Published in final edited form as:
J Stat Plan Inference. 2008 February 1; 138(2): 387–404. doi:10.1016/j.jspi.2007.06.007.

A new class of mixture models for differential gene expression in DNA microarray data

Ming-Hui Chen[a], **Joseph G. Ibrahim**[b,*], and **Yueh-Yun Chi**[c]

[a]Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

[b]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

[c]Department of Biostatistics, University of Washington, Seattle, WA 98101, USA

Abstract

One of the fundamental issues in analyzing microarray data is to determine which genes are expressed and which ones are not for a given group of subjects. In datasets where many genes are expressed and many are not expressed (i.e., underexpressed), a bimodal distribution for the gene expression levels often results, where one mode of the distribution represents the expressed genes and the other mode represents the underexpressed genes. To model this bimodality, we propose a new class of mixture models that utilize a random threshold value for accommodating bimodality in the gene expression distribution. Theoretical properties of the proposed model are carefully examined. We use this new model to examine the problem of differential gene expression between two groups of subjects, develop prior distributions, and derive a new criterion for determining which genes are differentially expressed between the two groups. Prior elicitation is carried out using empirical Bayes methodology in order to estimate the threshold value as well as elicit the hyperparameters for the two component mixture model. The new gene selection criterion is demonstrated via several simulations to have excellent false positive rate and false negative rate properties. A gastric cancer dataset is used to motivate and illustrate the proposed methodology.

Keywords

Bayesian inference; Empirical Bayes; Gene selection criteria; Prior elicitation; Random threshold mixture model; Simulation study

1. Introduction

Statistical methodologies for differential gene expression have been rapidly developing in recent years, both from a frequentist and Bayesian framework. Recent frequentist methods include Tusher et al. (2001), Storey and Tibshirani (2003), Kerr et al. (2000), Dudoit et al. (2002), Lonnstedt and Speed (2002), Olshen and Jain (2002), Chen et al. (1997), and Lee et al. (2002). Bayesian approaches include Efron et al. (2001), Baldi and Long (2001), Ibrahim et al. (2002), Parmigiani et al. (2002), Newton et al. (2001, 2004),Newton and Kendziorski (2003),West (2003), Ishwaran and Rao (2003, 2005), Tadesse et al. (2003), Mueller et al. (2004), Liu et al. (2004), Do et al. (2005), and Hein et al. (2005). An excellent review article on statistical methods in genomics is Sebastiani et al. (2003). An recent edited book on the analysis of microarray data is Parmigiani et al. (2003). We refer the reader to the book and the

*Corresponding author. Tel.: +1 919 843 2715. *E-mail address:* ibrahim@bios.unc.edu (J.G. Ibrahim).

review article for more detailed discussions on various methodologies and additional references.

One of the fundamental issues in analyzing microarray data is trying to determine which genes are expressed and which ones are not, and in particular, determining a threshold value for which any expression level above the threshold value will be deemed as expressed and any expression value below the threshold value would be deemed not expressed, hence underexpressed. In datasets where many genes are expressed and many are not expressed, a bimodal distribution for the gene expression levels often results. Two component mixture distributions can be quite useful for this type of modeling problem. A related problem, which can be viewed as a generalization of the bimodal problem, is to also model genes that are under expressed, expressed, and over expressed, leading to a three component mixture. Bayesian model-based methods for DNA microarray analysis are now becoming quite popular since complex models can be fit in a relatively straightforward fashion and Bayesian hierarchical models can be especially useful for this type of problem.

To motivate the proposed modeling, we consider a cDNA dataset in gastric cancer published in Chen et al. (2003). Gastric cancer, which is a form of stomach cancer, is the second most common cause of cancer death worldwide (Parkin et al., 1999). The dataset contains 90 tumor samples and 22 normal samples. A total of 6688 genes were available for analysis. The goal for these data is to determine which genes are differentially expressed between the two groups. An exploratory analysis of these data shows that for each group, the distribution of gene expression appears to be bimodal. For example, Fig. 1 shows nonparametric density estimates for nine selected genes in the tumor sample group. The horizontal axis in Fig. 1 corresponds to the logarithm of the red to green channel, log(*R*/*G*), which is the measure of gene expression. The vertical axis corresponds to the density value. Each plot represents a gene, and the nonparametric density estimate for each gene is based on the $n_1 = 90$ tumor samples. We see from Fig. 1 the apparent bimodality in the gene expression distribution. This bimodality may be due to the fact that certain genes are expressed for certain subjects and not expressed for other subjects, thereby creating the bimodality.

To model this bimodality, a threshold value can be defined such that all gene expression levels above this threshold value are deemed expressed and all expression values below the threshold are classified as underexpressed. One of the big challenges in this approach is how to determine the threshold value and whether the threshold value should be treated as fixed or random. Towards these goals, we develop a new class of mixture models that utilize a random threshold value for accommodating bimodality in the gene expression distribution, such as that encountered in Fig. 1. The model is then used to determine which genes are differentially expressed between the two groups (tumor vs. normal). The random threshold value can be viewed as a latent variable in the modeling process, in which a novel distribution is specified for it. Then the joint posterior distribution of the parameters and the threshold value is used for inference. Specification of this threshold mixture model has several advantages over a standard two component mixture model, as discussed in Sections 2 and 3. One of its greatest advantages is that it leads to an identifiable two component mixture model and it facilitates a straightforward prior elicitation scheme via empirical Bayes methodology. Prior elicitation using empirical Bayes methods (see Ibrahim et al., 2002;Efron et al., 2001) is now widely recognized as critical in parametric modeling of DNA microarray data since such models are highly parameterized and conventional prior elicitation strategies using noninformative or improper priors lead to either weakly or nonidentifiable models as well as computational instability. These issues are elaborated upon in Sections 2 and 3. The proposed methodology, in some sense, generalizes previous work by Ibrahim et al. (2002) in that (i) the threshold parameter is assumed unknown and random, and a distribution is specified for it using a multiple imputation technique, (ii) a probability model is posited for the underexpressed genes

as well as the expressed genes, and (iii) we allow general classes of distributions for the gene expression data, in which the log-normal, Box–Cox transformations of a normal random variable, and several others are special cases. We mention that other approaches for dealing with low expression levels include left censoring the gene expression data at the truncation value as in Tadesse et al. (2003). However, such methods assume that the threshold value at which to censor is known, and thus are not a general as the methodology considered here.

In addition to the new threshold mixture model, prior distributions for the model parameters are proposed as well as a new criterion for determining which genes are differentially expressed. This new criterion is demonstrated through several simulations to have excellent false positive rate and false negative rate properties. The proposed methodology is also compared to a fully frequentist procedure called PERMAX developed by Mutter et al. (2001), the static significance analysis of microarray (SAM) model, proposed by Tusher et al. (2001), and the parametric empirical bayes methods for microarray data (EBarrays), proposed by Kendziorski et al. (2003).

The rest of this article is organized as follows. In Section 2, we present a new mixture model for modeling expressed and underexpressed genes using a single threshold value. In Section 3, we introduce a class of hierarchical priors for the random threshold parameter as well as the other parameters. Since the threshold value is random, we propose an algorithm for eliciting prior distributions via a standard multiple imputation technique. In Section 4, a new criterion for determining which genes are differentially expressed is developed. In Section 5, we present extensive simulation results illustrating the operating characteristics of the proposed methodology, and in Section 6, we illustrate the methodology on the gastric cancer dataset, and carry out prediction analysis of microarrays (PAM), which is a procedure for cross-validation proposed by Tibshirani et al. (2002).We conclude the article with a brief discussion in Section 7.

2. New threshold mixture model

The proposed model is constructed as follows: Let $j = 1, 2$ index the tissue type (normal vs. tumor) and let y_{jgi} denote the gene expression random variable for the jth tissue type and the gth gene for the ith individual, $i = 1, 2, \ldots, n_{jg}$ and $g = 1,\ldots, G$. Let p_{jg} = probability that the gth gene is not expressed for tissue type j. Here, we do not assume that the raw gene expression levels y_{jgi} follow a particular distribution, but rather Assume that $h(y_{jgi})$ is a known differentiable transformation of y_{jgi} to achieve normality. For example, $h(.)$ could be the Box–Cox class of transformations in which

$$h(x)=\begin{cases} \frac{x^{\gamma}-1}{\gamma}, & \gamma \neq 0, \\ \log(x), & \gamma=0. \end{cases}$$

$h(.)$ can also represent other classes of parametric transformations that achieve normality. Consider the model,

$$p(y_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg})=p_{jg}\ p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg})+(1-p_{jg})\ p_2(y_{jgi}|\mu_{jg},\sigma^2_{jg}), \tag{2.1}$$

where

$$p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg})=(2\pi)^{-1/2}|h\prime(y_{jgi})|\tau^{-1}_{jg}\ \exp\left\{-\frac{\mathbf{1}}{\mathbf{2\tau^2_{jg}}}(\mathbf{h(y_{jgi})}-\alpha_{\mathbf{jg}})^{\mathbf{2}}\right\}$$

and

$$p_2(y_{jgi}|\mu_{jg},\sigma^2_{jg})=(2\pi)^{-1/2}|h\prime(y_{jgi})|\sigma^{-1}_{jg}\exp\left\{-\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}_{\mathbf{jg}}}(\mathbf{h}(\mathbf{y_{jgi}})-\mu_{\mathbf{jg}})^{\mathbf{2}}\right\}.$$

Note that in (2.1), $p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg})$ and $p_2(y_{jgi}|\mu_{jg},\sigma^2_{jg})$ are the distributions for y_{jgi} for the underexpressed and expressed genes, respectively.

One of the potential disadvantages with (2.1) is that it is virtually impossible to develop data-based prior specifications for $p_{jg},\alpha_{jg},\tau^2_{jg},\mu_{jg},$ and σ^2_{jg} since one does not know in advance which group each gene expression level belongs to. In order to solve this dilemma, we can construct a cut-off value, or *threshold*, in the prior elicitation strategy so that all gene expression levels below a certain threshold belong to one group and all gene expression levels above this threshold belong to the other group. Once a threshold value is defined, empirical Bayes prior elicitation would proceed in a straightforward fashion.

In our model development, we wish to introduce a threshold while at the same time retaining the two component structure of (2.1). Towards this goal, we consider an alternative but equivalent model of (2.1). Let c_{jgi} denote a *random threshold parameter* such that the gene is not expressed if $y_{jgi}\leq c_{jgi}$ for the j th tissue type, the ith individual, and the gth gene. Assume that the c_{jgi} 's are i.i.d. with distribution

$$p(c_{jgi}|c_{0jg},\varsigma^2_{jg})=(2\pi)^{-1/2}|h\prime(c_{jgi})|\varsigma^{-1}_{jg}\exp\left\{-\frac{\mathbf{1}}{\mathbf{2}\varsigma^{\mathbf{2}}_{\mathbf{jg}}}(\mathbf{h}(\mathbf{c_{jgi}})-\mathbf{h}(\mathbf{c_{0jg}}))^{\mathbf{2}}\right\}. \tag{2.2}$$

Let $A_{jgi}=\{y: y\leq c_{jgi}\}$ and A^c_{jgi} denotes the complement of A_{jgi}. We consider the following joint distribution for (y_{jgi}, c_{jgi}):

$$\begin{aligned}&p\,(y_{jgi},c_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg},c_{0jg},\varsigma^2_{jg})\\&=[\,p_{jg}\,p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg})1_{A_{jgi}}(y_{jgi})/q_{y_{jgi}}\\&\quad+(1-p_{jg})\,p_2(y_{jgi}|\mu_{jg},\sigma^2_{jg})1_{A^c_{jgi}}(y_{jgi})/(1-q_{y_{jgi}})]\;p(c_{jgi}|c_{0jgi},\varsigma^2_{jg}),\end{aligned} \tag{2.3}$$

where $q_{y_{jgi}}=\int_{y_{jgi}}^{\infty}p(c_{jgi}|c_{0jg},\varsigma^2_{jg})\mathrm{d}c_{jgi}$. Now we are led to a useful identity which relates (2.3) to (2.1).

Identity 2.1

The marginal distribution of (2.3) for y_{jgi} reduces to the distribution, $p(y_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg})$ given in (2.1). That is,

$$\int_0^{\infty}p\,(y_{jgi},c_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg},c_{0jg},\varsigma^2_{jg})\,\mathrm{d}c_{jgi}=p(y_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg}).$$

Proof—Given y_{jgi}, we have

$$\begin{aligned}&\int_0^\infty p\,(y_{jgi}, c_{jgi}|\alpha_{jg}, \tau^2_{jg}, \mu_{jg}, \sigma^2_{jg}, c_{0jg}, \varsigma^2_{jg})\mathrm{d}c_{jgi}\\&=\int_{y_{jgi}}^\infty [\,p_{jg}\; p_1(y_{jgi}|\alpha_{jg}, \tau^2_{jg})/q_{y_{jgi}}]\; p\,(c_{jgi}|c_{0jg}, \varsigma^2_{jg})\,\mathrm{d}c_{jgi}\\&\quad+\int_0^{y_{jgi}} [\,(1-p_{jg})\; p_2(y_{jgi}|\mu_{jg}, \sigma^2_{jg})/(1-q_{y_{jgi}})]\; p\,(c_{jgi}|c_{0jg,}\varsigma^2_{jg})\,\mathrm{d}c_{jgi}\\&=p_{jg}\; p_1(y_{jgi}|\alpha_{jg}, \tau^2_{jg})+(1-p_{jg})\; p_2(y_{jgi}|\mu_{jg}, \sigma^2_{jg}),\end{aligned}$$

which establishes the identity.

The main purpose of (2.3) and (2.1) is that by introducing the latent variable c_{jgi}, we are able to (i) obtain an identifiable model and (ii) construct the same two-component mixture model (2.1) such that once c_{jgi} is given, we then immediately know which genes belong to which group, and this is what facilitates a straightforward data-based prior elicitation scheme. We note that model (2.1) as it stands is not identifiable since the labels of all of the parameters are arbitrary. However, by introducing the threshold parameter c_{jgi}, we induce an ordering in the parameters of the two component mixture model that immediately yields an identifiable model. We mention here that there are alternative approaches to the model development and inference scheme proposed here. One such approach is to deal with the two component mixture model directly and make it identifiable by placing constraints on the means, and then estimate the parameters using the EM algorithm. In this framework, however, estimation of standard errors would be much more difficult than the fully Bayesian approach we adopt here. Identity 2.1 also demonstrates that (2.1) and (2.3) are indeed equivalent.

Let $\delta_{jgi} = 1$ if $y_{jgi} \leq c_{jgi}$ and 0 otherwise. Since c_{jgi} is random, δ_{jgi} is an unobserved latent variable. However, given the value of c_{jgi} and y_{jgi}, the value of δ_{jgi} is completely known. We present another useful identity which relates p_{jg} to the event $\{y_{jgi} \leq c_{jgi}\}$.

Identity 2.2

The probability p_{jg} in (2.1) *is the probability that the gene is not expressed, that is, $y_{jgi} \leq c_{jgi}$ under the mixture distribution* (2.3). *Specifically, we have*

$$P(\delta_{jgi}=1)=P(y_{jgi} \leq c_{jgi})=p_{jg} \quad \textit{and} \quad P(\delta_{jgi}=0)=P(y_{jgi}>c_{jgi})=1-p_{jg}. \tag{2.4}$$

Proof—It is sufficient to show $P(\delta_{jgi} = 1) = P\,(y_{jgi} \leq c_{jgi}) = p_{jg}$. Based on the definition of the joint distribution of (y_{jgi}, c_{jgi}), we obtain

$$\begin{aligned}P\,(y_{jgi} \leq c_{jgi}) &=\int_{y_{jgi}\leq c_{jgi}} p(y_{jgi}, c_{jgi}|\alpha_{jg}, \tau^2_{jg}, \mu_{jg}, \sigma^2_{jg}, c_{0jg}, \varsigma^2_{jg})\,\mathrm{d}c_{jgi}\,\mathrm{d}y_{jgi}\\&=\int_0^\infty\int_{y_{jgi}}^\infty [\,p_{jg}\; p_1(y_{jgi}|\alpha_{jg,}\; \tau^2_{jg})/q_{y_{jgi}}]p(c_{jgi}|c_{0jg}, \varsigma^2_{jg})\,\mathrm{d}c_{jgi}\mathrm{d}y_{jgi}\\&=\int_0^\infty p_{jg}p_1(y_{jgi}|\alpha_{jg}, \tau^2_{jg})\,\mathrm{d}y_{jgi}=p_{jg,}\end{aligned}$$

which establishes the identity.

Identity 2.2 shows the relationship between p_{jg} and the threshold parameter c_{jgi}, and thus we see that p_{jg} simply corresponds to the probability that the gene expression level is below the threshold in (2.3).

We now construct the likelihood as follows. Let $\delta=(\delta_{\ 111},\ldots, \delta_{2,G,n_{2G}})$, $\boldsymbol{c}= (c_{111},\ldots, c_{2,G,n_{2G}})$, $\boldsymbol{c}_0=(c_{011}, c_{021},\ldots,c_{01G}, c_{02G})$, $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{21},\ldots, \alpha_{1G}, \alpha_{2G})$
$\boldsymbol{\tau}^2=(\tau^2_{11}, \tau^2_{21},\ldots, \tau^2_{1G}, \tau^2_{2G}), \mu=(\mu_{11}, \mu_{21},\ldots, \mu_{1G}, \mu_{2G}), \boldsymbol{\sigma}^2=(\sigma^2_{11}, \sigma^2_{21},\ldots, \sigma^2_{1G}, \sigma^2_{2G}),$

$\boldsymbol{\varsigma}^2=(\varsigma_{11}^2,\varsigma_{21}^2,\ldots,\varsigma_{1G}^2,\varsigma_{2G}^2)$, $\boldsymbol{p}=(p_{11},p_{21},\ldots,p_{1G},p_{2G})$, and $\boldsymbol{\theta}=(\boldsymbol{\alpha},\boldsymbol{\tau}^2,\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{c}_0,\varsigma^2,\boldsymbol{p})$. Let $\boldsymbol{D}=(y_{111},\ldots,y_{2,G,n_{2G}},\boldsymbol{c})$ denote the complete data and $\boldsymbol{D}_{obs}=(y_{111},\ldots,y_{2,G,n_{2G}})$ denote the observed data.

The likelihood function for θ based on the *complete data* $D=(y_{111},\ldots,y_{2,G,n_{2G}},\boldsymbol{c})$ is thus given by

$$L(\theta|D) = \prod_{j=1}^{2}\prod_{g=1}^{G}\prod_{i=1}^{n_{jg}}\{[p_{jg}/q_{y_{jgi}}]^{\delta_{jgi}}[(1-p_{jg})/(1-q_{y_{jgi}})]^{1-\delta_{jgi}}p_1(y_{jgi}|\alpha_{jg},\tau_{jg}^2)^{\delta_{jgi}} \times p_2(y_{jgi}|\mu_{jg},\sigma_{jg}^2)^{1-\delta_{jgi}}\}. \quad (2.5)$$

An interesting special case of the general model (2.5) is obtained by taking $c_{jgi}=c_{jg}$, that is to let the random threshold parameter be free of the sample (subject). This special case is attractive, since (i) the y_{jgi}'s share the same random threshold parameter c_{jg} for the same tissue type and the same gene, and (ii) the y_{jgi}'s are correlated across the same tissue type and the same gene.

3. Prior specifications

Following Ibrahim et al. (2002), the empirical Bayes methodology is carried out by specifying a data-based *guide value* for all of the hyperparameters of the priors. We first elicit the parameters $(c_{0jg},\varsigma_{jg}^2)$ given in (2.2). The guide values for c_{0jg} and ς_{jg}^2 are $h(c_{0jg})=\frac{1}{n_{jg}}\sum_{i=1}^{n_{jg}}h(y_{jgi})$ and $\varsigma_{jg}^2=\frac{\kappa_0}{G-1}\sum_{g=1}^{G}(h(c_{0jg})-\bar{h}(c_{0j}))^2$, where $\bar{h}(c_{0j})=\frac{1}{G}\sum_{g=1}^{G}h(c_{0jg})$ and κ_0 is a fixed parameter. A default choice of κ_0 is 1. Here, c_{jgi} is an unobserved latent variable. Using multiple imputation, we independently generate c_{jgi}^b from (2.2) for $b=1,2,\ldots,B$. For each b, let $\delta_{jgi}^{*b}=1$ if $y_{jgi}\le c_{jgi}^b$ and 0 otherwise, $\delta_{jgi1}^{*b}=\delta_{jgi}^{*b}$ and $\delta_{jgi2}^{*b}=1-\delta_{jgi}^{*b}$ for $b=1,2,\ldots,B$. We then take

$$\bar{n}_{jk}=\frac{1}{BG}\sum_{b=1}^{B}\sum_{g=1}^{G}\sum_{i=1}^{n_{jg}}\delta_{jgik}^{*b}$$

for $k=1,2$.

We follow the same ideas as in Ibrahim et al. (2002) for specifying priors for the rest of the parameters. For α_{jg}, we take

$$\alpha_{jg}|\tau_{jg}^2,\alpha_{j0}\sim N(\alpha_{j0},\tau_0\tau_{jg}^2/\bar{n}_{j1}),$$

where $\tau_0>0$ is a specified scalar, and

$$\tau_{jg}^2\sim\mathcal{IG}(a_{j01},b_{j01}),$$

where (a_{j01},b_{j01}) are hyperparameters, for $j=1,2$. Similarly, we take

$$\mu_{jg}|\sigma^2_{jg},\mu_{j0} \sim \mathrm{N}(\mu_{j0}, \sigma_0\sigma^2_{jg}/\bar{n}_{j2}),$$

where $\sigma_0 > 0$ is a specified scalar and

$$\sigma^2_{jg} \sim \mathcal{IG}(a_{j02}, b_{j02}),$$

where (a_{j02}, b_{j02}) are hyperparameters, for $j = 1, 2$. We further take $\alpha_{j0} \sim \mathrm{N}(m_{j01}, v^2_{j01}), \mu_{j0} \sim \mathrm{N}(m_{j02}, v^2_{j02})$, $j = 1, 2$, and we take a_{j0k} fixed and b_{j0k} random for our hierarchical prior. Specifically, we take a gamma prior for b_{j0k}, i.e., $b_{j0k} \sim \mathcal{G}(q_{j0k}, t_{j0k})$, where (q_{j0k}, t_{j0k}) are specified hyperparameters for $k = 1, 2$ and $j = 1, 2$.

For the p_{jg}'s, we specify the prior as follows. We first let

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

and then specify a normal prior on the e_{jg}'s, therefore inducing a prior on the p_{jg}'s. Thus, we take

$$e_{jg} \sim \mathrm{N}(u_{j0}, k_{j0}w^2_{j0}), \quad j=1,2$$

and for the prior for e_{jg}, we take

$$u_{j0} \sim \mathrm{N}(\widehat{u}_{j0}, h_{j0}w^2_{j0}), \quad j=1,2.$$

The hyperparameters $k_0 = (k_{10}, k_{20})$, $h_0 = (h_{10}, h_{20})$, and w^2_{j0}, $j = 1, 2$, are prespecified.

The guide values for all the hyperparameters are specified as follows. For m_{j0k}, we take

$$m_{j0k}=\frac{1}{BN_{jk}}\sum_{b=1}^{B}\sum_{g=1}^{G}\sum_{i=1}^{n_{jg}}\delta^{*b}_{jgik}h(y_{jgi}),$$

where $N_{jk}=\frac{1}{B}\sum_{b=1}^{B}\sum_{g=1}^{G}\sum_{i=1}^{n_{jg}}\delta^{*b}_{jgik}$, for $k = 1, 2$ and $j = 1, 2$. For υ^2_{j0k} we take

$$\upsilon^2_{j0k}=\eta_{j0k}\mathrm{MSG}_{jk},$$

where

$$\mathrm{MSG}_{jk}=\frac{1}{G-1}\sum_{g=1}^{G}n_{jgk}\,(m_{jg0k}-m_{j0k})^2,$$

$$m_{jg0k}=\frac{\sum_{b=1}^{B}\sum_{i=1}^{n_{jg}}\delta_{jgik}^{*b}h\,(y_{jgi})}{\sum_{b=1}^{B}\sum_{i=1}^{n_{jg}}\delta_{jgik}^{*b}},$$

$n_{jgk}=\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n_{jg}}\delta_{jgik}^{*b}$, for $k = 1, 2$ and $j = 1, 2$, and $\eta_0 = (\eta_{101}, \eta_{201}, \eta_{102}, \eta_{202})$ is a vector of chosen scalers. A guide value for t_{j0k} is $t_{j0k}^{-1}=d_{j0k}\mathrm{MSE}_{jk}$ where

$$\mathrm{MSE}_{jk}=\frac{1}{B(N_{jk}-G)}\sum_{b=1}^{B}\sum_{g=1}^{G}\sum_{i=1}^{n_{jg}}\delta_{jgik}^{*b}(h(y_{jgi})-m_{jg0k})^2$$

for $k = 1, 2$ and $j = 1, 2$, and $d_0 = (d_{101}, d_{201}, d_{102}, d_{202})$ is a vector of chosen scalars. We see that MSE_{jk} is just the mean square error for the expressed or underexpressed gene expression levels for tissue type j. Finally, we elicit the guide values based on the sample proportion of underexpressed gene expression values for $\hat{u}_{j0}$ and w_{j0}^2. For $\hat{u}_{j0}$, we propose a guide value of

$$\widehat{u}_{j0}=\log\left[\frac{1}{BG}\sum_{b=1}^{B}\sum_{g=1}^{G}\widehat{p}_{jg}^{b}\Big/\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{BG}}\sum_{\mathbf{b=1}}^{\mathbf{B}}\sum_{\mathbf{g=1}}^{\mathbf{G}}\widehat{\mathbf{p}}_{\mathbf{jg}}\right)\right],$$

where $\hat{p}_{jg}$ is the sample proportion of underexpressed gene expression values over all of the individuals for the jth tumor type in the bth imputed sample. This guide value for $\hat{u}_{j0}$ seems quite suitable based on the definition of e_{jg}. Finally, for w_{j0}^2, we take a guide value of the form

$$w_{j0}^2=\left\{\left(\frac{\mathbf{1}}{\mathbf{BG}}\sum_{\mathbf{b=1}}^{\mathbf{B}}\sum_{\mathbf{g=1}}^{\mathbf{G}}\widehat{\mathbf{p}}_{\mathbf{jg}}^{\mathbf{b}}\right)\left(\mathbf{1}-\frac{\mathbf{1}}{\mathbf{BG}}\sum_{\mathbf{b=1}}^{\mathbf{B}}\sum_{\mathbf{g=1}}^{\mathbf{G}}\widehat{\mathbf{p}}_{\mathbf{jg}}^{\mathbf{b}}\right)\right\}^{-1}.$$

Thus we see that the guide value for w_{j0}^2 is just the frequentist variance of $\frac{1}{BG}\sum_{b=1}^{B}\sum_{g=1}^{G}\widehat{p}_{jg}^{b}$.

To gain a better understanding of the prior distributions and their associated hyperparameters, a directed acyclic graph (DAG) of the prior elicitation scheme is given in Fig. 2.

4. Gene selection criteria

To discriminate between the normal and tumor tissues, we follow Ibrahim et al. (2002) and let

$$\psi_{jg}=E_y[y_{jgi}|p_{jg},\alpha_{jg},\tau_{jg}^2,\mu_{jg},\sigma_{jg}^2],$$

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

$$\psi_{jg}=E[y_{jgi}|\alpha_{jg},\tau^2_{jg}]\,p_{jg}+(1-p_{jg})\,E[y_{jgi}|\mu_{jg},\sigma^2_{jg}]. \tag{4.1}$$

If $h(y_{jgi}) = \log(y_{jgi})$, then

$$\psi_{jg}=p_{jg}\exp\left\{\alpha_{\mathbf{jg}}+\frac{\tau^{\mathbf{2}}_{\mathbf{jg}}}{\mathbf{2}}\right\}+(1-p_{jg})\exp\left\{\mu_{\mathbf{jg}}+\frac{\sigma^{\mathbf{2}}_{\mathbf{jg}}}{\mathbf{2}}\right\}. \tag{4.2}$$

The primary reason why we focus on $\psi_{jg}=E_y[y_{jgi}|p_{jg},\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg}]$ as a gene selection criterion is that this quantity provides combined information on both the location and scale parameters in the specified model for y_{jgi}. In contrast, if for example, y_{jgi} has a log-normal distribution and we consider the expected value of $\log(y_{jgi})$ as the gene selection criterion, we can immediately see that this expectation is simply the weighted average of the location parameters. Thus, the expected value of $\log(y_{jgi})$ is not as informative as the expected value of y_{jgi} since it does not use information on both the location and scale parameters.

To compare the gene expression level means between the normal and tumor tissues, we follow Ibrahim et al. (2002) and define

$$\xi_g=\psi_{2g}/\psi_{1g},\quad g=1,2,\ldots,G. \tag{4.3}$$

Then, we propose the following algorithm for determining which genes are differentially expressed between the two groups:

Step 1. We first compute the posterior distributions of all the ξ_g's, $g = 1, 2, \ldots, G$, and for each ξ_g, we compute $\gamma_{g21} = P(\xi_g > 2|D)$ or $\gamma_{g22} = P(\xi_g < 0.5|D)$ (the 2-criterion), as well as $\gamma_{g31} = P(\xi_g > 3|D)$ or $\xi_g 32 = P(\xi_g < 1/3|D)$ (the 3-criterion).

Step 2. We select a cut-off value, denoted γ_0, for γ_{gjk} for determining which genes are different. Possible values of γ_0 might be $\gamma_0 = 0.7$, $\gamma_0 = 0.8$, and $\gamma_0 = 0.9$.

Step 3. We declare gene g different for the two tissue types if $\gamma_{g21}\geq\gamma_0$ or $\gamma_{g22}\geq\gamma_0$ (the 2-criterion), or if $\gamma_{g31}\geq\gamma_0$ or $\gamma_{g32}\geq\gamma_0$ (the 3-criterion).

We note that computing $P(\xi_g > 2|D)$ (or $P(\xi_g > 3|D)$) is quite straightforward since it is a by-product of Markov chain Monte Carlo (MCMC) sampling. Specifically, suppose $\{\xi_g^q, q=1,2,\ldots,Q\}$ is an MCMC sample from the posterior distribution. Then, an Monte Carlo estimate of $P(\xi_g > 2|D)$ is simply

$$\widehat{P}(\xi_g>2|D)=\frac{1}{Q}\sum_{q=1}^{Q}1_{\{\xi_g^q>2\},}$$

where $1_{\{\xi_g^q>2\}}$ is the indicator function. In using (4.3), Ibrahim et al. (2002) only considered the "one-criterion" for determining which genes are differentially expressed, that is, $P(\xi_g>1|D)$. Our experience shows that the 2 and 3-criteria yield much better false positive and false negative rates as opposed to the 1-criterion, and thus we use these for determining which genes are differentially expressed.

5. A simulation study

We conducted a simulation study to investigate the operating characteristics of the threshold mixture model in (2.5) in the context of differential gene expression, and to also compare the performance of the proposed model to frequentist methods for differential gene expression based on Significance Analysis of Microarrays (SAM, Tusher et al., 2001), parametric empirical bayes methods for microarray data (EBarrays, Kendziorski et al., 2003), and permutation methods based on *t* statistics (PERMAX, Mutter et al., 2001). Towards these goals, we simulated data from the log-normal model in (2.1). The simulation assumes two groups, in which $n_1 = n_2 = 25$ and $G = 1000$ genes. The data was simulated so that 50 genes are in truth "differentially expressed" (i.e., the expression levels are simulated from two different log-normal distribution with different location and scale parameters), and 950 genes are in truth "not differentially expressed" (i.e., the gene expression levels are generated from identical log-normal distributions). Specifically, the data was simulated from the log-normal ($h(y) = \log(y)$) mixture model in (2.1) with $p_{jg} = 0.4$, $\alpha_{jg} = 1$, $\tau^2_{jg} = 0.25$, $\mu_{jg} = 4$, and $\sigma^2_{jg} = 2$, $j = 1, 2$, for the 950 "similar" genes, and, $p_{jg} = 0.4$, $\alpha_{1g} = 1$, $\alpha_{2g} = 1.5$, $\tau^2_{jg} = 0.25$, $j = 1, 2$, $\mu_{1g} = 3$, $\mu_{2g} = 7$, and $\sigma^2_{jg} = 2$, $j = 1, 2$, for the 50 "different genes".

Table 1 summarizes the false positive rates (FPR) and false negative rates (FNR) based on three different priors: (I) noninformative with $(\eta_0, d_0, k_0, h_0) = (100, 100, 50, 50)$, (II) moderately informative with $(\eta_0, d_0, k_0, h_0) = (1, 1, 1, 1)$, or (III) informative with $(\eta_0, d_0, k_0, h_0) = (0.01, 0.01, 0.01, 0.01)$. The results shown in Table 1 are based on 500 simulations. We see from Table 1 that the performance of the proposed 2-criterion and 3-criterion is quite good, and appears to behave best with $\gamma_0 \geq 0.80$. For example, for $\gamma_0 = 0.80$ under noninformative priors, the mean FPR is 0.038 and 0.007 and the mean FNR is 0.0003 and 0.001 under the 2-criterion and the 3-criterion, respectively. Moreover, the FPR and FNR are quite robust with respect to the choice of the prior. We see that we get essentially the same rates for all three priors for several different values of γ_0. These results are very encouraging and show that our gene selection algorithm described in Section 4 along with both the 2-criterion and the 3-criterion have good properties.

Table 2 shows the false positive and false negative rates based on the noninformative prior and various combinations of (n_1, n_2). In Table 2, the results are based on the 3-criterion. We also compared our procedure to PERMAX, SAM, and EBarrays. 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, the permutation distribution of the most extreme statistics over all genes is used. Since the distributions of the *t* statistics are not symmetric with unequal group sizes, this is done separately in each tail. Assuming positive values of t_g indicate higher values in normal tissues, and letting $t^{(p)}$ be the maximum statistic over all 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. SAM is now a well known statistical technique for finding significant genes in a set of microarray experiments. It uses repeated permutations of the data to determine if the expression of any genes are significantly related to the response variable (the grouping variable in the context of this paper). EBarrays assumes a hierarchical mixture model to account for differences among genes in their average expression levels, differential expression for a given gene among groups, and measurement fluctuations. Posterior probabilities of patterns of differential expression across groups can be computed and used to determine significant genes.

Table 2 shows the PERMAX procedure based on 50,000 permutations, as well as the SAM model based on 20,000 repeated permutations. We see from Table 2 that our method provides more satisfactory false negative and false positive rate results compared to the other three approaches. For a given γ_0 and criterion (2 or 3-criterion) based on our approach, both the false positive and negative rates increase as the two sample size decreases. Although the PERMAX procedure gives an excellent FPR, the false negative rates based on the PERMAX procedure are extremely high. For example, for $(n_1, n_2) = (25, 25)$ using a significance level of 0.05, the false negative rate is 0.715. The false negative rates based on the PERMAX procedure increase as the sample size decreases. Based on controlling the false discovery rate (FDR), SAM yields reasonably good false positive and false negative rates; however, the false negative rate increases substantially when $(n_1, n_2) = (5, 5)$. Like the PERMAX procedure, the false negative rates increase as the sample size decreases. For the EBarrays results, we assume a log-normal mixture model, and use either 0.5 or 0.7 as the threshold posterior probability to identify genes that are differentially expressed. We can clearly see that under almost all simulation scenarios, EBarrays produces the highest false positive rates. An exception occurs when $(n_1, n_2)=(5, 5)$ and the threshold posterior probability is 0.7; however, in this case, the false negative rate is extremely high.

In addition, we conducted a study of the robustness of the log-normal distribution. We simulated data from a gamma mixture model with all shape parameters equal to 3 and means matching the ones specified for the log-normal models in the simulation design discussed earlier. Again, the data were simulated so that 50 genes are in truth differentially expressed and 950 genes are in truth not differentially expressed. The results shown in Table 3 are based on noninformative priors and the 3-criterion. Our proposed 3-criterion outperforms the PERMAX, SAM and EBarrays procedures. Although SAM gives impressive results as shown in Table 2, it is far less robust with respect to model assumptions compared to our log-normal hierarchical model in (2.5). Overall, these simulation results show that our mixture model (2.5) along with the 3-criterion (or 2-criterion) gene selection algorithm is very promising.

6. Analysis of gastric cancer data

In this section, we revisit the gastric cancer dataset of Chen et al. (2003). The dataset contains 90 tumor and 22 normal examples, and a total of 6688 genes were available for analysis. Here, we carry out an analysis of these data with the proposed model and compare it to the analysis of Chen et al. (2003), which is reported on the website (http://genome-www.stanford.edu/Gastric_Cancer2), and SAM. Table 4 shows the number of selected genes under the 2 and 3 criteria, respectively. As expected, we see that as γ_0 increases, the number of differentially expressed genes becomes smaller, and also the 3-criterion yields fewer differentially expressed genes than the 2-criterion. Table 4 also lists the percentage of differentially expressed genes that matched between our list and Chen et al.'s (2003) list, as well as the number of matches between our list and the list from SAM. We see that our list and Chen et al.'s list matched for at least 80% of the genes for all combinations of γ_0 and type of criterion (2 or 3-criterion), with the highest percentage of matches occurring for $\gamma_0 = 0.90$ along with the 2-criterion. Chen et al. (2003) identified 3329 genes as differentially expressed using a "nonparametric *t*-test" with a *p*-value cutoff of 0.001 or 0.002 based on 10,000 random "column permutations". Further details of this statistical procedure can be found in Troyanskaya et al. (2002). The percentage of matched genes identified as differentially expressed between our method and SAM is at least 96%. Despite this high percentage, the number of genes identified by SAM as differentially expressed is substantially larger than our approach. We also note that the gastric cancer dataset consists of a substantial number of missing gene expression measurements. From the 90 tumor samples, only 8.49% of the genes are completely observed, and up to 35.36% of the genes have more than five missing gene expression measurements. As for the 22 normal samples, only 43.26% of the genes are fully

observed and 7.64% of the genes are missing at least five gene expression measurements. Our proposed mixture model intrinsically allows for unbalanced data, and is capable of handling the missing data properly, whereas the SAM method naively imputes missing values via a *k*-nearest neighbor algorithm. This ad-hoc method of handling of missing data may yield misleading results and makes SAM less appropriate as a tool in analyzing microarray data when there is a significant amount of missing data. We mention here that EBarrays cannot handle missing data. More specifically, in the presence of missing gene expression data, the number of genes for each subject becomes different and EBarrays is not able toaccommodate this setting, and hence not applicable for non-rectangular data. Therefore, EBarrays cannot be applied to the gastric cancer data.

Fig. 3 shows nonparametric density estimates for nine genes for the tumor and normal samples that were deemed differentially expressed by our proposed method as well as Chen et al. (2003). From Fig. 3, we see the clear separation in distributions, thereby correctly identifying the genes as differentially expressed, as well as the apparent bimodality in each distribution. Fig. 4 shows nine genes that were not deemed differentially expressed by Chen et al. (2003) but were deemed differentially expressed by us using either the 2 or 3 criteria or $\gamma_0 \geq .70$. Fig. 4 is striking. Although it appears that there is much overlap between the distributions in each panel, we see that there are great differences in the tails between the two distributions in each panel, and in particular, there is often bimodality in the tails. Chen et al.'s method cannot pick up these differences as this method is primarily aimed at detecting differences in location, and is unable to detect bimodality or large differences in the tails. In contrast, (2.3) is very well suited to pick up these types of differences. The nine genes reported in Fig. 4 are identified by SAM as differentially expressed. Fig. 5 shows a plot of the posterior probability for the 3-criterion (see Step 1 in Section 4) versus the posterior mean of $\log(\xi_g)$ for the 3329 genes that were deemed differentially expressed by Chen et al. (2003). We see from Fig. 5 that many such genes have small posterior probability (less than 0.70) as well as a small $\log(\xi_g)$ according to our proposed criterion, and thus such genes might be inaccurately claimed to be differentially expressed. Similarly, Fig. 6 shows a plot of the posterior probability for the 3-criterion versus the posterior mean of $\log(\xi_g)$ for the 3329 genes that were not deemed differentially expressed by Chen et al. (2003).We see from Fig. 6 that many such genes actually have a large $\log(\xi_g)$ as well as a large posterior probability according to our proposed method, implying that Chen et al. may be inaccurately declaring certain genes as not differentially expressed, when in reality they may be. Again, these false declarations may be due to the inability of the nonparametric *t*-test to detect differences in tail behavior and bimodality in the gene expression distributions.

In addition, we have done extensive sensitivity analyses to examine the robustness of the proposed methodology. For the gastric cancer dataset, we considered analyses using $\kappa_0 = 0.5$, 1, 2, with the 2 and 3-criteria, along with $\gamma_0 = 0.70, 0.80, 0.90$. The results were remarkably consistent with each other and for a given γ_0 and criterion, the number of differentially expressed genes were nearly identical for all three values of κ_0. Table 4 is based on $\kappa_0 = 1$, yielding for example, 762, 613, and 411 differentially expressed genes for κ_0=0.7, 0.8, 0.90 based on the 2-criterion. For $\kappa_0 = 0.5$, the number of differentially expressed genes was 76, 611, and 408, corresponding to a matching percentage rate (with $\kappa_0 = 1$) of 99.74%, 99.67%, and 99.50%, respectively. Similar results were obtained for $\kappa_0 = 2$ and the 3-criterion. In fact, the matching percentage was always at least 99.32% for any combination of γ_0, type of criterion (2-or 3-criterion) and $\kappa_0 = 0.5, 2$.

Finally, we compared both the classification and predictive accuracy of our proposed model via cross-validation using the genes selected by our method and the genes selected by SAM. Cross-validation was carried out using the PAM, proposed by Tibshirani et al. (2002).We use 10-fold cross-validation, and determine the degree of shrinkage in the calculation of the nearest shrunken centroids by minimizing the cross-validated and test errors. The results show that

both our method and SAM are quite capable of producing excellent classification and predictive power.

7. Discussion

We have proposed a useful two component mixture model for analyzing gene expression data. The proposed model is especially useful in situations where bimodality exists in the gene expression distributions. Such bimodality is common when there are many non-expressed as well as expressed genes for a given tissue type. The proposed methodology has been shown to outperform other well known methods for detecting differentially expressed genes. Future work includes further evaluations of this methodology on real datasets to see if the proposed methods can uncover differentially expressed genes when the biological truth is known. Another point of further evaluation is to apply the proposed methodology on the affymetrix Latin square database to see if one can further recover genes that are known to be differentially expressed in various settings.

Extensions of our proposed model to 3 or more components is straightforward, accommodating situations involving underexpressed, expressed, and overexpressed genes in a given dataset. For example, suppose we consider the following three component mixture:

$$p(y_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg},\phi_{jg},\eta^2_{jg}) = p_{1jg}\, p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg}) + p_{2jg}\, p_2(y_{jgi}|\mu_{jg},\sigma^2_{jg}) + p_{3jg}\, p_3(y_{jgi}|\phi_{jg},\eta^2_{jg}), \tag{7.1}$$

where $p_{ljg} \geq 0, l=1,2,3, p_{1jg}+p_{2jg}+p_{3jg}=1,$

$$p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg}) = (2\pi)^{-1/2}|h'(y_{jgi})|\tau^{-1}_{jg} \exp\left\{-\frac{1}{2\tau^2_{jg}}(\mathbf{h(y_{jgi})} - \alpha_{jg})^2\right\},$$

$$p2(y_{jgi}\,|\mu_{jg},\sigma^2_{jg}) = (2\pi)^{-1/2}|h'(y_{jgi})|\,\sigma^{-1}_{jg} \exp\left\{-\frac{1}{2\sigma^2_{jg}}(\mathbf{h(y_{jgi})} - \mu_{jg})^2\right\},$$

and

$$p3(y_{jgi}|\phi_{jg},\eta^2_{jg}) = (2\pi)^{-1/2}|h'(y_{jgi})|\eta^{-1}_{jg} \exp\left\{-\frac{1}{2\eta^2_{jg}}(\mathbf{h(y_{jgi})} - \phi_{jg})^2\right\}.$$

In (7.1), $p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg})$, $p2(y_{jgi}|\mu_{jg},\sigma^2_{jg})$, and $p3(y_{jgi}|\phi_{jg},\eta^2_{jg})$ are the distributions for y_{jgi} for the underexpressed, expressed, and over expressed genes, respectively. An equivalent model to (7.1) with random threshold parameters can be constructed as follows. Let $c_{jgi} = (c_{1jgi}, c_{2jgi})'$ denote a vector of two *random threshold parameters* such that the gene is underexpressed if $y_{jgi} \leq c_{1jgi}$ and expressed if $c_{1jgi} < y_{jgi} \leq c_{2jgi}$ for the jth tissue type, ith individual, and the gth gene. Assume that the c_{jgi} 's are i.i.d. with a continuous bivariate distribution $p(\mathbf{c}_{jgi}|\mathbf{c}_{0cjg}, \Sigma_{0cjg})$ with support $\Omega_{\mathbf{c}} = \{0<c_{1jgi} < c_{2jgi} < \infty\}$. Let $A_{1jgi} = \{y: y \leq c_{1jgi}\}$, $A_{2jgi} = \{y: c_{1jgi} < y \leq c_{2jgi}\}$, and $A_{3jgi} = \{y: y > c_{2jgi}\}$, and suppose $(y_{jgi}, \mathbf{c}_{jgi})$ have the following joint distribution:

$$\begin{aligned} p(y_{jgi,}\mathbf{c}_{jgi}|\alpha_{jg},\tau^2_{jg},\mu_{jg},\sigma^2_{jg},\phi_{jg},\eta^2_{jg},\mathbf{c}_{0jg},\Sigma_{0jg}) \\ = [\, p_{1jg}\, p_1(y_{jgi}|\alpha_{jg},\tau^2_{jg}) 1_{A1jgi}(y_{jgi})/q_{1y_{jgi}} + p_{2jg}\, p_2(yjgi|\mu_{jg},\sigma^2_{jg}) 1_{A_{2jgi}}(y_{jgi})/q_{2y_{jgi}} \\ \times\; p_{3jg}\, p_{3\,(yjgi}|\phi_{jg},\eta^2_{jg}) 1_{A3jgi}(y_{jgi})/q3_{y_{jgi}}]p(\mathbf{c}_{jgi}|\mathbf{c}_{0jg,}\Sigma_{0cjg}), \end{aligned} \tag{7.2}$$

where $q_{1y_{jgi}} = \int y_{jgi} \le c_{1jgi} < c_{2jgi}\, p(\mathbf{c}_{jgi} | \mathbf{c}_{0jg}, \Sigma_{0cjg})\, d\mathbf{c}_{jgi}$, $q_{2y_{jgi}} = \int c_{1jgi} < y_{jgi} \le c_{2jgi}\, p(\mathbf{c}_{jgi} | \mathbf{c}_{0jg}, \Sigma_{0cjg})\, d\mathbf{c}_{jgi}$, and $q_{3y_{jgi}} = \int c_{1jgi} < c_{2jgi} < y_{jgi}\, p(\mathbf{c}_{jgi} | \mathbf{c}_{0jg}, \Sigma_{0cjg})\, d\mathbf{c}_{jgi}$. Following the proof of Identity 2.1, we can show that (7.2) reduces to (7.1) after integrating out $\mathbf{c}_{jgi}$. Similar to the two component mixture model (2.1), we take

$$p(\mathbf{c}_{jgi} | \mathbf{c}_{0jg}, \sum\nolimits^{2}_{0cjg}) \propto |h\prime(c_{1jgi}) h\prime(c_{2jgi})| |\Sigma_{0cjg}|^{-1/2} \times \exp\left\{ -\tfrac{1}{2} (\mathbf{h}(\mathbf{c_{jgi}}) - \mathbf{h}(\mathbf{c_{0jg}}))\prime \sum\nolimits^{-1}_{\mathbf{0cjg}} (\mathbf{h}(\mathbf{c_{jgi}}) - \mathbf{h}(\mathbf{c_{0jg}})) \right\},\ 0 < c_{1jgi} < c_{2jgi} < \infty,$$

where $h(\mathbf{c}_{jgi}) = (h(c_{1jgi}), h(c_{2jgi}))'$ and $h(c_{0jg}) = (h(c_{01jg}), h(c_{02jg}))'$. We then elicit the hyperparameters c_{0jg} and $\sum^{2}_{0cjg}$ using the summary statistics of the quantiles of the y_{jgi}'s.

References

Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bioinformatics 2001;17:509–519. [PubMed: 11395427]

Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics 1997;4:364–374.

Chen X, Leung S, Yuen ST, Chu K-M, Ji J, Li R, Chan SY, Law S, Troyanskaya OG, Wong J, Samuel S, Botstein D, Brown PO. Variation in gene expression patterns in human gastric cancers. Molelcular Biol. Cell 2003;14:3208–3215.

Do K-A, Mueller P, Tang F. A bayesian mixture model for differential gene expression. Appl. Statistics 2005;54:611–626.

Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Statist. Sinica 2002;12:111–139.

Efron B, Tibshirani R, Storey J, Tusher VG. Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc 2001;96:1151–1160.

Hein A-M, Richardson S, Cuaston HC, Graeme AK, Green PJ. BGX: a fully bayesian integrated approach to the analysis of affymetrix GeneChip data. Biostatistics 2005;6:349–373. [PubMed: 15831583]

Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression with DNA microarray data. J. Amer. Statist. Assoc 2002;97:88–99.

Ishwaran H, Rao JS. Detecting differentially expressed genes in microarrays using Bayesian model selection. J. Amer. Statist. Assoc 2003;98:438–455.

Ishwaran H, Rao S. Spike and slab gene selection of multigroup microarray data. J. Amer. Statist. Assoc 2005;100:764–780.

Kendziorski CM, Newton MA, Lan H, Gould MN. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. Statist. Med 2003;22:3899–3914.

Kerr MK, Martin M, Churchhill GA. Analysis of variance for gene expression microarray data. J. Comput. Biol 2000;7:819–837. [PubMed: 11382364]

Lee MLT, Lu W, Whitmore GA, Beier D. Models for microarray gene expression data. J. Biopharm. Statist 2002;12:1–19.

Liu, D.; Parmigiani, G.; Caffo, B. Screening for differentially expressed genes: are multilevel models helpful?. Technical Report, Department of Biostatistics, Johns Hopkins University; 2004.

Lonnstedt I, Speed T. Replicated microarray data. Statist. Sinica 2002;12:31–46.

Mueller P, Parmagiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. J. Amer. Statist. Assoc 2004;99:990–1001.

Mutter GL, Baak JPA, Fitzgerald JT, Gray RJ, Neuberg D, Kust GA, Gentleman R, Gullens SR, Wei LJ, Wilcox M. Global expression changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation. Gynecol. Oncol 2001;83:177–185. [PubMed: 11606070]

Newton, MA.; Kendziorski, CM. The Analysis of Gene Expression Data: An Overview of Methods and Software. New York: Springer; 2003. Parametric empircal bayes methods for microarrays; p. 254-271.

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J. Comput. Biol 2001;8:37–52. [PubMed: 11339905]

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 2004;5:155–176. [PubMed: 15054023]

Olshen AB, Jain AN. Deriving quantitative conclusions from microarray data. Bioinformatics 2002;18:961–970. [PubMed: 12117794]

Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. J. Roy. Statist. Soc. Ser. B 2002;64:717–736.

Parmigiani, G.; Garrett, ES.; Irizarry, RA.; Zeger, SL., editors. The Analysis of Gene Expression Data: An Overview of Methods and Software. New York: Springer; 2003.

Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. Internat. J. Cancer 1999;80:827–841.

Sebastiani P, Gussoni E, Kohane IS, Ramoni MF. Statistical challenges in functional genomics (with discussion). Statist. Sci 2003;18:33–70.

Storey, J.; Tibshirani, R. The Analysis of Gene Expression Data: An Overview of Methods and Software. New York: Springer; 2003. SAM thresholding and false discovery rates for detenting differential gene expression in DNA microarrays; p. 272-290.

Tadesse MG, Ibrahim JG, Mutter G. Identification of differentially expressed genes in high-density oligoneucleotide arrays accounting for the quantification limits of the technology. Biometrics 2003;59:542–554. [PubMed: 14601755]

Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat. Acad. Sci 2002;99:6567–6572. [PubMed: 12011421]

Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002;18:1454–1461. [PubMed: 12424116]

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci 2001;98:5116–5121. [PubMed: 11309499]

West, AP. Bayesian factor analysis regression for models in the “Large *p*, Small *m*” Paradigm. In: Bernardo, JM.; Bayarri, MJ.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM.; West, M., editors. Bayesian Statistics 7. Oxford: Oxford University Press; 2003. p. 733-742.

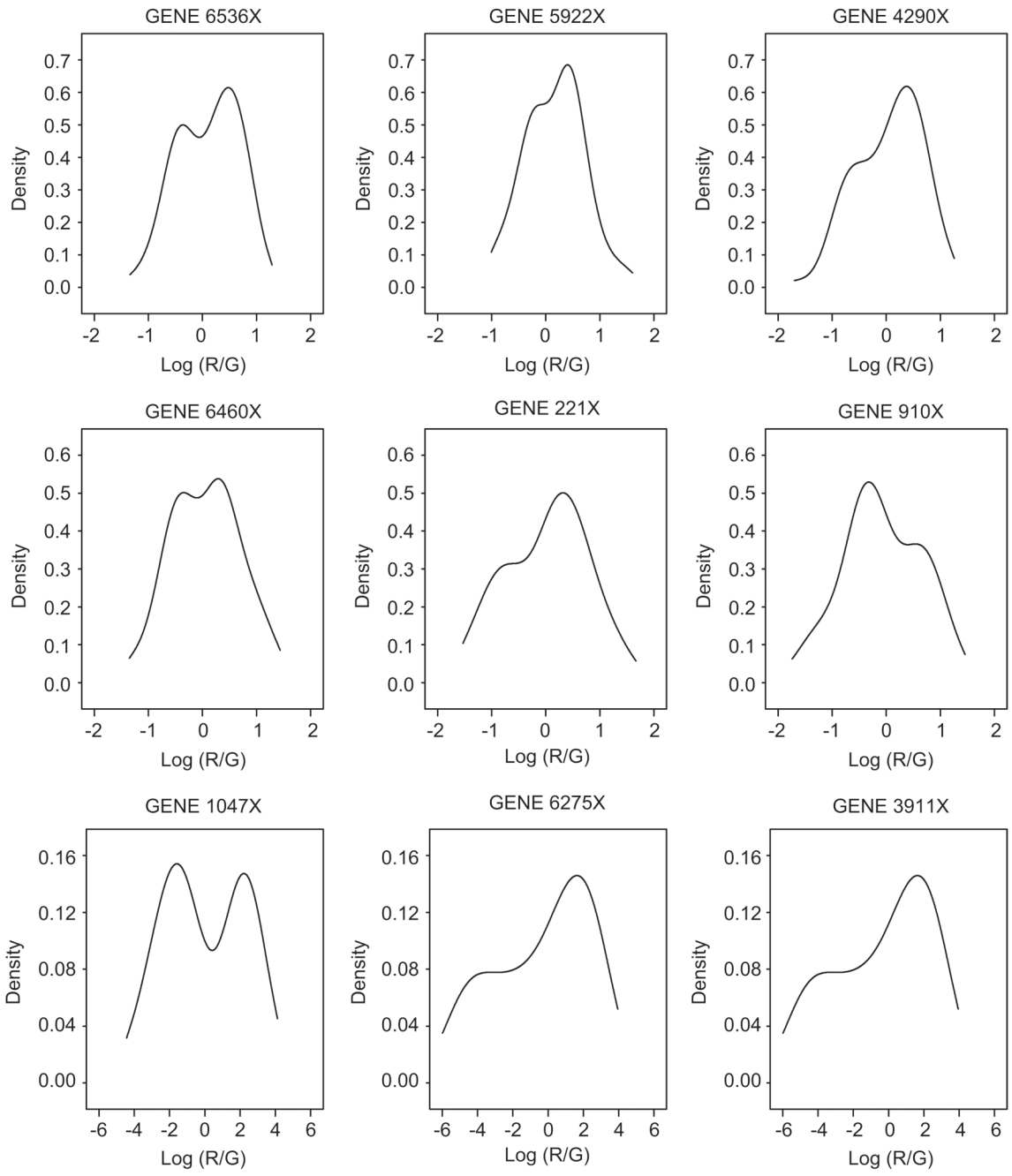

Fig 1.
Densities for nine selected genes.

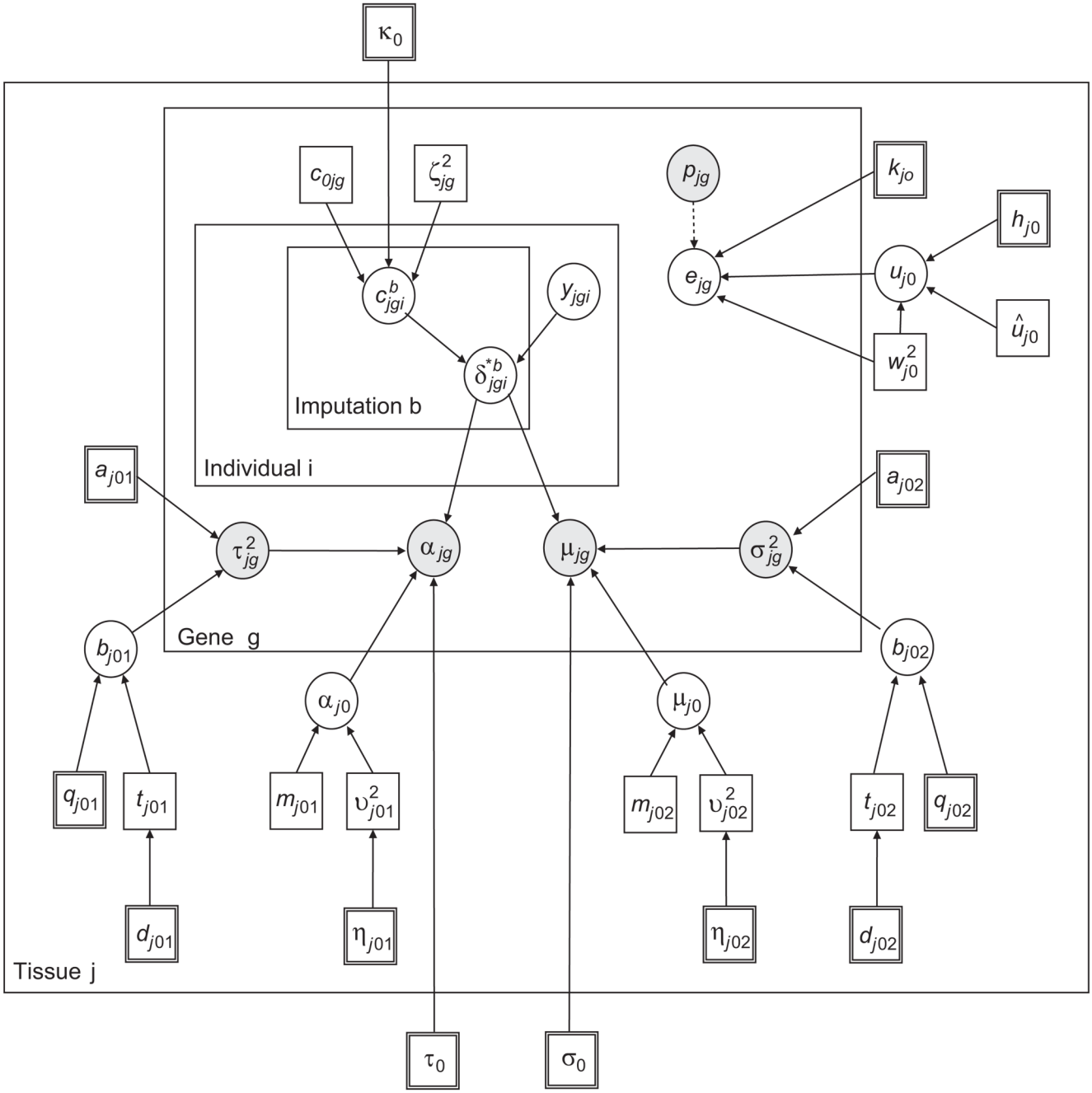

Fig 2.

Graphical display in prior specification. Elements in circles are stochastic, while elements in squares are empirically specified hyperparameters. Shaded circles indicate parameters of interest. Double squares correspond to prespecified scalar hyperparameters.

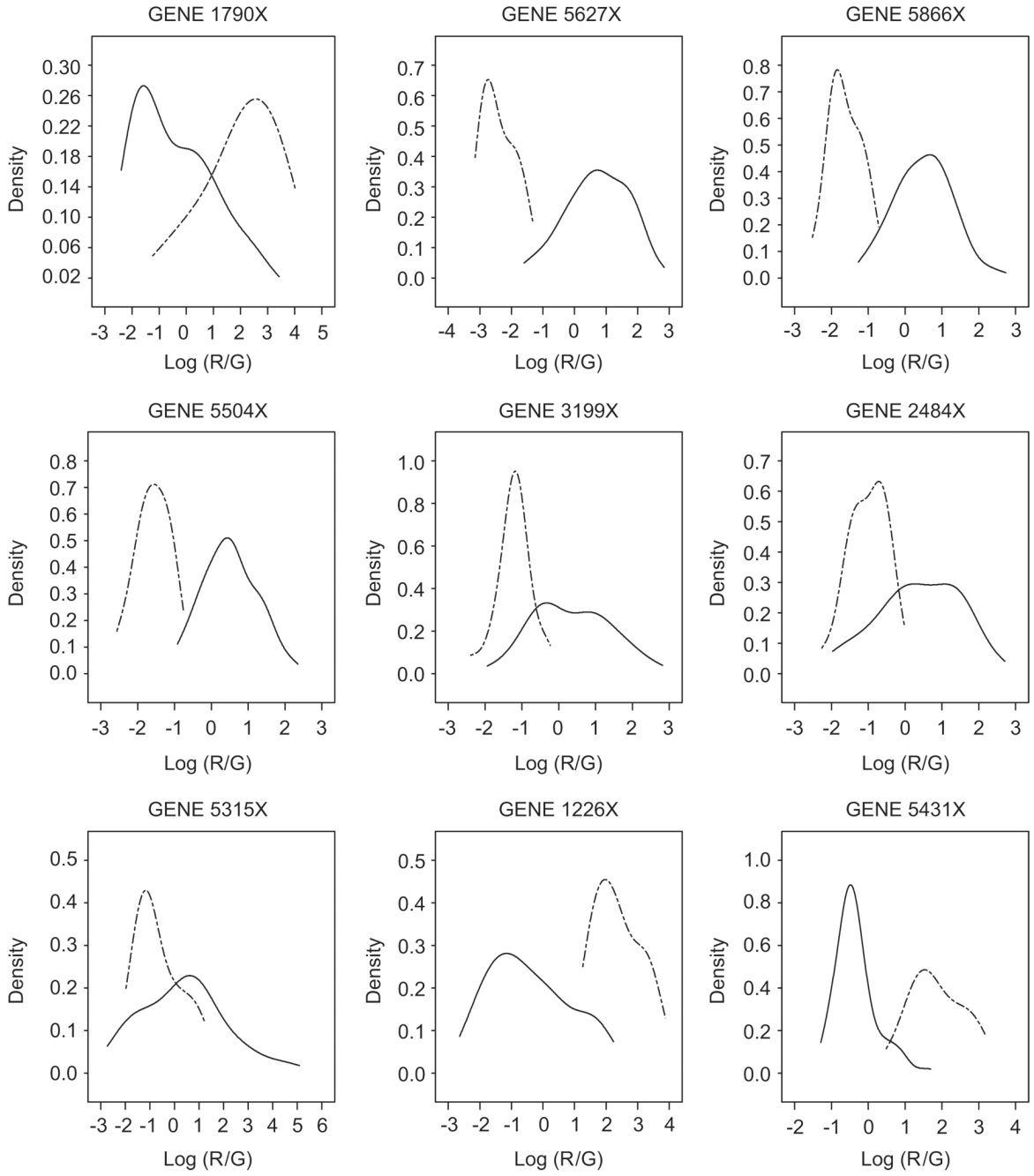

Fig 3.

Densities for nine genes deemed differentially expressed by the proposed methodology as well as by Chen et al. (2003) (solid: tumor, dashed: normal).

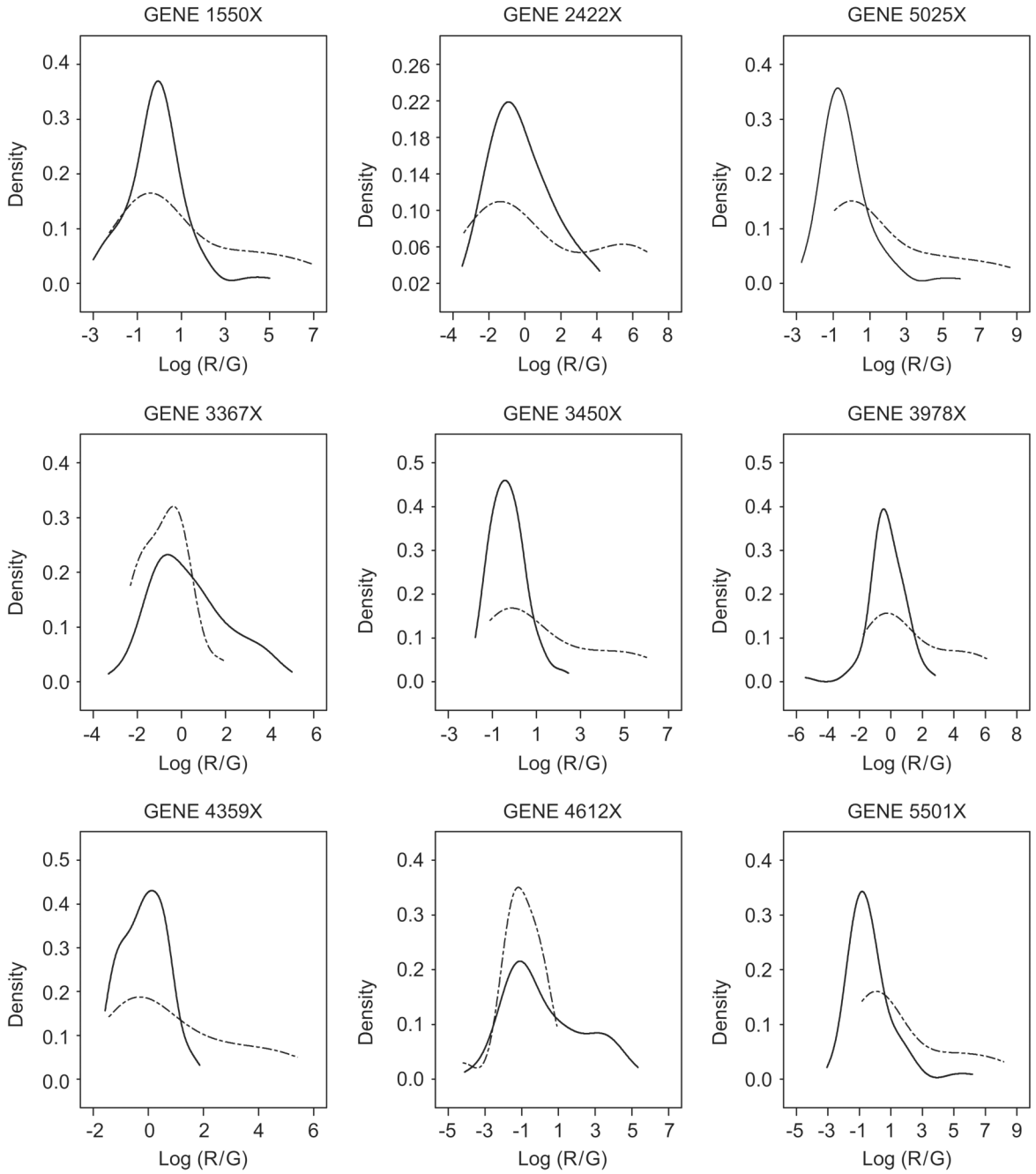

Fig 4.

Densities for nine genes that were deemed differentially expressed by the proposed methodology but not deemed differentially expressed by Chen et al. (2003) (solid: tumor, dashed: normal).

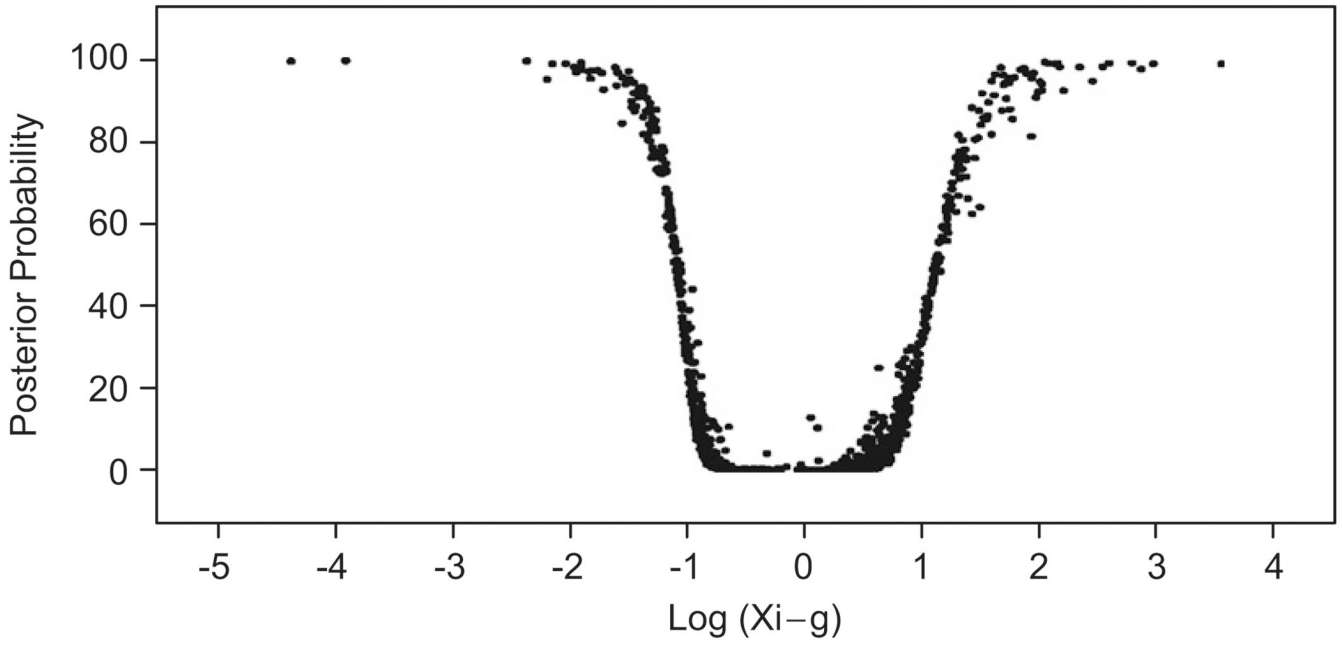

Fig 5.
Posterior probability ($\max\{\gamma_{g31}, \gamma_{g32}\}$) versus the posterior mean of $\log(\xi_g)$ for 3329 genes selected by Chen et al. (2003).

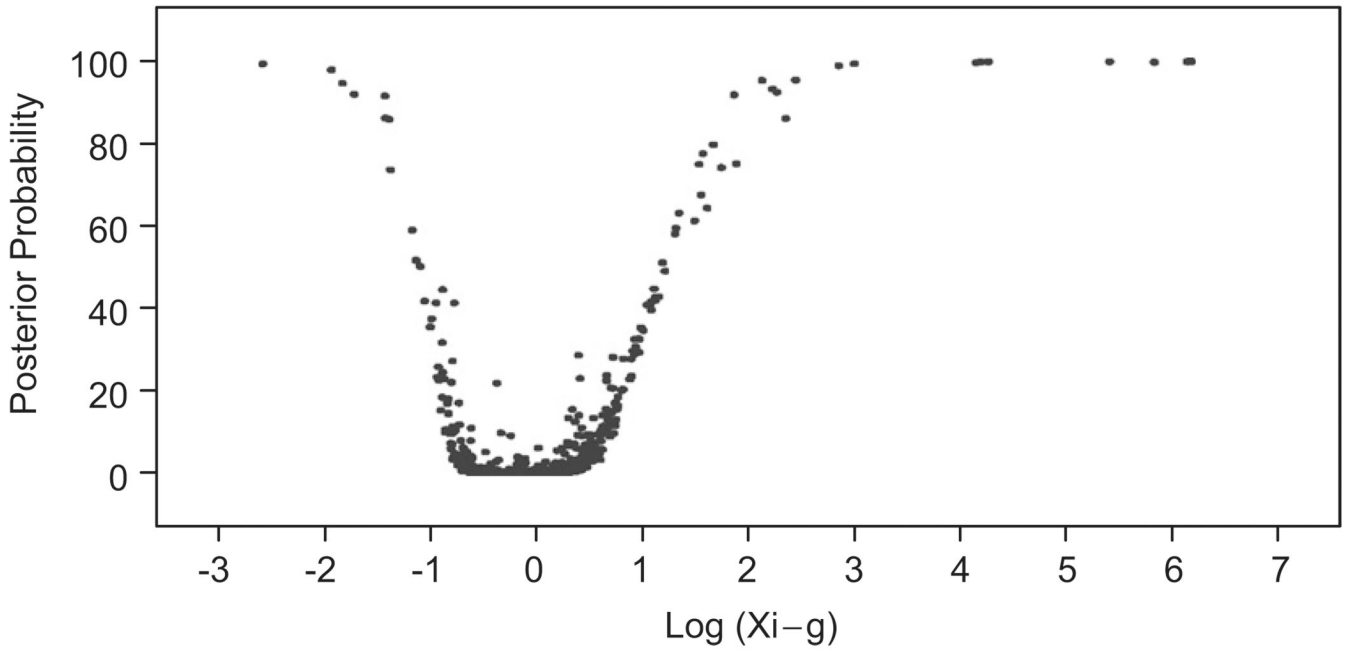

Fig 6.
Posterior probability ($\max\{\gamma_{g31}, \gamma_{g32}\}$) versus the posterior mean of $\log(\xi_g)$ for genes that were not selected by Chen et al. (2003).

Table 1

False negative rate and false positive rate of the proposed criterion under model (2.5)

Prior	γ_o	2-criterion				3-criterion			
		Mean		Mean		Mean		Mean	
		FNR	SD	FNR	SD	FNR	SD	FNR	SD
I	0.70	0.0001	0.0013	0.0965	0.0343	0.0004	0.0029	0.0241	0.0173
	0.80	0.0003	0.0025	0.0381	0.0157	0.0014	0.0052	0.0070	0.0052
	0.85	0.0007	0.0037	0.0198	0.0094	0.0026	0.0071	0.0030	0.0026
	0.90	0.0019	0.0061	0.0079	0.0046	0.0060	0.0114	0.0009	0.0011
	0.95	0.0019	0.0061	0.0016	0.0015	0.0202	0.0191	0.0001	0.0004
II	0.70	0.0001	0.0009	0.0790	0.0094	0.0003	0.0024	0.0164	0.0043
	0.80	0.0003	0.0024	0.0299	0.0058	0.0012	0.0050	0.0045	0.0022
	0.85	0.0006	0.0035	0.0153	0.0042	0.0026	0.0072	0.0019	0.0015
	0.90	0.0017	0.0058	0.0057	0.0025	0.0062	0.0114	0.0006	0.0008
	0.95	0.0017	0.0058	0.0012	0.0011	0.0210	0.0206	0.0001	0.0003
III	0.70	0.0003	0.0025	0.0700	0.0083	0.0017	0.0056	0.0121	0.0035
	0.80	0.0011	0.0046	0.0284	0.0055	0.0046	0.0093	0.0038	0.0019
	0.85	0.0028	0.0072	0.0153	0.0041	0.0078	0.0125	0.0017	0.0014
	0.90	0.0058	0.0105	0.0067	0.0026	0.0153	0.0180	0.0006	0.0008
	0.95	0.0058	0.0105	0.0016	0.0013	0.0410	0.0268	0.0001	0.0003

J Stat Plan Inference. Author manuscript; available in PMC 2009 August 10.

Table 2

Comparison of three methods

Method	(n_1, n_2)	γ_0	Mean FNR	SD	Mean FPR	SD
Proposed criterion	(25, 25)	0.70	0.0004	0.0029	0.0241	0.0173
under model (2.5)		0.80	0.0014	0.0052	0.0070	0.0052
		0.90	0.0060	0.0114	0.0009	0.0011
SAM (FDR ≤ 0.05)			0.0000	0.0000	0.0013	0.0011
SAM (FDR ≤ 0.10)			0.0000	0.0000	0.0038	0.0022
PERMAX (α = 0.05)			0.7150	0.0627	0.0000	0.0002
PERMAX (α = 0.10)			0.6382	0.0688	0.0001	0.0003
EBarrays (PP > 0.5)			0.0000	0.0000	0.9999	0.0003
EBarrays (PP > 0.7)			0.0000	0.0000	0.9835	0.0247
Proposed criterion	(20, 20)	0.70	0.0020	0.0061	0.0341	0.0201
under model (2.5)		0.80	0.0058	0.0105	0.0103	0.0061
		0.90	0.0197	0.0198	0.0014	0.0014
SAM (FDR ≤ 0.05)			0.0003	0.0025	0.0015	0.0012
SAM (FDR ≤ 0.10)			0.0001	0.0015	0.0038	0.0023
PERMAX (α = 0.05)			0.8462	0.0510	0.0001	0.0003
PERMAX (α = 0.10)			0.7871	0.0597	0.0001	0.0004
EBarrays (PP > 0.5)			0.0000	0.0000	0.9999	0.0002
EBarrays (PP > 0.7)			0.0004	0.0029	0.9864	0.0159
Proposed criterion	(10, 10)	0.70	0.0283	0.0238	0.0709	0.0334
under model (2.5)		0.80	0.0570	0.0349	0.0235	0.0125
		0.90	0.1400	0.0565	0.0033	0.0028
SAM (FDR ≤ 0.05)			0.0855	0.0528	0.0014	0.0013
SAM (FDR ≤ 0.10)			0.0409	0.0361	0.0040	0.0022
PERMAX (α = 0.05)			0.9890	0.0142	0.0000	0.0001
PERMAX (α = 0.10)			0.9799	0.0196	0.0001	0.0003
EBarrays (PP > 0.5)			0.0006	0.0037	0.9995	0.0011
EBarrays (PP > 0.7)			0.1319	0.1385	0.4254	0.3458
Proposed criterion	(5, 5)	0.70	0.1074	0.0532	0.1403	0.0943
under model (2.5)		0.80	0.1822	0.0693	0.0563	0.0466

Method	(n_1, n_2)	γ_0	Mean FNR	SD	Mean FPR	SD
		0.90	0.3441	0.0902	0.0110	0.0114
SAM (FDR ≤ 0.05)			0.6104	0.1443	0.0007	0.0010
SAM (FDR ≤ 0.10)			0.4634	0.1624	0.0023	0.0020
PERMAX (α = 0.05)			0.9978	0. 0067	0.0000	0.0002
PERMAX (α = 0.10)			0.9957	0.0101	0.0001	0.0003
EBarrays (PP > 0.5)			0.0124	0.0184	0.9922	0.0115
EBarrays (PP > 0.7)			0.9114	0.1327	0.0026	0.0104

Table 3

Sensitivity analysis

Method	(n_1, n_2)	γ_0	Mean FNR	SD	Mean FPR	SD
Proposed criterion	(25, 25)	0.70	0.0000	0.0000	0.0963	0.1224
under model (2.5)		0.80	0.0000	0.0000	0.0202	0.0351
		0.90	0.0001	0.0023	0.0012	0.0021
SAM (FDR ≤ 0.05)			0.4290	0.1043	0.0014	0.0015
SAM (FDR ≤ 0.10)			0.2800	0.0834	0.0040	0.0024
PERMAX (α = 0.05)			0.8046	0.0556	0.0000	0.0002
PERMAX (α = 0.10)			0.7428	0.0621	0.0001	0.0003
EBarrays (PP > 0.5)			0.0000	0.0000	0.9999	0.0004
EBarrays (PP > 0.7)			0.0063	0.0128	0.8935	0.1248
Proposed criterion	(20, 20)	0.70	0.0000	0.0009	0.1217	0.1408
under model (2.5)		0.80	0.0001	0.0013	0.0286	0.0476
		0.90	0.0002	0.0018	0.0020	0.0040
SAM (FDR ≤ 0.05)			0.5992	0.0886	0.0011	0.0011
SAM (FDR ≤ 0.10)			0.4646	0.1042	0.0029	0.0021
PERMAX (α = 0.05)			0.8980	0.0430	0.0001	0.0003
PERMAX (α = 0.10)			0.8588	0.0501	0.0001	0.0003
EBarrays (PP > 0.5)			0.0001	0.0015	0.9997	0.0006
EBarrays (PP > 0.7)			0.0304	0.0381	0.6714	0.2652
Proposed criterion	(10, 10)	0.70	0.0024	0.0072	0.2145	0.2037
under model (2.5)		0.80	0.0058	0.0109	0.0718	0.1004
		0.90	0.0147	0.0180	0.0100	0.0170
SAM (FDR ≤ 0.05)			0.9176	0.0469	0.0007	0.0009
SAM (FDR ≤ 0.10)			0.8909	0.0685	0.0011	0.0012
PERMAX (α = 0.05)			0.9890	0.0141	0.0000	0.0002
PERMAX (α = 0.10)			0.9798	0.0194	0.0001	0.0003
EBarrays (PP > 0.5)			0.0027	0.0074	0.9970	0.0039
EBarrays (PP > 0.7)			0.4563	0.1507	0.0688	0.06111
Proposed criterion	(5, 5)	0.70	0.0265	0.0269	0.2989	0.2343
under model (2.5)		0.80	0.0510	0.0372	0.1301	0.1388

Method	(n_1, n_2)	γ_0	Mean FNR	SD	Mean FPR	SD
		0.90	0.0991	0.0478	0.0308	0.0334
SAM (FDR ≤ 0.05)			0.9724	0.0226	0.0011	0.0011
SAM (FDR ≤ 0.10)			0.9717	0.0252	0.0011	0.0012
PERMAX (α = 0.05)			0.9996	0. 0028	0.0000	0.0002
PERMAX (α = 0.10)			0.9990	0.0045	0.0001	0.0003
EBarrays (PP > 0.5)			0.0209	0.0316	0.9764	0.0335
EBarrays (PP > 0.7)			0.9020	0.1032	0.0067	0.0102

Table 4

Number of genes differentially expressed using $\kappa_0 = 1$

	$\gamma_0 = 0.70$	$\gamma_0 = 0.80$	$\gamma_0 = 0.90$
2-*criterion*			
Number of genes selected	762	613	411
Number of genes matched with Chen et al. (2003)	695	563	379
Matched percentage	91.21%	91.84%	92.21%
Number of genes matched with SAM (FDR ≤ 0.05)	739	598	403
Matched percentage	96.98%	97.55%	98.05%
Number of genes matched with SAM (FDR ≤ 0.10)	744	602	406
Matched percentage	97.64%	98.21%	98.78%
3-*criterion*			
Number of genes selected	188	145	98
Number of genes matched with Chen et al. (2003)	160	123	79
Matched percentage	85.11%	84.83%	80.61%
Number of genes matched with SAM (FDR ≤ 0.05)	183	141	96
Matched percentage	97.34%	97.24%	97.96%
Number of genes matched with SAM (FDR ≤ 0.10)	186	143	97
Matched percentage	98.94%	98.62%	98.98%

Note that when FDR ≤ 0.05, 4511 genes were identified differentially expressed, while when FDR≤0.1, 5082 genes were identified different