nu$. Based on their theoretical exploration, $\nu = 1/2$ is a desirable and justifiable choice for model selection.

It can be shown that (19) can be expressed as a posterior expectation, so that

$$L = \sum_{g=1}^{G}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\{E_{\boldsymbol{\theta}|D}(E[(z_{jig})^2|\boldsymbol{\theta}]) - \mu_{jig}^{(\mathbf{z})}\} + \nu\sum_{g=1}^{G}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(\mu_{jig}^{(\mathbf{z})} - x_{jig})^2, \tag{20}$$

where $\mu_{jig}^{(\mathbf{z})} = E_{\boldsymbol{\theta}|D}[E(z_{jig}|\boldsymbol{\theta})]$. Thus (20) can be computed by sampling from the posterior distribution of $\boldsymbol{\theta}$ via MCMC methods. Once the posterior samples of $\boldsymbol{\theta}$ are obtained, (20) can be easily evaluated. The L measure in (20) is finite as long as the first two moments of $\mathbf{z}|\boldsymbol{\theta}$ are finite, and the expectation of $p(\mathbf{z}|\boldsymbol{\theta})$ is finite with respect to the posterior distribution $p(\boldsymbol{\theta}|D)$.

Here we mention that the lognormal component of the model given in (2) has location and scale parameters depending on tissue type ($j$) and gene ($g$). This is a general development that allows both the mean and the variance of the expression level to depend on the tissue type as well as on the gene. Because our main goals here are to assess differential gene expression level across tissue types and to characterize the genes that have similar expression levels, specifying a general model that allows for the mean and the variance of the expression level to depend on the tissue type as well as the gene is critical for our development. A simpler parameterization would not allow us to assess the general gene expression patterns that arise between the tissue types and across the different genes. In addition, (3) can be viewed as the "full model" in some sense, because in our gene selection algorithm we do simplify the parameterization of the model through the L measure statistic, which is based on the data. Thus we prefer to let the data dictate the simplification in the parameterization via the L measure statistic rather than make this assumption at the outset. The L measure is a very valuable model assessment statistic that lets the data dictate a simpler parameterization for our model.

In addition, the parameters $\mu_{jg}, \sigma_{jg}$, and $p_{jg}$ can be viewed as random effects in our model, and all have prior distributions whose hyperparameters do not depend on the gene index $g$, thus simplifying the parameterization a lot. Thus, viewing our model as a random-effects type model with appropriate

hierarchical priors, we can see that the model indeed is not highly parameterized. Also, as long as the priors are proper for our model, the joint posterior distribution of all of the parameters will always be proper. We also note that for each $(j, g)$ combination, we have $n_j$ observations, so that there is "replication" in some sense, and thus the model in (3) is in fact not overparameterized. We also note that at first sight, the model appears to be highly parameterized. However, we note that the parameters of primary inferential interest in our model are $\mu_{j0}, v_{j0}, m_{j0}, a_{j0}$, and $b_{j0}$ (none of which depend on the gene index $g$), and the remaining parameters are nuisance parameters. Thus we have a total of 10 primary parameters of interest. In this sense, our model is not highly parameterized at all, and in fact is not overparameterized.

## 7. CASE STUDY IN ENDOMETRIAL CANCER

Here we consider the endometrial cancer dataset described in Section 2. Our main focus in this application is to identify subsets of genes that have markedly different expression levels between the normal and cancer tissues. We do this using our proposed model, the gene selection algorithm, prior distributions, and the L measure. In addition, we demonstrate how various choices of prior parameters affect the number of genes selected and model choice in general.

For these data, the normalization was carried out as follows. The raw average difference values, regardless of the call or the sign of the value, were rescaled within each sample to sum to 3,000,000 over all of the genes. Then values below 20 and values with a call of A or M were set to $C_0 = 20$. First, we consider the cases for which $\gamma_0 = .7, .80, .90$, and .95, and under each case identify which genes are different. Table 2 shows the number of genes declared to be different under each criterion. For Tables 2 and 3, the prior hyperparameters are taken to be the guide values. Thus for $\mu_{j0}$, we take $m_{10} = 6.26$, $m_{20} = 6.18$, $v_{10}^2 = 4.50\eta_{10}$, and $v_{20}^2 = 10.34\eta_{20}$. The hyperparameters for $b_{j0}$ are $t_{10}^2 = .169d_{l0}$ and $t_{20}^2 = .268d_{20}$, and the hyperparameters for $e_{j0}$ are $\hat{u}_{10} = -.806$, $\hat{u}_{20} = -.602$, $w_{10}^2 = 4.69$, and $w_{20}^2 = 4.37$. In both Tables 2 and 3, we take $\eta_{10} = \eta_{20} = d_{10} = d_{20} = k_{10} = k_{20} = h_{10} = h_{20}$. Thus $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = .01$ implies that $\eta_{10} = \eta_{20} = d_{10} = d_{20} = k_{10} = k_{20} = h_{10} = h_{20} = .01$.

From Table 2 we see that as $\gamma_0$ is increased, fewer genes are identified as being different. For example, under moderately informative priors, when $\gamma_0 = .95$, 178 genes were declared as different, and when $\gamma_0 = .70$, 1,350 genes were declared as different. Table 3 shows the posterior means and standard deviations of the values of $\mu_{j0}$ and the L measure values (using $\nu = 1/2$) for the full model using various choices of prior parameters. We see from Table 3, that if the prior becomes too informative, i.e., $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ is very small, then the L measure value for the model is large. Despite the difference in L measure values, the posterior means and standard deviations of $\mu_{j0}$ are very robust as the prior parameters are varied. We see that the posterior mean [with standard deviation (SD)] for $\mu_{10}$ is 6.08 .013 and the posterior mean for $\mu_{20}$ is approximately 6.016 (.0099) for nearly all of the choices of $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ reported SD in Table 3.

*Table 2. Number of Genes Declared to be Different Based on Several Choices of Hyperparameter Values for $\mu_{j0}$ and $e_{jg}$ With Various Choices of $\gamma_0$*

| | $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ | | | |
|---|---|---|---|---|
| $\gamma_0$ | .01 | .02 | .05 | .10 |
| .95 | 178 | 167 | 154 | 115 |
| .90 | 316 | 290 | 271 | 283 |
| .80 | 695 | 668 | 629 | 674 |
| .70 | 1,350 | 1,266 | 1,191 | 1,209 |

*Table 3. Posterior Summaries for the Full Model Using Several Choices of Hyperparameter Values for $\mu_{j0}$ and $e_{jg}$*

| | Mean | | |
|---|---|---|---|
| $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ | Normal tissue | Cancer tissue | L measure |
| .01 | 6.08 (.013) | 6.013 (.0097) | 180,837 |
| .02 | 6.08 (.013) | 6.015 (.0099) | 177,047 |
| .05 | 6.08 (.013) | 6.016 (.0099) | 167,057 |
| .10 | 6.08 (.013) | 6.016 (.0099) | 155,624 |

NOTE: Standard deviations are in parentheses.

Tables 4 and 5 compare our results with the PERMAX criterion developed by Mutter et al. (2001). In PERMAX, standard pooled variance $t$ statistics for comparing normal tissues to tumor tissues are computed for each gene. We let $t_g$ denote the $t$ statistic for the $g$th gene. To nonparametrically determine the significance of each gene while controlling the overall error rate, we use the permutation distribution of the most extreme statistics over all the genes. Because the distributions of the $t$ statistics are not symmetric with unequal group sizes, this is done separately in each tail. Assuming that positive values of $t_g$ indicate higher values in normal tissues, and letting $t^{(p)}$ be the maximum statistic over all the genes for the $p$th permutation, the $p$ value for gene $g$ in the direction of higher expression in normal tissues is the proportion of permutations where the observed $t_g$ is $\geq t^{(p)}$, with a similar calculation in the opposite tail for differences in the opposite direction. Software for PERMAX analysis is available at http://biowww.dfci.harvard.edu/~gray. For the analysis of the endometrial data using PERMAX, the expression levels were normalized and truncated as described earlier. Only the genes with at least 2 values above 20 were considered (3,214 genes), because the $t$ statistic is undefined for genes with all values equal to 20, and the statistic is either 1.69 or $-.62$ when only one value is not equal to 20. Log expression levels were used in the statistics, because it is natural to think of differences between tissue types as a multiplicative effect of fold increase or decrease. Genes with $p$ values below .25 in either direction were selected as being the most interesting for possible further examination.

For our proposed model, the hyperparameters for the informative and moderate priors used in Tables 4 and 5 are $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = .01$ and $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = 1$. We see that the PERMAX criterion identifies 47 genes as being different in the tumor and normal tissues. The first column in Table 4 gives a blinded gene code. The next four columns are indicators which show the results for the 95%, 90%, 80%, and 70% criteria; in these columns where a 1 means that the gene satisfied

*Table 4. Comparisons With PERMAX Criterion: Informative Priors ($\boldsymbol{\eta}_0 = .01$, $\mathbf{d}_0 = .01$, $\mathbf{k}_0 = .01$, $\mathbf{h}_0 = 0.01$)*

| Gene code | 95% | 90% | 80% | 70% | Probability |
|---|---|---|---|---|---|
| x1047 | 0 | 1 | 1 | 1 | .0764 |
| x1195 | 0 | 0 | 0 | 1 | .7791 |
| x1222 | 0 | 0 | 0 | 0 | .6276 |
| x1223 | 1 | 1 | 1 | 1 | .9615 |
| x1271 | 1 | 1 | 1 | 1 | .9967 |
| x1469 | 0 | 0 | 1 | 1 | .1080 |
| x151 | 0 | 0 | 0 | 1 | .2598 |
| x1553 | 1 | 1 | 1 | 1 | .9937 |
| x1629 | 0 | 0 | 0 | 1 | .7100 |
| x1661 | 0 | 1 | 1 | 1 | .9156 |
| x197 | 0 | 0 | 1 | 1 | .1765 |
| x2140 | 0 | 0 | 0 | 0 | .4043 |
| x2192 | 0 | 0 | 1 | 1 | .8468 |
| x2341 | 0 | 0 | 1 | 1 | .8579 |
| x2404 | 0 | 0 | 0 | 0 | .6488 |
| x2590 | 0 | 0 | 1 | 1 | .8103 |
| x2719 | 0 | 0 | 0 | 1 | .7111 |
| x2797 | 0 | 0 | 0 | 0 | .4557 |
| x3108 | 0 | 0 | 0 | 1 | .7065 |
| x4293 | 0 | 1 | 1 | 1 | .9496 |
| x4495 | 1 | 1 | 1 | 1 | .0128 |
| x4535 | 0 | 0 | 0 | 0 | .4968 |
| x4685 | 0 | 0 | 0 | 0 | .6931 |
| x4693 | 0 | 1 | 1 | 1 | .9263 |
| x4880 | 0 | 0 | 0 | 1 | .7280 |
| x4985 | 0 | 0 | 0 | 1 | .2566 |
| x5023 | 0 | 0 | 0 | 0 | .4504 |
| x5238 | 1 | 1 | 1 | 1 | .9637 |
| x5284 | 0 | 0 | 0 | 1 | .7539 |
| x5442 | 0 | 0 | 1 | 1 | .1464 |
| x5612 | 1 | 1 | 1 | 1 | .0057 |
| x5624 | 1 | 1 | 1 | 1 | .0027 |
| x5757 | 1 | 1 | 1 | 1 | .0001 |
| x5912 | 1 | 1 | 1 | 1 | .9998 |
| x6197 | 0 | 0 | 0 | 1 | .7038 |
| x6235 | 0 | 0 | 1 | 1 | .8232 |
| x6335 | 1 | 1 | 1 | 1 | .9950 |
| x6741 | 0 | 0 | 1 | 1 | .8231 |
| x7019 | 1 | 1 | 1 | 1 | .9792 |
| x7065 | 0 | 1 | 1 | 1 | .9181 |
| x724 | 0 | 0 | 1 | 1 | .8563 |
| x860 | 0 | 0 | 0 | 0 | .6391 |
| x879 | 0 | 0 | 1 | 1 | .1199 |
| x881 | 0 | 0 | 0 | 1 | .7943 |
| x935 | 1 | 1 | 1 | 1 | .9767 |
| x723 | 1 | 1 | 1 | 1 | .0034 |
| x2670 | 1 | 1 | 1 | 1 | .0279 |
| Sum | 14 | 19 | 29 | 39 | |

*Table 5. Comparisons With PERMAX Criterion: Moderate Priors ($\boldsymbol{\eta}_0 = 1$, $\mathbf{d}_0 = 1$, $\mathbf{k}_0 = 1$, $\mathbf{h}_0 = 1$)*

| Gene code | 95% | 90% | 80% | 70% | Probability |
|---|---|---|---|---|---|
| x1047 | 0 | 0 | 1 | 1 | .1110 |
| x1195 | 1 | 1 | 1 | 1 | .0385 |
| x1222 | 1 | 1 | 1 | 1 | .0279 |
| x1223 | 1 | 1 | 1 | 1 | .9614 |
| x1271 | 1 | 1 | 1 | 1 | .9957 |
| x1469 | 1 | 1 | 1 | 1 | .0097 |
| x151 | 1 | 1 | 1 | 1 | .0173 |
| x1553 | 1 | 1 | 1 | 1 | .9932 |
| x1629 | 1 | 1 | 1 | 1 | .0404 |
| x1661 | 0 | 1 | 1 | 1 | .9216 |
| x197 | 1 | 1 | 1 | 1 | .0157 |
| x2140 | 1 | 1 | 1 | 1 | .0347 |
| x2192 | 0 | 1 | 1 | 1 | .0623 |
| x2341 | 1 | 1 | 1 | 1 | .9699 |
| x2404 | 0 | 1 | 1 | 1 | .0595 |
| x2590 | 1 | 1 | 1 | 1 | .0462 |
| x2719 | 0 | 1 | 1 | 1 | .0784 |
| x2797 | 0 | 1 | 1 | 1 | .9488 |
| x3108 | 1 | 1 | 1 | 1 | .0322 |
| x4293 | 1 | 1 | 1 | 1 | .9870 |
| x4495 | 1 | 1 | 1 | 1 | .0024 |
| x4535 | 1 | 1 | 1 | 1 | .0438 |
| x4685 | 0 | 1 | 1 | 1 | .0781 |
| x4693 | 0 | 1 | 1 | 1 | .9367 |
| x4880 | 1 | 1 | 1 | 1 | .0312 |
| x4985 | 0 | 0 | 0 | 1 | .7156 |
| x5023 | 0 | 1 | 1 | 1 | .0870 |
| x5238 | 1 | 1 | 1 | 1 | .0355 |
| x5284 | 1 | 1 | 1 | 1 | .0355 |
| x5442 | 1 | 1 | 1 | 1 | .0414 |
| x5612 | 1 | 1 | 1 | 1 | .0178 |
| x5624 | 1 | 1 | 1 | 1 | .0128 |
| x5757 | 1 | 1 | 1 | 1 | .0004 |
| x5912 | 1 | 1 | 1 | 1 | .9997 |
| x6197 | 1 | 1 | 1 | 1 | .0372 |
| x6235 | 1 | 1 | 1 | 1 | .0397 |
| x6335 | 1 | 1 | 1 | 1 | .9940 |
| x6741 | 0 | 0 | 0 | 0 | .5881 |
| x7019 | 1 | 1 | 1 | 1 | .9765 |
| x7065 | 0 | 1 | 1 | 1 | .9266 |
| x724 | 0 | 0 | 0 | 1 | .2258 |
| x860 | 0 | 0 | 1 | 1 | .1076 |
| x879 | 1 | 1 | 1 | 1 | .0257 |
| x881 | 1 | 1 | 1 | 1 | .0368 |
| x935 | 1 | 1 | 1 | 1 | .9740 |
| x723 | 1 | 1 | 1 | 1 | .0059 |
| x2670 | 1 | 1 | 1 | 1 | .0069 |
| Sum | 33 | 42 | 44 | 46 | |

the criterion. Thus for gene x1047 for example, the 95% criterion is 0, which means that the posterior probability of $\xi_{\text{x1047}}$ was not greater than .95 or less than .05; in fact, it was .0764. The last row in Table 4, labeled "Sum," gives the sum of the genes from the PERMAX criterion that satisfy the 95%, 90%, 80%, and 70% criteria. That is, this row gives us the number of genes that were identified both by the PERMAX criterion and our criteria—the number of genes that overlapped. For the 95% criterion, 14 genes overlapped; for 90%, 19 genes overlapped; for 80%, 29 genes overlapped; and for 70%, 39 genes overlapped. These results thus show that the degree of overlap is monotonically increasing with a decreasing percentage criterion.

Table 5 gives the same comparisons but with moderate priors. We see that when more moderate priors are considered, we get a much higher degree of overlap, because more genes are identified as being different as the prior becomes less informative. For the 95% criterion, 33 genes overlapped; for 90%, 42 genes overlapped; for 80%, 44 genes overlapped; and for 70%, 46 genes overlapped. Thus the moderate priors give more overlapped genes with the PERMAX criterion. To further assess the sensitivity to the choice of prior, we considered a small perturbation of $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ for the full model. The results are robust to minor perturbations of $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$. For example, with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (.9, .9, .9, .9)$, the number of genes that overlap with the PERMAX criterion are 32, 42, 44, and 46 under the 95%, 90%, 80%, and 70% criteria. Moreover, the ratio of the L measures for the model with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (.9, .9, .9, .9)$ to the model with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1, 1, 1, 1)$ is 1.011 using $\nu = 1/2$. So the L measures are quite close under these two choices of hyperparameters. Also, with these hyperparameters, the posterior

mean and standard deviation of $\mu_{10}$ (normal tissues) are 6.083 and .013. For the tumor tissues, the posterior mean and standard deviation of $\mu_{20}$ are 6.017 and .010. Similar results were observed with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1.1, 1.1, 1.1, 1.1)$. Specifically, the number of genes overlap with the PERMAX criterion are 33, 42, 44, and 46 under the 95%, 90%, 80%, and 70% criteria. Moreover, the ratios of the L measures for the model with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1.1, 1.1, 1.1, 1.1)$ to the model with $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1, 1, 1, 1)$ is .991, so the L measures are quite close under these two choices of hyperparameters. Also, under these hyperparameters, the posterior mean and standard deviation of $\mu_{10}$ are 6.083 and .013. For the tumor tissues, the posterior mean and standard deviation of $\mu_{20}$ are 6.017 and .010. We conclude that minor perturbations in $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ lead to quite robust parameter estimates, L measure values, and gene selection criteria.

Table 6 gives L measure values with $\nu = 1/2$ based on several models using the hyperparameters $\eta_{10} = \eta_{20} = 10$, $d_{10} = d_{20} = 10$, and $k_{10} = k_{20} = h_{10} = h_{20} = 50$. With these hyperparameters, the model based on the 70% criterion is the best-fitting model according to the L measure, and the models based on the 80% and 90% criteria gives similar fits. We also note that the PERMAX criterion has the largest L measure. Table 6 demonstrates how our gene selection algorithm of Section 5 works. We see that the larger the $\gamma_0$, the fewer the number of genes declared to be different. The smaller the $\gamma_0$, the greater the number of genes that are declared different. Thus a criterion using $\gamma_0 = .95$ or $\gamma_0 = .90$, for example, would declare fewer genes to be different than would a criterion using a smaller value of $\gamma_0$. Table 6 indicates that there is a trade-off between the number of genes declared to be different and the goodness of fit (and complexity) of the model as indicated by the L measure. Thus a model with $\gamma_0 = .90$ declares fewer genes to be different, but at the same time is a more parsimonious model with fewer parameters than a model with $\gamma_0 = .70$. Thus the L measure for the model with $\gamma_0 = .70$ is smaller than that of the model with $\gamma_0 = .90$. For this reason, our gene selection algorithm, through $\gamma_0$ and the L measure, tries to strike a balance between the number of genes declared to be different and the goodness of fit of the model. The results from Tables 2–6 suggest that many genes are differentially expressed between tumor tissues and normal tissues. Although PERMAX selected fewer genes, it is designed to select a small subset of the most significant genes for further study. In any case, there was a great deal of overlap from the selected genes from our gene selection algorithm and the PERMAX criterion. We include the comparison of PERMAX and our Bayesian model because we feel that it is important to compare our method with a frequentist technique, and there is no other well-known frequentist competitor, and we believe PERMAX to be a reasonable frequentist criterion based on $t$ statistics, like many other criteria currently being used for DNA microarray data. It is possible that our Bayesian model may produce more similar results to PERMAX if we change our gene selection algorithm of Section 5 in step 1 to be of the form $\gamma = P(\xi_g > a|D)$, where $a = 2, 3$, and so on. Here we have chosen $a = 1$. This issue needs further investigation, however, and is currently being examined.

Table 7 shows the L measure for the full model with $\nu = 1/2$ based on several choices of prior parameters. For Table 6, we take both of the components for each of $\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0$, and $\mathbf{h}_0$ to be the same. For example, $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (20, 20, 50, 50)$ implies that $\eta_{10} = \eta_{20} = 20$, $d_{10} = d_{20} = 20$, $k_{10} = k_{20} = 50$, and $h_{10} = h_{20} = 50$. We see that as the prior becomes less informative [i.e., $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ are increased], an improvement in the L measure results. That is, as less strength is borrowed across genes, the overall fit of the model is improved. However, we also note that there is a threshold region for which the L measure actually worsens if the prior becomes too noninformative. That is, there is a region of informativeness in the priors for which the L measure is improved, and there is a threshold for the choices in the prior parameters, so that making the prior too informative or too noninformative results in a poorer fit and hence yields larger L measure values. Overall, from Table 7 we see that the L measure is quite robust with respect to several choices of prior parameters. For example $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (50, 50, 50, 50)$, (20, 20, 50, 50), (100, 100, 50, 50) yield nearly identical L measure values.

To assess the assumption of the lognormal distribution for modeling the gene expression levels, we constructed normal probability plots for each individual (Fig. 1). We see from Figure 1 that the normal probability plots are quite linear for each individual, implying that the assumption of the lognormal distribution is indeed tenable. To further assess the overall fit of the full model, we checked to see if the posterior estimates of $p_{jig}$ and $\xi_g$ are comparable in some sense to the observed data. Figure 2 shows boxplots of $\Delta_{p1} = \hat{p}_{1g} - E(p_{1g}|D)$ and $\Delta_{p2} = \hat{p}_{2g} - E(p_{2g}|D)$ under informative $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (.01, .01, .01, .01)$, moderate $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1, 1, 1, 1)$, and vague $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (100, 100, 50, 50)$ priors. Thus $\Delta_{p1}$ and $\Delta_{p2}$ are the differences in the sample proportions and the posterior means of $p_{jg}$, for $j = 1, 2$. From these boxplots, we see that for vague priors, $\Delta_{p1}$ and $\Delta_{p2}$ are highly concentrated near 0, whereas for informative and moderate priors there is more dispersion, but the medians of $\Delta_{p1}$ and $\Delta_{p2}$ are near 0.

*Table 6. L Measure for Various Values of $\gamma_0$*

| Criterion | L measure | Number of genes declared to be different |
|---|---|---|
| Full model | 98,305 | 3,214 |
| 70% | 97,932 | 2,055 |
| 80% | 98,905 | 1,505 |
| 90% | 102,017 | 1,004 |
| PERMAX | 110,809 | 47 |

*Table 7. L Measure for the Full Model Using Different Values of the Hyperparameters for $\mu_{j0}$ and $e_{jg}$*

| $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0)$ | L measure |
|---|---|
| (1,1,1,1) | 116,246 |
| (10, 10, 10, 10) | 101,326 |
| (20, 20, 20, 20) | 99,699 |
| (100, 100, 20, 20) | 99,690 |
| (50, 50, 50, 50) | 98,307 |
| (20, 20, 50, 50) | 98,307 |
| (100, 100, 50, 50) | 98,307 |
| (10, 10, 50, 50) | 98,305 |

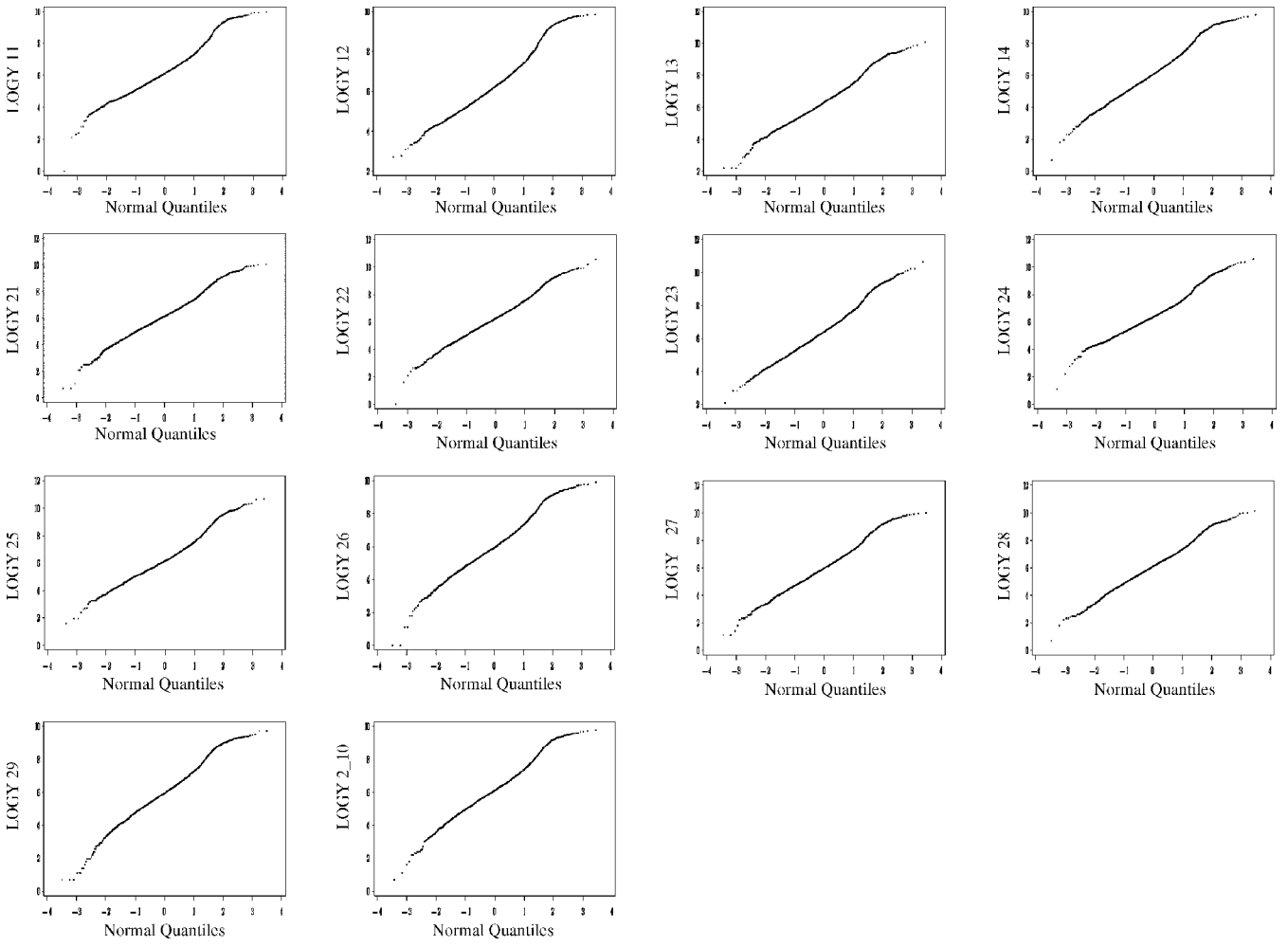


*Figure 1. Normal Probability Plots of log* $\mathbf{y}_{ji} = (\log(y_{jig} - 20): y_{ijg} > 20)'$ *for* $j = 1, 2$ *and* $i = 1, 2, \ldots, n_j$.

Figure 2 thus demonstrates that the observed data are quite consistent with the posterior estimates of $p_{jg}$ obtained from the fitted full model, thus validating our choice of the functional form of the model. Figure 3 shows a boxplot of

$$\Delta_\xi = \log\left(\frac{(1/n_2)\sum_{i=1}^{n_2} y_{2ig}}{(1/n_1)\sum_{i=1}^{n_1} y_{1ig}}\right) - E(\log(\xi_g)|D)$$

for $g = 1, 2, \ldots, G$, under the informative, moderate, and vague priors used for Figure 2. Again in this boxplot we observe a great deal of consistency between the posterior estimates and the observed data. For each case in Figure 3, we see that the boxplot is concentrated near 0, with the dispersion increasing as the prior becomes more informative. Figure 3 again demonstrates that the observed data are quite consistent with the posterior estimates of $\xi_g$ obtained from fitting the full model, thus validating our choice of a model.

We have also examined the impact of certain observations on our model and model robustness in general. We note that if $m_{j0}$ has a moderate to vague prior, then the model is quite robust, and the impact of the omission of a single observation is in fact minimal. However, if $m_{j0}$ has a sharp prior, then the model becomes less robust, and certain observations can greatly influence the model. Thus if we borrow too much strength (sharp prior for $m_{j0}$), then the model becomes less robust, and certain observations can be quite influential. However, if we borrow low to moderate strength, then the model is quite robust. This seems to be a nice property of our model. A similar but opposite phenomenon occurs with $e_{jg}$. If the prior on $e_{jg}$ is too vague, then the model becomes less robust. However, a moderate to sharp prior for $e_{jg}$ improves the robustness of the model, and observations are less influential with moderate to sharp priors for $e_{jg}$. Thus for $e_{jg}$, if we do not borrow enough strength, then the model becomes less robust.

## 8. DISCUSSION

We have developed a new parametric Bayesian model for analyzing DNA microarray data. This model along with the L measure allow us to formally assess differences in genes for the two tissue types and make inferences about the magnitude of the differences. We have proposed classes of hierarchical priors that allow us to build a correlation structure between the genes. In addition, we have proposed a novel gene selection algorithm for producing various models that then can be formally assessed with the L measure. Thus the gene selection algorithm coupled with the L measure provide us with new tools for analyzing gene expression data. Our proposed methodology provides an attractive approach to these types

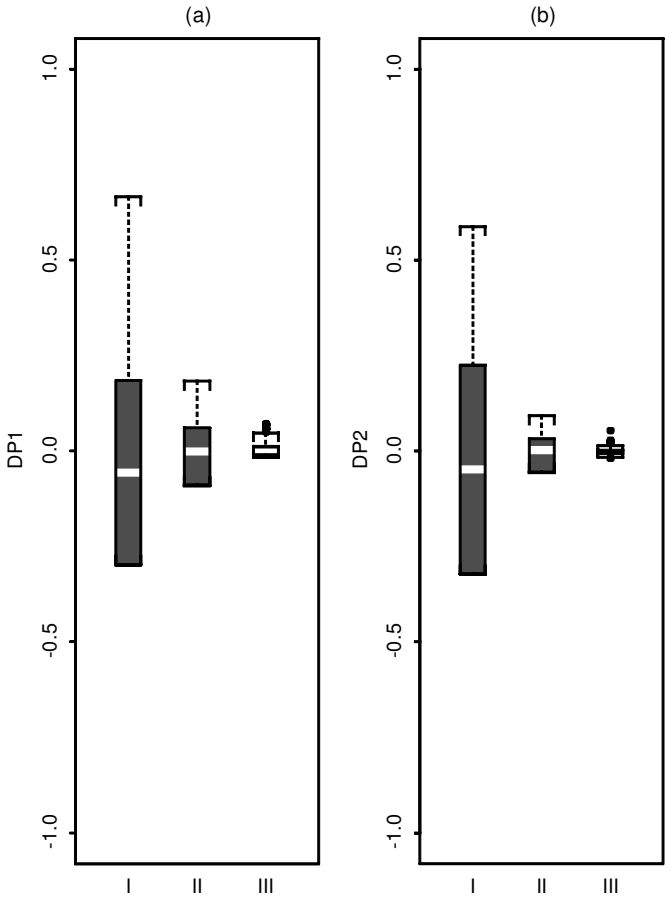


*Figure 2. Boxplots for $\Delta_{p1}$ and $\Delta_{p2}$ Using Informative $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (.01,.01,.01,.01)$ Priors, Moderate $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1,1,1,1)$ Priors, and Vague $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (100,100,50,50)$ Priors.*

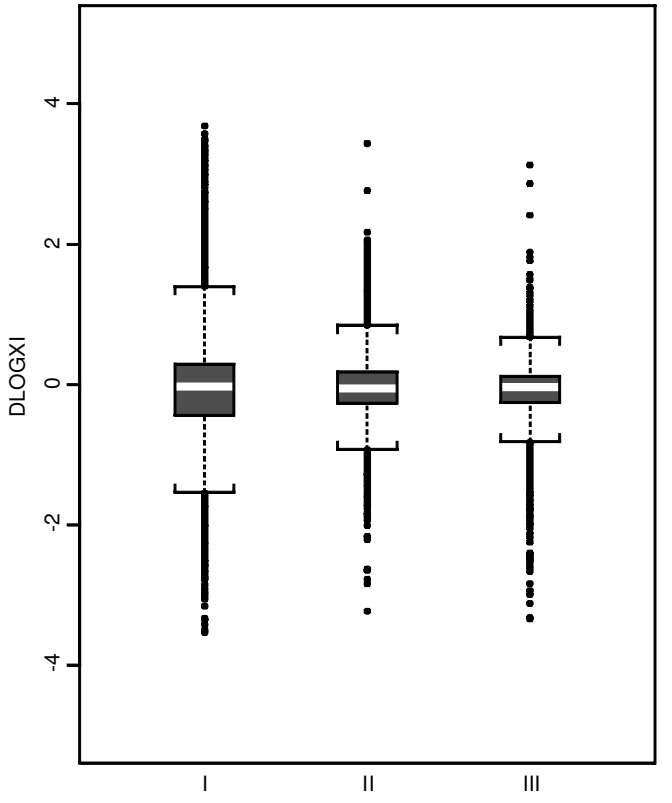


*Figure 3. Boxplots for $\Delta_\xi$ Using Informative $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (.01,.01,.01,.01)$ Priors, Moderate $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (1,1,1,1)$ Priors, and Vague $(\boldsymbol{\eta}_0, \mathbf{d}_0, \mathbf{k}_0, \mathbf{h}_0) = (100,100,50,50)$ Priors.*

of problems. Software to fit our proposed model is available from the authors upon request.

Our approach is also quite different from the clustering algorithms proposed in the literature, because it provides us with much more flexibility for making inferences about the gene differences and characterizing the gene expression patterns in the two tissue types. Here we note that our model development and prior distributions can be easily extended to handle three or more tissue types. We have focused on two tissue types mainly for clarity of exposition and also because this is perhaps one of the most common settings in cancer.

Here we mention that analyses of gene expression data are sometimes carried out by completely ignoring the absolute call. Doing this here results in a completely different dataset that is not at all comparable to a dataset that incorporates the absolute call into the analysis. We will consider such analyses in future work. In addition, more research is needed to develop more general classes of priors that borrow strength within and across tissue types. Such classes of priors are currently being investigated. Finally, further investigation of the gene selection criterion will be carried out; in particular, we will study other criteria related to $P(\xi > 1|G)$ that may be suitable for gene selection. The promising results given here will lead to fruitful research in this direction.

## APPENDIX: COMPUTATIONAL ALGORITHMS

We use the Gibbs sampling algorithm to sample from the joint posterior distribution of the parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$, $\mathbf{b}_0 = (b_{10}, b_{20})$, $\boldsymbol{\mu}_0 = (\mu_{10}, \mu_{20})$, $\mathbf{e} = (e_{11}, e_{21}, \ldots, e_{1G}, e_{2G})$, and $\mathbf{u}_0 = (u_{10}, u_{20})$. Although the joint posterior distribution of the model parameters does not have a closed form, the conditional posteriors have a closed form or the conditional posterior densities are log-concave, and thus implementation of the Gibbs sampler is straightforward. Let $D$ denote the data, and let rest denote the remaining parameters. Then the following apply:

1. If $\delta_{jg} = \sum_{i=1}^{n_j} \delta_{jig} = n_j$, then
$$[\mu_{jg} \mid D, \text{rest}] \sim N(\mu_{j0}, \tau_0 \sigma_{jg}^2 / \bar{n}_j),$$
where $\bar{n}_j$ is defined by (7), and if $\delta_{jg} < n_j$,
$$[\mu_{jg} \mid D, \text{rest}] \sim N(\mu_{jg}^*, \sigma_{jg}^{*2}),$$
where $\sigma_{jg}^{*2} = \{\bar{n}_j/(\tau_0\sigma_{jg}^2) + (n_j - \delta_{jg})/\sigma_{jg}^2\}^{-1}$ and
$$\mu_{jg}^* = \sigma_{jg}^{2*}\left\{\sum_{i=1}^{n_j}(1-\delta_{jig})\log(y_{jig})/\sigma_{jg}^2 + \mu_{j0}\bar{n}_j/(\tau_0\sigma_{jg}^2)\right\}$$
for $j = 1, 2$ and $g = 1, 2, \ldots, G$.
2. $[\sigma_{jg}^2 \mid D, \text{rest}] \sim \text{IG}(a_{jg}^*, b_{jg}^*)$, where
$$a_{jg}^* = \frac{1}{2}(n_j - \delta_{jg} + 1) + a_{j0}$$
and
$$b_{jg}^* = \frac{1}{2}\left\{\sum_{i=1}^{n_j}(1-\delta_{jig})(\log(y_{jig}) - \mu_{jg})^2 + \bar{n}_j(\mu_{jg} - \mu_{j0})^2/\tau_0\right\} + b_{j0}$$
for $j = 1, 2$ and $g = 1, 2, \ldots, G$.

3. $[\mu_{j0} \mid D, \text{rest}] \sim N(\mu^*_{j0}, \sigma^{*2}_{j0})$, where

$$\sigma^{*2}_{j0} = \left\{ 1/v^2_{j0} + \sum_{g=1}^{G} \bar{n}_j/(\tau_0 \sigma^2_{jg}) \right\}^{-1}$$

and

$$\mu^*_{j0} = \sigma^{*2}_{j0} \left\{ \sum_{g=1}^{G} \bar{n}_j \mu_{jg}/(\tau_0 \sigma^2_{jg}) + m_{j0}/v^2_{j0} \right\}$$

for $j = 1, 2$.

4. $[b_{j0} \mid D, \text{rest}] \sim \gamma(q^*_{j0}, t^*_{j0})$, where $q^*_{j0} = Ga_{j0} + q_{j0}$ and $t^*_{j0} = \sum_{g=1}^{G} 1/\sigma^2_{jg} + t_{j0}$ for $j = 1, 2$.

5. For $j = 1, 2$ and $g = 1, 2, \ldots, G$, the logarithm of the conditional posterior density is of the form

$$\log p(e_{jg} \mid D, \text{rest}) \propto \delta_{jg} e_{jg} - n_j \log(1 + \exp(e_{jg})) - (e_{jg} - u_{j0})^2/(2k_{j0} w^2_{j0}),$$

and it can be shown that $p(e_{jg} \mid D, \text{rest})$ is log-concave.

6. $[u_{j0} \mid D, \text{rest}] \sim N(u^*_{j0}, w^{*2}_{j0})$, where

$$w^{*2}_{j0} = \{1/(h_{j0} w^2_{j0}) + G/(k_{j0} w^2_{j0})\}^{-1}$$

and

$$u^*_{j0} = w^{*2}_{j0} \left\{ \sum_{g=1}^{G} e_{jg}/(k_{j0} w^2_{j0}) + \hat{u}_{j0}/(h_{j0} w^2_{j0}) \right\}$$

for $j = 1, 2$.

Thus, for 1–4 and 6, the generation is straightforward, whereas for 5, we can use an adaptive rejection algorithm of Gilks and Wild (1992), because the corresponding conditional posterior densities are log-concave.

*[Received May 2000. Revised August 2001.]*

## REFERENCES


Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997), "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Journal of Biomedical Optics*, 4, 364–374.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2001), "Microarrays and Their Use in a Comparative Experiment," submitted.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868.

Geisser, S. (1993), *Predictive Inference: An Introduction*, London: Chapman and Hall.

Gelfand, A., and Ghosh, S. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, l–13.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P., Ross, D., Scherf, U., Weinstein, J., Alizadeh, A., Staudt, L., and Botstein, D. (2000), "Gene Shaving: A New Class of Clustering Methods for Expression Arrays," submitted.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), "Criterion Based Methods for Bayesian Model Assessment," *Statistica Sinica*, 11, 419–443.

Ibrahim, J. G., and Laud, P. W. (1994), "A Predictive Approach to the Analysis of Designed Experiments," *Journal of the American Statistical Association*, 89, 309–319.

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 57, 247–262.

Mutter, G. L., Baak, J. P. A., Fitzgerald, J. T., Gray, R., Neuberg, D., Kust, G. A., Gentleman, R., Gallans, S. R., Wei, L. I., Wilcox, M. (2001) Global express changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation. *Gynecologic Oncology*, in press.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. B., Botstein, D., and Futcher, B. (1998), "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell*, 9, 3273–3297.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., and Dmitrovsky, E. (1999), "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences*, 96, 2907–2912.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., and Brown, P. (2000), "Clustering Methods for the Analysis of DNA Microarray Data," submitted.

West, M, Nevins, J. R., Marks, J. R., Spang, R., and Zuzman, H. (2000), "Bayesian Regression Analysis in the 'Large p, Small n' Paradigm With Application in DNA Microarray Studies," submitted.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with $g$-Prior Distributions," In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243.