A Bayesian Mixture Model for Differential Gene Expression

Kim-Anh Do, Peter Müller and **Feng Tang**[1]

Abstract

We propose model-based inference for differential gene expression, using a non-parametric Bayesian probability model for the distribution of gene intensities under different conditions. The probability model is essentially a mixture of normals. The resulting inference is similar to the empirical Bayes approach proposed in Efron et al. (2001). The use of fully model-based inference mitigates some of the necessary limitations of the empirical Bayes method. However, the increased generality of our method comes at a price. Computation is not as straightforward as in the empirical Bayes scheme. But we argue that inference is no more difficult than posterior simulation in traditional nonparametric mixture of normal models. We illustrate the proposed method in two examples, including a simulation study and a microarray experiment to screen for genes with differential expression in colon cancer versus normal tissue (Alon et al., 1999).

KEY WORDS: Density Estimation; Dirichlet Process; Gene Expression; Microarrays; Mixture Models; Nonparametric Bayes.

[1]Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX.
Research partially supported by NIH/NCI grant 2 R01 CA75981-04A1, the University of Texas SPORE in Prostate Cancer grant CA90270, and the Early Detection Research Network grant CA99007.
The sequence of authors is strictly alphabetical.

1 INTRODUCTION

We discuss the use of nonparametric Bayesian inference to analyze data from microarray experiments conducted to screen for differential gene expression over conditions of interest. We define a model-based version of the empirical Bayes approach proposed in Efron et al. (2001). The probability model is a variation of traditional Dirichlet process (DP) mixture models. The model includes an additional mixture corresponding to the assumption that observed transcription levels arise as a mixture over non-differentially and differentially expressed genes. Inference proceeds as in DP mixture models, with an additional set of latent indicators to resolve the additional mixture.

1.1 Background

With the recent advent in DNA array technologies, a new class of large data sets emerge from microarray experiments that allow researchers to measure the relative expression of thousands of genes simultaneously. Microarrays measure mRNA concentrations by labeling the sample with a dye and then allowing them to hybridize to spots on the array. Each spot contains either DNA oligomers (typically 25 nucleotides) or a longer DNA sequence (hundreds of nucleotides long) designed to be complementary to a particular messenger RNA of interest. There are two main types of arrays: Oligonucleotide arrays generated by photolithography techniques to synthesize oligomers directly on the glass slide (primarily Affymetrix arrays); and cDNA arrays generated by mechanical gridding, where prepared material is applied to each spot by ink-jet or physical deposition. For oligonucleotide microarrays, cross-hybridization may occur, that is, multiple genes may hybridize to the same spot. Therefore oligonucleotide arrays must measure each gene with a probe set of oligomers (for example, Affymetrix arrays use probe set sizes of 32-40), and the identification of a gene is only made if "positive" hybridization can be detected in the majority of the probes in the set. Oligonucleotide arrays are manufactured with probes that form a perfect match (PM) and a mismatch (MM) with the target polynucleotide of interest. The PM oligo probe will contain a segment of a wild-type allele (creating a perfect complementary match with a segment of the target polynucleotide of interest), while the MM oligo probe will be a copy of the PM oligo that has been altered by one base at a central position, usually the thirteenth position. In current practice, Affymetrix oligonucleotide arrays measure a single sample at a time with a single type of dye. In contrast, cDNA microarrays can use two or more different fluorescent dyes to label different samples, thus allowing simultaneous monitoring of multiple samples on the same array. See Wu (2001) for a gentle introductory review of microarray technologies. Statistical methods applicable to the analysis of such data have an important role to play in the discovery, validation, and understanding of various classes and subclasses of cancer. See, for example, Eisen et al. (1998), Alizadeh et al. (2000), Ben-Dor et al. (1999, 2000), Alon et al. (1999), Golub et al. (1999), Moler et al. (2000), and Xing and Karp (2001), among other. The different stages of a microarray experiment (not necessarily distinct) include experimental design, image analysis, graphical presentation and normalization, identification of differentially expressed genes, and finally clustering or classification of the gene expression profiles. See Smyth et al. (2002) for a review of statistical issues and corresponding methods for these numerous stages. In this article, we use the term "expression level" to refer to a summary measure of relative red to green channel intensities in a fluorescence-labeled cDNA array or a summary difference of the PM and MM scores from an oligonucleotide array.

1.2 Inference for Differential Expression

Recently, statisticians and researchers in bioinformatics have focused much attention on the development of statistical methods to identify differentially expressed genes, with special emphasis on those methods that identify genes that are differentially expressed between two conditions. There are two main approaches to distinguish real signals from noise in a chip-to-chip comparison: thresholding or replicate analysis. The latter approach is desirable but expensive. The former approach involves imposing an arbitrary threshold of signal difference, or fold-change ratio, between experimental and control samples, above which differences are considered to be real. Often differential expression is assessed by taking ratios of expression levels of different samples at a spot on the array and flagging those where the magnitude of the fold difference exceeds some threshold, possibly allowing for the fact that the variability of these ratios is not constant across mean expression levels. For example, an increase or decrease of at least twofold may be considered as significant. The use of fold-change ratios can be inefficient and erroneous. The uncertainty associated with dividing two intensity values further increases overall errors (Newton et al., 2001; Yang et al., 2001; Miles, 2001). The methods are often variants of Student's t-test that conduct a hypothesis test at each gene and subsequently correct for multiple comparisons. Earlier simple methods were discussed in Schena et al. (1995), Schena et al. (1996), DeRisi et al. (1996) and and Lönnstedt and Speed (2002). Chen et al. (1997) considered a less arbitrary threshold by using replicated housekeeping genes; and more recently methods that implicitly assume non-constant coefficient of variation were discussed by Baggerly et al. (2001), Newton et al. (2001), and Rocke and Durbin (2001).

A recent strategy for the detection of differentially expressed genes, called significance analysis of microarrays (SAM), has been described by Tusher et al. (2002). SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific t-tests. The approach incorporates means and standard deviations across experimental conditions in the computation of a relative difference in gene expression. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of the repeated measurements for this specific gene corresponding to different experimental conditions. To prevent the denominator of the t-statistic from getting too small, Tusher et al. (2002) proposed a refinement of the t statistic by adding a constant term a_0 to the denominator of the standardized average. The constant term a_0 can be taken to equal the nth percentile of the standard errors of all the genes, as suggested by Efron et al. (2001), or as a value that minimizes the coefficient of variation of the t-statistic, as suggested by Tusher et al. (2002).

A number of researchers have employed mixture modeling approaches in the analysis of microarray data. For example, McLachlan et al. (2002) developed the software EMMIX-GENE that includes a mixture of t distributions, using the Student-t family as a heavy-tailed alternative to the normal distribution. The use of a mixture of normal distributions as a flexible and powerful tool to estimate the two distributions related to gene expression has been discussed, for example, in Pan et al. (2002). They use a parametric bootstrap technique to choose a cut point in declaring statistical significance for identified genes while controlling for the number of false positives. Efron et al. (2001) discuss, in addition to other methods, the use of density estimates to approximate the distribution of transcription levels for differentially and non-differentially expressed genes. The main thrust of the discussion in Efron et al. (2001) is an empirical Bayes approach. They compute the posterior probability of differential expression by substituting estimates of relevant parameters and (ratios of) densities based on the empirical distribution of observed transcription levels. We propose an extension of the method discussed by Efron et al. (2001). Analogous to Efron's empirical Bayes approach, our approach starts with assuming that the observed expression scores are generated from a mixture of two distributions that can be interpreted as distributions for affected and unaffected genes, respectively. The desired inference about differential expression for a particular gene amounts to solving a deconvolution problem corresponding to this mixture. While Efron et al. (2001) proceed by plugging in point estimates, we choose a fully model-based approach. We construct a probability model for the unknown mixture, allowing investigators to deduce the desired inference about differential expression as posterior inference in that probability model. We choose Dirichlet process mixture models to represent the probability model for the unknown distributions. We develop Markov chain Monte Carlo (MCMC) posterior simulation to generate samples from the relevant posterior and posterior predictive distributions.

2 DATA

Alon et al. (1999) used Affymetrix oligonucleotide arrays to monitor expressions of over 6,500 human gene expressions in 40 tumor and 22 normal colon tissue samples. The samples were taken from 40 different patients, with 22 patients supplying both a tumor and normal tissue sample. Alon et al. (1999) focused on the 2,000 genes with highest minimal intensity across the samples, and it is these 2,000 genes that comprise our data set. The microarray data matrix thus has $n = 2,000$ rows and $p = 62$ columns. We have rearranged the data so that the tumors are labeled 1 to 40 and the normals 41 to 62. The first 11 columns report tumor tissue samples collected under protocol P1 (using a poly detector), columns 12-40 are from tumor tissue samples collected under protocol P2 (using total extraction of RNA), columns 41-51 are normal tissue samples collected under P1 from the same patients as columns 1-11, and columns 52-62 are normal tissue samples collected under protocol P2 from the same patients as columns 12-22.

From the data matrix we construct two difference matrices, D and d. The first matrix, D, contains all the possible differences between tumor and normal tissues within the same protocol (P1 or P2), with the i-th row of D defined as the vector of all differences for the i-th gene. The other matrix, d, contains all possible differences within the same conditions and same protocol, i.e., differences between all pairs of tumor columns, and between all pairs of normal columns collected under the same protocol. Also, in constructing D, we exclude differences of paired columns corresponding to the same patient. Including such differences would require the introduction of patient specific random effects to model the difference in variation between differences of paired and independent columns, respectively. Thus patient to patient variation as well as any other noise is included in both, d and D. But possible effects due to differential expression in tumor versus normal tissues are included only in D. We refer to d as the null sample, reporting only measurement error, and D as the mixed sample, including noise plus a tumor versus normal effect for differentially expressed genes. The goal of the upcoming discussion is to identify those genes that are differentially expressed across the two conditions and separate the mixed sample into a subset of non-differentially expressed genes for which D reports only noise as in d, and differentially expressed genes that show an additional effect in D.

Let $\overline{D_i}$ and $\overline{d_i}$ denote the average of all elements in D and d, respectively, corresponding to gene i, i.e., the average in the i-th row D_i and d_i, respectively. Similar to Efron et al. (2001), we construct two sets of Z scores, Z^{null} and Z^{mix}, obtained as

$$\begin{aligned} Z_i^{\text{mix}} &= \overline{D_i}/(\alpha_0 + S_i) \\ Z_i^{\text{null}} &= \overline{d_i}/(\alpha_0' + s_i) \end{aligned}$$

where S_i and s_i are respectively the standard deviations of D_i and d_i. The offsets α_0 and α_0' are the correction scores. We use $\alpha_0 = \alpha_0' = 0$.

3 A MIXTURE MODEL FOR GENE EXPRESSION

We assume that a gene is either affected or unaffected by the condition of interest. Hence we can write the distribution of expression scores Z_i^{mix} as a mixture of two density functions, f_0 and f_1, representing the density function under unaffected and affected conditions, respectively. For $Z \in \{Z_i^{\text{mix}},\ i = 1, \ldots, n\}$ we assume $Z \sim f(Z)$ with

$$f(Z) = p_0 f_0(Z) + (1 - p_0) f_1(Z) \tag{1}$$

where p_0 is the proportion of genes that are not differentially expressed across the two experimental conditions. The main inference question of interest is about the probability of differential expression. Using Bayes' rule for given (f_0, f_1, p_0) we find from (1) the posterior probability of differential expression

$$P_1(Z \mid f_0, f_1, p_0) = (1 - p_0) f_1(Z)/f(Z), \tag{2}$$

and the complementary probability of non-differential expression $P_0(Z \mid f_0, f_1, p_0) = p_0 f_0(Z)/f(Z)$. Both posterior probabilities are conditional on assumed (f_0, f_1, p_0).

Efron et al. (2001) propose to estimate P_0 by an empirical Bayes approach, substituting point estimates for f_0/f and p_0. To estimate f_0/f they construct a logistic regression experiment, set up such that the odds are $\pi(Z) = f(Z)/(f(Z) + f_0(Z))$. The resulting estimate $\widehat{\pi}$ gives an implied estimate $\widehat{q} = (1 - \widehat{\pi})/\widehat{\pi}$ for $q = f_0/f$. To derive a point estimate for p_0 they observe that non-negativity of P_1 implies $p_0 \leq \min_Z f(Z)/f_0(Z)$, and propose to substitute the bound as point estimate $\widehat{p}_0 \equiv \min_Z f(Z)/f_0(Z)$. We will refer to the resulting estimates $\widehat{P}_0(Z) = \widehat{p}_0\, \widehat{q}(Z)$ and $\widehat{P}_1(Z) = 1 - \widehat{P}_0(Z)$ as empirical Bayes estimates. The bound $\widehat{p}_0$ overestimates p_0 and hence introduces a corresponding bias in $\widehat{P}_0(Z)$ and $\widehat{P}_1(Z)$.

This limitation of the empirical Bayes approach can be overcome by a fully model-based Bayesian approach that introduces a probability model on (f_0, f_1, p_0) and computes posterior probabilities of differential expression as appropriate marginal posterior probabilities.

3.1 Non-parametric Bayesian Approach (NPBA)

Defining a prior probability model for the unknown quantities in (1), and combining this with the relevant sampling distributions we can derive a posterior distribution for the unknown f_0, f_1 and p_0. The implied posterior distribution on $P_1 = (1 - p_0)f_1/f$ provides the desired probabilities of differential expression. The key advantages of this approach are that it replaces point estimates for f_0/f and p_0 by a full description of uncertainties and appropriately accounts for these uncertainties. Also, the approach inherits other relevant advantages of coherent posterior inference. In particular, once we introduce a joint probability model across all genes and samples, we can provide joint inference on multiple genes, and we include accounting for multiplicities in the usual Bayesian fashion. We illustrate these two issues in the context of examples in section 5.

Defining a prior probability model for inference in (2) requires investigators to choose a probability model for the unknown densities f_0 and f_1. Bayesian inference for random distributions, like f_0 and f_1, is known as nonparametric Bayesian inference (Walker et al., 1999). By far the most popular nonparametric Bayesian model is the Dirichlet process (DP). A random probability distribution G is generated by a DP if for any partition $A_1, \ldots, A_k$ of the sample space the vector of random probabilities $G(A_i)$ follows a Dirichlet distribution: $(G(A_1), \ldots, G(A_k)) \sim \text{Dir}(M\, G^\star(A_1), \ldots, M\, G^\star(A_k))$. We denote this by $G \sim \text{DP}(M, G^\star)$. Two parameters need to be specified: a scalar parameter M, and the base measure $G^\star$. The base measure $G^\star$ defines the expectation, $E\{G(B)\} = G^\star(B)$, and M is a precision parameter that defines variance. Properties and definition of the DP are discussed, among others, in Ferguson (1973) or Antoniak (1974). A useful result is the construction by Sethurman (1994). Let δ_x denote a point mass at x. Any $G \sim \text{DP}(M, G^\star)$ can be represented as $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_h}(\cdot)$ with

$$\mu_h \overset{i.i.d.}{\sim} G^\star, \text{ and } w_h = U_h \prod_{j<h}(1 - U_j) \text{ with } U_h \overset{i.i.d}{\sim} \text{Beta}(1, M). \tag{3}$$

In words, realizations of the DP are almost surely discrete. The locations μ_h of the point masses are a sample from $G^\star$, and the random weights w_h are generated by a "stick-breaking" procedure. The almost sure discreteness is inappropriate in many applications. A simple extension to remove the constraint to discrete measures is to introduce an additional convolution, representing a random probability measure F as

$$F(x) = \int f(x|\theta)\, \mathrm{d}G(\theta) \qquad \text{with} \qquad G \sim \mathrm{DP}(M, G^\star). \tag{4}$$

Such models are known as DP mixtures (MDP) (Escobar, 1988; MacEachern, 1994; Escobar and West, 1995). Using a Gaussian kernel, $f(x|\mu, S) = \phi_{\mu,S}(x) \propto \exp[-(x-\mu)^T S^{-1}(x-\mu)/2]$, and mixing with respect to $\theta = (\mu, S)$ we obtain density estimates resembling traditional kernel density estimation. See, for example, MacEachern and Müller (2000) for a recent review of MDP models and posterior simulation methods.

We use DP mixture models as in (4) to define prior models for f_0 and f_1. Let $\mathrm{N}(x;\ m, S)$ denote a normal probability density function for the random variable x with moments (m, S). We assume

$$f_j(z) = \int \mathrm{N}(z; \mu, \sigma^2)\, \mathrm{d}G_j(\mu) \text{ and } G_j \sim \mathrm{DP}(M, G_j^\star), \text{ for } j = 0, 1. \tag{5}$$

Using the stick-breaking representation (3), we can write model (5) equivalently as

$$f_0(z) = \sum_{h=1}^{\infty} \omega_h \mathrm{N}(z; \mu_h, \sigma^2) \tag{6}$$

with locations μ_h and weights ω_h generated by the stick-breaking prior (3). An analogous representation holds for f_1. Thus, similar to McLachlan et al. (2002) and Pan et al. (2002), we represent the unknown densities f_0 and f_1 as mixtures of normals. For the base measures $G_j^\star$ we use

$$G_0^\star = \mathrm{N}(b, \sigma_b^2) \text{ and } G_1^\star = 0.5\, \mathrm{N}(-b_1, \sigma_{b_1}^2) + 0.5\, \mathrm{N}(b_1, \sigma_{b_1}^2).$$

The base measure for the null scores is unimodal, centered at zero. The base measure for scores from differentially expressed genes is symmetric bimodal, reflecting the prior belief that differential expression (on the log scale) in either direction is equally likely.

For notational convenience, we relabel the data as $z_i, i = 1, \cdots, n$, for the null sample Z^{null} and $z_i, i = n+1, \cdots, 2n$, for the mixed sample Z^{mix}. Let $f = p_0\, f_0 + (1 - p_0)\, f_1$ denote the sampling distribution for the mixed sample. For the null sample, the sampling distribution is f_0 itself, without the additional mixture in f. In summary, the likelihood is

$$p(z_1, \cdots, z_{2n} \mid f_0, f_1, p_0) = \prod_{i=1}^{n} f_0(z_i) \prod_{i=n+1}^{2n} f(z_i). \tag{7}$$

We complete the model with a prior probability model for p_0 and the parameters of the base measures G_0^* and G_1^*, using fixed hyperparameters m, m_1, τ^2, τ_1^2, α_σ, β_σ, α_τ, β_τ, α_1, and β_1. We assume a uniform prior $p_0 \sim \text{Unif}(0.05, 1)$, conjugate normal priors on hyperparameters b and b_1, $b \sim \text{N}(m, \tau^2)$, and $b_1 \sim \text{N}(m_1, \tau_1^2)$, and inverse gamma priors on the variance parameters, $\sigma^2 \sim \text{IG}(\alpha_\sigma, \beta_\sigma)$, $\tau^2 \sim \text{IG}(\alpha_\tau, \beta_\tau)$, and $\tau_1^2 \sim \text{IG}(\alpha_1, \beta_1)$. The total mass parameter M in the DP priors is fixed as $M = 1$.

4 POSTERIOR INFERENCE

Posterior inference in the proposed model is carried out using MCMC simulation (Tierney, 1994). Implementation is greatly simplified by two computational devices. Firstly, as usual with DP models, posterior simulation is based on the the marginal posterior, after marginalizing with respect to the unknown random functions f_0 and f_1. See, for example, MacEachern (1998). In other words, we do not represent the actual random functions f_0 and f_1. Later, in section 4.2 we discuss an algorithm that allows us to add inference on f_0 and f_1 in a straightforward manner. The second computational strategy that simplifies implementation is related to the mixtures appearing at various levels of the proposed model. One mixture appears in equation (5) when we construct the DP mixture of normal models for f_0 and f_1. A second mixture appears in the representation of the sampling distribution f for the mixed sample in equation (1). MCMC in mixture models usually proceeds by deconvoluting the mixtures via the introduction of latent variables (Diebolt and Robert, 1994; Robert, 1996). We will follow the same strategy here. The resulting MCMC scheme is no more difficult than posterior simulation in standard MDP models with DP mixtures of normals, as described, e.g. in MacEachern (1998). In fact, the only difference is that minor modifications are required in calculating the resampling probabilities for some of the indicators. We elucidate these details below.

4.1 Markov Chain Monte Carlo Simulation

Posterior simulation is implemented by a Gibbs sampling scheme, iterating over draws from the complete conditional posterior distributions (Tierney, 1994). For the construction of the Gibbs sampler, it is convenient to consider an equivalent representation of the involved mixtures as hierarchical models. The mixtures in (1) and (5) are replaced by a hierarchical model

$$z_i \sim \mathrm{N}(\mu_i, \sigma^2) \text{ and } \mu_i \sim \begin{cases} G_0 & \text{if } r_i = 0, \\ G_1 & \text{if } r_i = 1; \end{cases}$$

with latent indicators $r_i \in \{0, 1\}$ defined by

$$Pr(r_i = 0) = \begin{cases} 1 & \text{for } i = 1, \cdots, n, \\ p_0 & \text{for } i = n+1, \cdots, 2n. \end{cases} \tag{8}$$

The latent variables μ_i break the DP mixtures assumed for f_0 and f_1. The latent indicators r_i break the additional mixture implied in the definition of f as $f = p_0 f_0 + (1 - p_0) f_1$. MCMC posterior simulation proceeds as usual in DP mixture models, with a slightly modified expression for the conditional posterior probabilities used to re-sample the latent μ_i. The key observation when considering the complete conditional posterior for μ_i is that the latent μ_i corresponding to the non-affected sample points $z_i, i = 1, \cdots, n$, can only have ties with other μ_j's that either correspond to other non-affected sample points, $j \in \{1, \ldots, n\}$, or are imputed as arising from f_0, i.e., $j \in \{n+1, \ldots, 2n\}$ and $r_j = 0$. However, $\mu_i, i = n+1, \cdots, 2n$, corresponding to sample points arising from the mixture can be matched with any other $\mu_j, j \neq i$. Let $g_0(\mu) \propto \mathrm{N}(z_i; \mu, \sigma^2)\, G_0^\star(\mu)$, $c_0 = \int \mathrm{N}(z_i; \mu, \sigma^2)\, G_0^\star(\mu)\, \mathrm{d}\mu$, and analogously for $g_1(\mu)$ and c_1. Below we write "$\cdots$" in the conditioning set to indicate the data and all other parameters except the parameter before the conditioning bar. For $i = 1, \cdots, n$, we find

$$(\mu_i | \cdots) = \begin{cases} \mu_j, j \neq i \text{ and } r_j = 0 & \text{with pr. } c\, \mathrm{N}(z_i; \mu_j, \sigma^2), \\ \sim g_0(\mu_i) & \text{with pr. } c\, c_0. \end{cases}$$

Here c is the common proportionality constant to ensure that the probabilities add up to one.

Let $n_0^- = \#\{h : h \neq i \text{ and } r_h = 0\}$ denote the number of data points different from z_i with r indicator equal 0, and analogously for n_1^-. For $i = n+1, \cdots, 2n$, we jointly update μ_i and r_i with

$$(\mu_i, r_i | \cdots) = \begin{cases} (\mu_j, 0), j \neq i, r_j = 0 & \text{with pr. } \gamma\, p_0\, \frac{1}{M+n_0^-}\, \mathrm{N}(z_i; \mu_j, \sigma^2), \\ \\ (\mu_j, 1), j \neq i, r_j = 1 & \text{with pr. } \gamma\, p_1\, \frac{1}{M+n_1^-}\, \mathrm{N}(z_i; \mu_j, \sigma^2), \\ \\ \mu_i \sim g_0(\mu_i) \text{ and } r_i = 0 & \text{with pr. } \gamma\, p_0\, \frac{M}{M+n_0^-}\, c_0, \\ \\ \mu_i \sim g_1(\mu_i) \text{ and } r_i = 1 & \text{with pr. } \gamma\, p_1\, \frac{M}{M+n_1^-}\, c_1. \end{cases}$$

Again, γ denotes the common proportionality constant. Actual implementation is further simplified by keeping track of a set of unique μ values $\{\mu_j^\star, j = 1, \cdots, k\}$ and corresponding indicators $\{r_j^\star, j = 1, \cdots, k\}$.

The remaining steps of the Gibbs sampler generate $p_0, \sigma^2, b, \tau^2, b_1$ and τ_1^2 from the respective complete conditional posterior distributions. Using the conjugate hyperpriors defined earlier, the complete conditional posterior distributions are a beta distribution for p_0, inverse gamma distributions for the variance parameters, and normal distributions for the location parameters b and b_1.

4.2 Inference on f_0 and f_1

The MCMC outlined in section 4.1 was greatly simplified by marginalizing with respect to the unknown distributions f_0 and f_1. However, the final goal of our analysis is inference about $P_1 = (1 - p_0)\, f_1/f$. Posterior inference on P_1 requires discussion of the posterior on f_0 and f_1. In general, inference on the unknown distribution in DP mixture models is challenging. See Gelfand and Kottas (2002) for a discussion. However, some important simplifications are possible. Let Y denote the observed data. The posterior means, $E(f_0|Y)$ and $E(f_1|Y)$, can be shown to be equivalent to the posterior predictive distribution in the MDP model. We exploit this to evaluate posterior estimates for f_0 and f_1. Using full conditional posterior distributions that are already evaluated in the course of the MCMC simulation we can further simplify the computation by using an ergodic average of these conditional predictive distributions. This allows computationally efficient evaluation of $E(f_0|Y)$ and $E(f_1|Y)$. However, for the desired full posterior inference about the probability of differential expression P_1 more information is required. We need posterior samples from the posterior $p(f_j|Y)$, $j = 0, 1$, on the unknown densities themselves. This is difficult in general. Below we describe a computational algorithm that allows easy (approximate) simulation in the context of the proposed model.

First, using f_0 as example, we note that the posterior mean $E(f_0|Y)$ is equal to the posterior predictive distribution. Let z_{2n+1} denote a new Z^{null} score. We find

$$p(z_{2n+1} \mid Y) = E[p(z_{2n+1} \mid Y, f_0) \mid Y] = E[f_0(z_{2n+1}) \mid Y].$$

Let θ denote the vector of all model parameters, and let $\theta^{(i)}$ denote the parameters imputed after i iterations of the MCMC simulation. We evaluate $p(z_{2n+1} \mid Y)$ as

$$p(z_{2n+1}|Y) = E[p(z_{2n+1} \mid Y, \theta) \mid Y] \approx \frac{1}{T} \sum_{i=1}^{T} p(z_{2n+1} \mid \theta^{(i)}, Y) = \frac{1}{T} \sum_{i=1}^{T} p(z_{2n+1} \mid \theta^{(i)}).$$

The terms in the last average are easily computed. Recall that $\{\mu_j^*,\ j=1,\ldots,k\}$ are the unique values of the latent variables μ_i, and r_j^* are the corresponding indicators r_i. Without loss of generality assume that the μ_j^* are arranged with $r_j^*=0$ for $j=1,\ldots,k_0$ and $r_j^*=1$ for $j=k_0+1,\ldots,k$. Let $n_j=\#\{i:\ \mu_i=\mu_j^*\}$ denote the number of μ_i equal to μ_j^* and let $N_0=\#\{i:\ r_i=0\}$ denote the number of $r_i=0$. We use a superindex $^{(i)}$ to identify the imputed parameter values after i iterations of the MCMC simulation. We find

$$p(z_{2n+1}\mid\theta^{(i)})\propto\sum_{j=1}^{k_0^{(i)}} n_j^{(i)}\mathrm{N}\left(z_{2n+1};\mu_j^{*(i)},\sigma^{2(i)}\right)+M\mathrm{N}\left(z_{2n+1};b^{(i)},\sigma^{2(i)}+\sigma^{2(i)}\right). \tag{9}$$

Uncertainty in f_0 is illustrated through posterior draws of f_0. For the following argument we consider augmenting the imputed parameter vector $\theta^{(i)}$ with the random distribution G_0 defined in (5). Given $\theta^{(i)}$ the conditional posterior for G_0 is a DP with updated parameters,

$$(G_0|\theta^{(i)},Y)\sim\mathrm{DP}(H_0,M+N_0^{(i)})\text{ with }H_0\propto M^{(i)}G_0^*+\sum_{j=1}^{k_0^{(i)}}n_j^{(i)}\delta_{\mu_j^{*(i)}}. \tag{10}$$

The large total mass parameter $M+N_0^{(i)}$ implies that the random measure G_0 is close to the conditional expectation H_0, the DP base measure in (10). We exploit this to approximate a posterior draw $G_0\sim p(G_0\mid\theta^{(i)},Y)$ as $G_0\approx H_0$, and thus a posterior draw for f_0 as $\int\mathrm{N}(\mu,S^{(i)})\,\mathrm{d}H_0(\mu)$, i.e.,

$$f_0(z)\propto M\int\mathrm{N}(z;\,\mu,\sigma^{2(i)})\,\mathrm{d}G_0^*(\mu)+\sum_{j=1}^{k_0^{(i)}}n_j^{(i)}\mathrm{N}(z;\,\mu_j^{*(i)},\sigma^{2(i)}).$$

The latter is simply the predictive distribution conditional on $\theta^{(i)}$ in (9). The same mechanism can be applied to obtain samples from $f_1(z_{2n+1}|Y)$.

5 SIMULATION STUDY AND APPLICATION

In section 5.1, we perform a small simulation study to illustrate the proposed approach. Results are compared with the known true parameter values in the simulation. In section 5.2, we analyze a colon data set from Alon (1999), and compare the results under the proposed nonparametric Bayesian model with inference obtained from the empirical Bayes approach.

5.1 Simulation Study

We simulate a sample of $n=400$ gene expression scores Z_i^{null}, $i=1,\ldots,n$, from $f_0=\mathrm{N}(0,1)$, and a sample Z_i^{mix}, $i=1,\ldots,n$, from $f=p_0\,f_0+(1-p_0)\,f_1$ with $f_1=0.5\,\mathrm{N}(-2,1)+0.5\,\mathrm{N}(2,1)$ and $p_0=0.8$.

Bayesian nonparametric inference is set up according to the proposed approach. We fixed the hyperparameters as $m = 0$, $m_1 = 1$, $\alpha_\sigma = 2$, $\beta_\sigma = 1$, $\alpha_\tau = 1$, $\beta_\tau = 0.2$, $\alpha_1 = 1$, and $\beta_1 = 0.2$. We summarize results here. We will use Y to generically denote the observed data.

Figure 1 shows the posterior mean curves $E(f_0|Y)$, $E(f_1|Y)$ and $E(f|Y)$, together with the true distributions used in the simulation. Posterior inference correctly recovers the true curves. Not surprisingly, the bias for f_1 is larger than for f_0 and f. While f_0 and f are easily estimated by the relatively large samples, the data gives only indirect evidence for f_1, implied by the deconvolution of (1). Figure 2 illustrates the uncertainty about the estimated

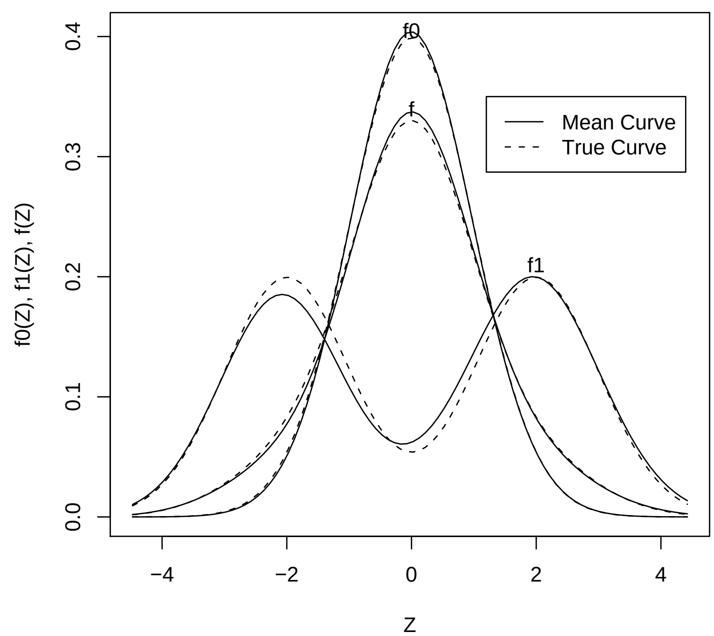


Figure 1: Posterior mean curves $E(f_0|Y)$, $E(f_1|Y)$ and $E(f|Y)$ (solid curves) and true distributions (dashed curves). The higher bias in estimating f_1 reflects that the data includes only indirect information about f_1. Inference has to be derived by deconvoluting the mixture $f = p_0\, f_0 + (1 - p_0)\, f_1$.

distributions.

Let $\overline{P}_1(z_i)$ denote the marginal posterior probability $E\{P_1(z_i|f_0, f_1, p_0)|Y\}$ for every gene, $i = 1, \ldots, n$. The structure of the proposed model implies that this marginal posterior probability of differential expression depends on the gene only through the observed score z_i, making it meaningful to consider $\overline{P}_1(z)$ as a function of z, as shown in Figure 3. For

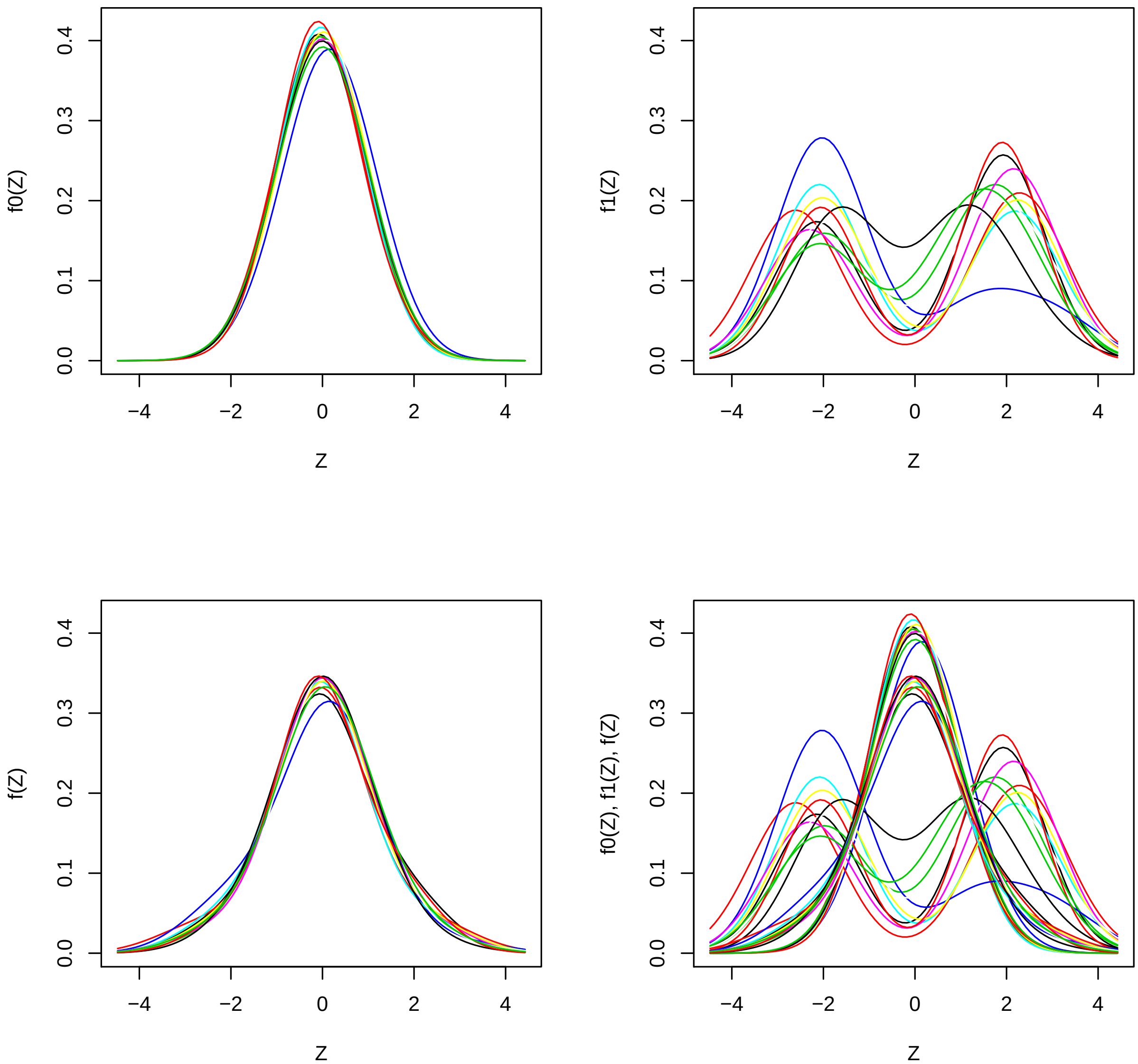


Figure 2: The four panels illustrate posterior uncertainty in f_0, f_1 and f. Panel (a) through (c) plot 10 draws from $f_0 \sim p(f_0|Y)$, $f_1 \sim p(f_1|Y)$ and $f \sim p(f|Y)$, respectively. To allow easier comparison, panel (d) combines plots (a) through (c). Notice the high uncertainty in f_1.

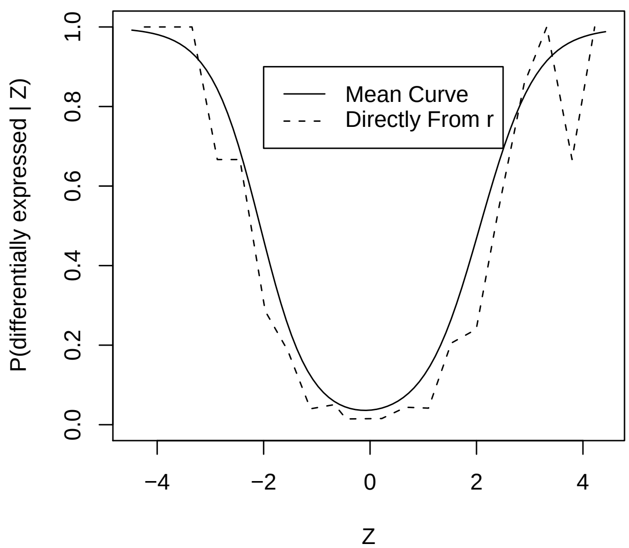


Figure 3: Posterior mean curve $\overline{P}_1(z) = E\{P_1(z|f_0, f_1, p_0) \mid Y\}$ for $P_1 = (1 - p_0)f_1/f$ (solid curve). For comparison the dashed curve plots average true indicators r_i (binned over z scores). The r_i indicators are defined as $r_i = 1$ if $z_i \sim f_1$ and $r_i = 0$ if $z_i \sim f_0$.

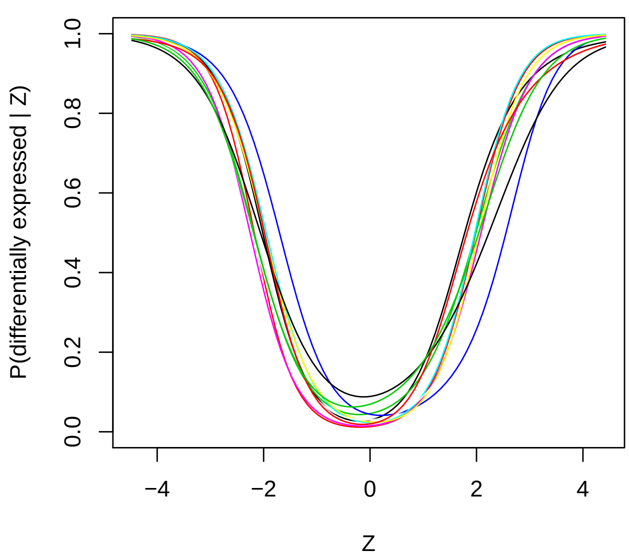


Figure 4: Posterior uncertainty about P_1. The plot shows 10 draws of $p(P_1|Y)$ to illustrate posterior uncertainty about P_1.

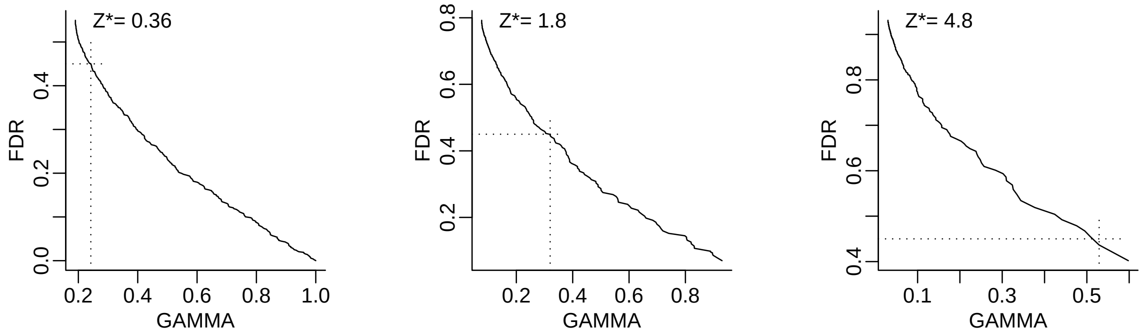
Table 1: Classification into differentially and non-differentially expressed genes. The table reports marginal posterior probabilities of differential expression $\overline{P}_1(z)$ across the three experiments (rows) and across z scores (columns). Posterior probabilities corresponding to rejection are highlighted in bold face. The rejection region is defined by a bound on the false discovery rate, $\overline{\text{FDR}} \leq \alpha$ (see the text for details). The first column reports the true value p_0 used in the simulations. Note how posterior inference automatically adjusts for the higher level of noise in the experiments with larger p_0.

	Observed z scores										
p_0	-5.00	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00	4.00	5.00
0.4	**1.00**	**1.00**	**0.98**	**0.87**	**0.46**	0.19	**0.43**	**0.85**	**0.98**	**1.00**	**1.00**
0.8	**0.94**	**0.90**	**0.75**	**0.41**	0.14	0.07	0.13	**0.44**	**0.81**	**0.93**	**0.96**
0.95	**0.46**	0.42	0.27	0.11	0.05	0.03	0.04	0.10	0.28	0.43	**0.50**

comparison, Figure 3 also shows the true proportion of differentially expressed genes for each z. Figure 4 shows the uncertainty in P_1.

The estimated posterior probabilities of differential expression can be used to carry out the multiple comparison to classify genes into affected and unaffected by the condition of interest. If we assume an underlying $(0, 1, c)$ hypothesis testing loss we would declare all genes with $\overline{P}_1(z_i) > c/(1 + c)$ as differentially expressed. Table 1, second row, reports the marginal posterior probabilities $\overline{P}_1(z)$ over a range of z values. In this example, using $\overline{P}_1 > 0.5$ (i.e., $c = 1$) leads to classifying genes with $|z| > 2.2$ as differentially expressed. The marginal posterior probabilities appropriately adjust for the observed level of noise. We illustrate this by considering two additional simulations with lower and higher proportions of non-differentially expressed genes, using (true) $p_0 = 0.4$ and with $p_0 = 0.95$, respectively. For $p_0 = 0.4$ the cutoff shifts to $|z| > 1.2$. For higher levels of noise, $p_0 = 0.95$, the cutoff shifts even further to $|z| > 5.2$ (see Table 1). Alternatively, the shifting cutoff can be thought of as a Bayesian adjustment for multiplicities. With higher p_0 there is an increasingly larger number of false comparisons. Posterior probabilities appropriately adjust.

A useful generalization of frequentist type-I error rates to multiple hypothesis testing is the false discovery rate (FDR) introduced in Benjamini and Hochberg (2002). Let δ_i denote an indicator for rejecting the i-th comparison, i.e., flagging gene i as differentially expressed. Recall from equation (8) the definition of r_i as indicators for true differential expression of



$p_0 = 0.4$ $p_0 = 0.8$ $p_0 = 0.95$

Figure 5: Posterior expected false discovery rate $\overline{\text{FDR}} = E(\text{FDR} \mid Y)$ for each of the three simulation experiments with low, medium and high true proportion negatives (i.e., non-differentially expressed genes). Assuming rejection regions of the type $\overline{P}_1(z_i) > \gamma$, the figures show $\overline{\text{FDR}}$ as a function of the cutoff γ. The dotted lines indicate the smallest cutoff γ^* to achieve $\overline{\text{FDR}} \le \alpha$ for $\alpha = 0.45$. In each figure the legend indicates the corresponding bound on $|z_i|$. See the text for more explanation.

gene i. FDR is defined as

$$\text{FDR} = \frac{\sum(1 - r_i)\delta_i}{\sum \delta_i},$$

the fraction of false rejections, relative to the total number of rejections. Applications of FDR to microarray analysis are discussed by Storey and Tibshirani (2003). Extensions are discussed by Genovese and Wasserman (2002, 2003), who also introduce the definition of posterior expected FDR as $\overline{\text{FDR}} = E(\text{FDR} \mid Y) = [\sum(1 - \overline{P}_1(z_i))\delta_i]/\sum \delta_i$. We consider decision rules that classify a gene as differentially expressed if $\overline{P}_1(z_i) > \gamma^*$. In analogy to classical hypothesis testing, we fix γ^* as the minimum value that achieves a certain pre-set false discovery rate, $\overline{\text{FDR}} \le \alpha$. It can be shown (Müller et al., 2002) that under several loss functions that combine false negative and false discovery counts and/or rates the optimal decision rule is of this form. Figure 5 shows how the cutoff is obtained for three simulations with true $p_0 = 0.4, 0.8$ and 0.95. We use $\alpha = 0.45$ (Since $\max \overline{P}_1(z_i) = 0.6$ for the third simulation, it is impossible to achieve any $\overline{\text{FDR}} \le 0.4$). Genes with $\overline{\text{FDR}}$ beyond the cutoff are highlighted in bold face in Table 1. As before, the rule adjusts to increasing levels of noise by defining increasingly more conservative cutoffs.

5.2 Gene Expression Profiling of Colon Cancer

We analyze the data set reported in Alon et al. (1999). The data format was described in Section 2. We compare inference under the proposed nonparametric Bayesian model with the empirical Bayes estimate (EMBA) discussed earlier.

The EMBA estimates posterior probabilities of differential expression $P_1(z|f_0, f_1, p_0)$ by substituting point estimates for f_0/f and p_0. The estimate for p_0 is derived as a bound on possible values for p_0. In this data set we find $\widehat{p}_0 = 0.39$. In contrast, Figure 6 shows the marginal posterior distribution $p(p_0|Y)$. The bound $\widehat{p}_0$ is far out in the tail of the posterior distribution, indicating that $\widehat{p}_0$ might lead to very conservative estimates for P_0 and P_1 (by underestimating P_1). Figures 7 through 9 show comparative inference for f_1, f_0, f and P_1. As expected, the posterior mean curve $\overline{P}_1(z)$ is estimated lower under the EMBA than under the proposed nonparametric Bayesian approach. Figure 8 summarizes the posterior distributions $p(f_0|Y)$, $p(f_1|Y)$ and $p(f|Y)$.

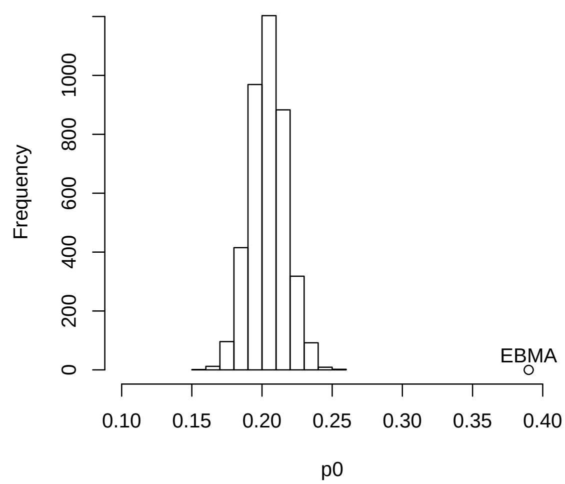


Figure 6: Analysis of colon cancer data. The histogram depicts the marginal posterior $p(p_0|Y)$ from the nonparametric Bayesian model. Compare with the point estimate $\widehat{p}_0 = 0.39$ under the empirical Bayes method.

The simulation-based implementation of posterior inference allows investigators to compute posterior probabilities for any event of interest under the posterior or posterior predictive distribution. The relevant probabilities are computed as appropriate ergodic averages under the proposed MCMC simulation. For example, it is possible to make joint inference

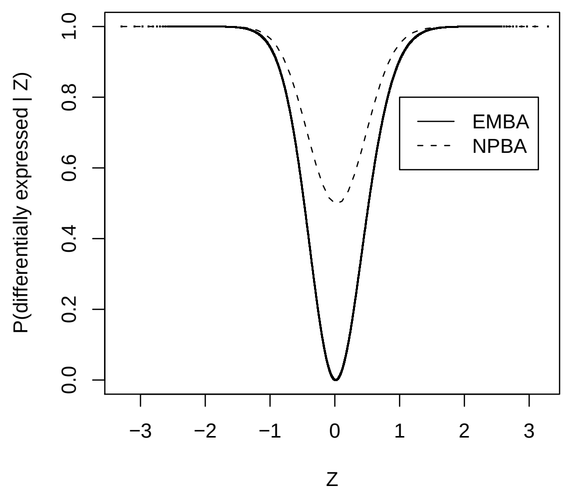


Figure 7: Estimated curve $\overline{P}_1(z)$ under both, the EMBA and the proposed nonparametric Bayesian approach, for the Alon colon cancer data. The lower estimate under the EMBA is a direct consequence of the overestimated $\widehat{p}_0$.

about a sub-group of genes being differentially expressed. Posterior simulation keeps track of the indicators r_i for all the genes. Evaluating the joint probability of a sub-group of genes being differentially expressed amounts to counting how often $r_i = 1$ for all genes in the subset of interest. For example, the probability of six smooth muscle genes (J02854, T60155, X12369, M63391, D31885, and X74295) or six ribosomal genes (T79152, T95018, T57633, T62947, T52185 and T57630) being joint differentially expressed is 0.61 and 0.22, respectively.

The posterior probabilities for the 2000 genes range from 0.498 to 1.0 corresponding to $|z_i|$ between 0.035 and 3.290, respectively. To estimate the number n_d of differentially expressed genes, the user can consider the ergodic average of the number of indicators r_i that equal unity. The marginal posterior distribution $p(n_d \mid Y)$ is shown in Figure 10. However, often in practice, statistical significance does not necessarily imply strong biological significance. Thus, investigators may wish to calibrate between a desired $\overline{\text{FDR}}$ and the number of significant genes, as discussed in Section 5.1. For fixed values of α ranging from 0.001 to 0.2, threshold values on the observed scores z_i, denoted as Z^*, and the smallest cutoff γ^* to achieve $\overline{\text{FDR}} \leq \alpha$ are depicted in Table 2, along with the estimated number of significant genes. As $|Z^*|$ increases, the number of genes identified as significant by NPBA decreases along with a decreasing FDR. Investigators can use Table 2 to calibrate the results

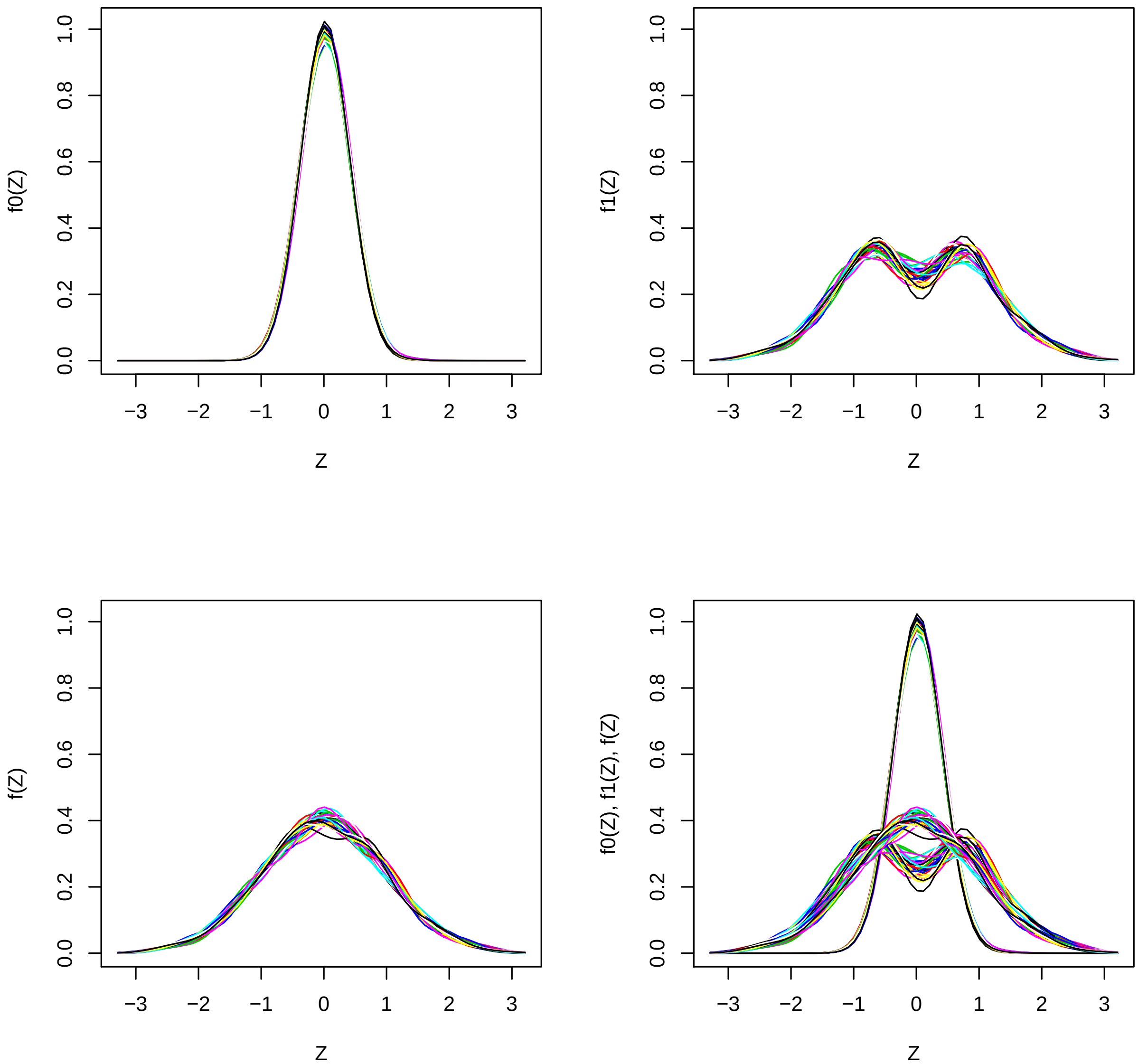


Figure 8: Posterior distributions for the unknown densities (Alon colon cancer data). The first three panels summarize the posterior distributions on f_0, f_1 and f, respectively, by showing 10 draws from $p(f_0|Y)$, $p(f_1|Y)$ and $p(f|Y)$, respectively. For easier comparison the fourth panel combines the plots from the first three panels into one figure.

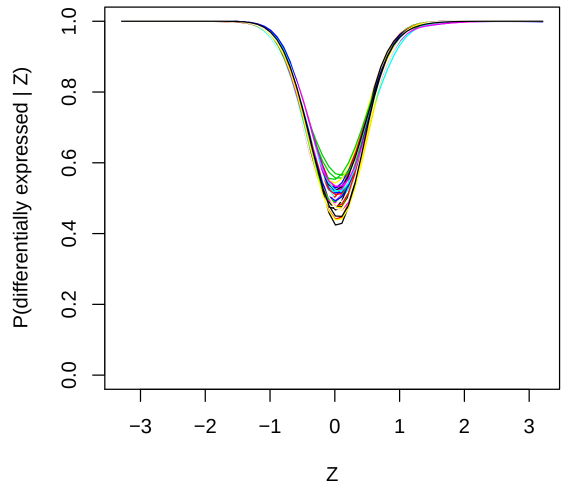


Figure 9: Posterior uncertainty for P_1 (Alon colon cancer data). The figure shows 10 draws from $p\{P_1|Y\}$. Compare with the posterior mean curve shown in Figure 7.

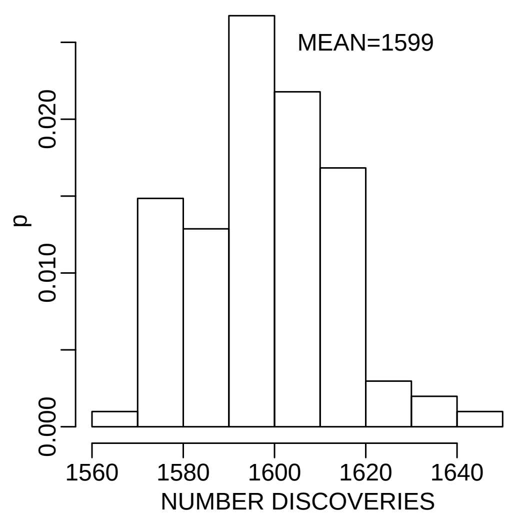


Figure 10: Posterior $p(n_d \mid Y)$ for the number of discoveries.

Table 2: Estimated number n_d of significant differentially expressed genes identified by NPBA for different values of FDR (Alon colon cancer data). Z^* denotes a threshold value for the observed z-scores; γ^* denotes the smallest cutoff to achieve $\overline{\text{FDR}} \leq \alpha$.

$\|Z^*\|$	γ^*	$\overline{\text{FDR}}$	$\widehat{n_d}$
0.008	0.500	0.200	1938
0.241	0.549	0.150	1667
0.360	0.655	0.100	1393
0.580	0.803	0.050	1083
1.000	0.967	0.010	579
1.200	0.990	0.005	422
1.302	0.995	0.001	346

that give the best biological interpretation.

6 CONCLUSION

We described a straightforward, albeit computer-intensive, model-based nonparametric Bayesian approach as an effective framework for studying the relative changes in gene expression for a large number of genes under two different conditions. It uses a variation of traditional Dirichlet process mixtures to model the population of affected and unaffected genes, thereby enabling full probabilistic posterior inference to be carried out through a deconvolution process of the mixtures in the MCMC simulation. Compared to the empirical Bayes approach of Efron et al. (2001) based on plug-in values of the density functions under the different conditions and corresponding prior probabilities for differential expression, we demonstrated via a simulation study and a colon cancer data set, that our method can avoid the bias inherent in the former when estimating the posterior probability of differential expression. We also addressed the multiple testing issues that arise when dealing with a large number of simultaneous tests (genes). Using an appropriate definition of posterior expected false discovery rate, a close connection can be established with the final estimates of the posterior probabilities that automatically adjust for the proportion of noisy data, or equivalently, the true number of differentially expressing genes. A strength of the approach we have presented is that the rejection regions can be adaptively chosen to accommodate a pre-specified and biologically meaningful FDR chosen by the investigator; thus an appropriate threshold value can be directly calculated for the summary expression score to declare significance.

A critical assumption is that gene expression scores are independently identically distributed where the important aspect of the variation of gene expression across tissue samples can be captured sufficiently well by a binary variable (affected versus unaffected). While it is obvious that complex interactions between expression levels of several genes are likely to be present in practice (for example, as a result of carcinogenic pathways for cancer data), the underlying independence approximation is still useful to determine whether expression level differences are significant solely on a gene-by-gene basis. The approach described here can be extended to the exploration of gene interactions. Consider the simple subset of just two interacting genes. A natural extension of our approach is to choose a mixture of an appropriate bivariate distribution for the dual non-differentially expressed components, and several other mutually exclusive bivariate distributions for the differentially expressed components. Our modeling framework allows for other kinds of elaboration including the combination of information across microarray technologies and gene-specific sensitivities (that may induce non-linearities in expression levels) due to different RNA preparations, different dyes, and different growth conditions. Further research into alternative prior structures to capture different sources of variation and potential interactions between genes will provide more precise estimates of differential gene expression and more accurate assessments of significant changes, thus reducing errors in downstream tasks, such as classification and cluster analysis.

Supplementary information and software updates are available at
`http://odin.mdacc.tmc.edu/~kim`

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I. S., Rosenwal, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G., Moore, T., Hudson Jr., J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, 403, 503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," in *Proceedings of the National Academy of Sciences*, vol. 96, pp. 6745–6750.

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2, 1152–1174.

Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and W., Z. (2001), "Identifying differentially expressed genes in cDNA microarray experiments," *Journal Computational Biology*, 8, 639–659.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., et al. (2000), "Tissue classification with gene expression profiles," *Journal Computational Biology*, 7, 559–584.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999), "Clustering gene expression patterns," *Journal Computational Biology*, 6, 281–297.

Benjamini, Y. and Hochberg, Y. (2002), "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

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*, 2(4), 364–374.

DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996), "Use of a cDNA microarray to analyze gene expression patterns in human cancer," *Nature Genetics*, 14, 457–460.

Diebolt, J. and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society, Series B*, 56, 163–175.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Eisen, M. B., Spellmann, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns," in *Proceedings of the National Academy of Sciences*, vol. 95, pp. 14863–14868.

Escobar, M. (1988), "Estimating the means of several normal populations by estimating the distribution of the means," Ph.D. thesis, Yale University.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Gelfand, A. E. and Kottas, A. (2002), "A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289–305.

Genovese, C. and Wasserman, L. (2002), "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society B*, 64, 499–518.

— (2003), *Bayesian Statistics 7*, Oxford: Oxford University Press, chap. Bayesian and Frequentist Multiple Testing, p. to appear.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeck, 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.

Lönnstedt, I. and Speed, T. P. (2002), "Replicated microarray data," *Statistica Sinica*, 12, 31–46.

MacEachern, S. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics: Simulation and Computation*, 23, 727–741.

MacEachern, S. N. and Müller, P. (2000), "Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models," in *Robust Bayesian Analysis*, eds. Ruggeri, F. and Ríos-Insua, D., New York: Springer-Verlag, pp. 295–316.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002), "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, 18, 413–422.

Miles, M. (2001), "Microarrays: lost in a storm of data?" *Nature Reviews Neurosciences*, 2, 441–443.

Moler, E. J., Chow, M. L., and Mian, I. S. (2000), "Analysis of molecular profile data using generative and discriminative methods," *Physiological Genomics*, 4, 109–126.

Müller, P., Parmigiani, G., Robert, C., and Rouseau, J. (2002), "Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays," Tech. rep., University of Texas, M.D. Anderson Cancer Center.

Newton, M. A., Kendziorsky, C. M., Richmond, C. S., R., B. F., and Tsui, K. W. (2001), "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *Journal Computational Biology*, 8, 37–52.

Pan, W., Lin, J., and T., L. C. (2002), "Model-based cluster analysis of microarray gene expression data," *Genome Biology*, 3(2), research009.1–research009.9.

Robert, C. P. (1996), "Mixture of Distributions:Inference and Estimation," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. R. and Spiegelhalter, D. J., London: Chapman & Hall, pp. 441–464.

Rocke, D. M. and Durbin, B. (2001), "A model for measurement error for gene expression arrays," *Journal Computational Biology*.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, 270, 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996), "Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes," *Proceedings of the National Academy of Sciences*, 93, 10614–10619.

Sethurman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Smyth, G. K., Yang, Y. H., and Speed, T. (2002), "Statistical Issues in cDNA Microarray Data Analysis," in *Functional Genomics: Methods and Protocols*, eds. Brownstein, M. J. and Khodursky, A. B., Totowa, NJ: Methods in Molecular Biology series, Humana Press.

Storey, J. S. and Tibshirani, R. (2003), "SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays," in *The analysis of gene expression data: methods and software*, New York: Springer.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *The Annals of Statistics*, 22, 1701–1762.

Tusher, V. G., R., T., and G., C. (2002), "Significance analysis of microarrays applied to the ionizing radiation response," in *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116–5121.

Walker, S., Damien, P., Laud, P., and Smith, A. (1999), "Bayesian nonparametric inference for distributions and related functions (with discussion)," *Journal of the Royal Statistical Society, Series B*, 61, 485–527.

Wu, T. D. (2001), "Analyzing gene expression data from DNA microarrays to identify candidate genes," *Journal of Pathology*, 195(1), 53–65.

Xing, E. P. and Karp, R. M. (2001), "CLIFF: Clustering of high-dimensional Microarray data via iterative feature filtering using normalized cuts," *Bioinformatics*, 17, Suppl 1: S306–S315.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001), "Normalization for cDNA microarray data," in *Microarrays: Optical Technologies and Informatics*, eds. Bittner, M. L., Chen, Y., Dorsel, A. N., and Dougherty, E. R., Bellingham, WA: SPIE, volume 4266 of Proceedings of SPIE.