Published in final edited form as:

J Biopharm Stat. 2010 March ; 20(2): 267–280. doi:10.1080/10543400903572746.

A Semiparametric Bayesian Approach for Estimating the Gene Expression Distribution

Fei Zou, Hanwen Huang, and Joseph G. Ibrahim

Department of Biostatistics University of North Carolina at Chapel Hill Chapel Hill, North Carolina 27599, U.S.A.

Abstract

Gene expression microarrays are powerful tools for global comparison and estimation of gene expression. Many microarray studies have demonstrated biologically plausible results with only a few arrays, leading to a mis-perception that a handful of hybridized arrays can always find something meaningful. From a statistical point of view, it is important to prospectively estimate required sample sizes prior to undertaking a microarray experiment. However, all sample size calculations need to directly or indirectly estimate the unknown distribution of the effect sizes of gene expression intensities. A parametric mixture model has been developed for relating the sample size directly to the false discovery rate (FDR), the most popular multiple comparison control criteria. In this paper, we extend the parametric mixture model and propose a robust semiparametric Dirichlet process mixture model, where the parametric distribution of gene expressions is no longer specified. This analysis is performed in a Bayesian inference framework using Markov chain Monte Carlo. The usefulness of the method is illustrated by simulations and a real murine lung study.

1 Introduction

Gene expression microarrays are powerful tools for global comparison and estimation of gene expression. One of main applications of microarray studies is to detect genes that are differentially expressed under two (or more) conditions. Early approaches had used simple thresholds for ratios of expression estimates under the two conditions (Chen et al 1997), while ordinary *t*-tests and Wilcoxon tests (Dudoit et al 2002 and Troyanskaya et al 2002) have later been used for more rigorous statistical analyses. More sophisticated ideas are further employed in regularized *t*-tests, for example by adding a constant to the variance estimate in the *t* denominator (Tusher et al 2001 and Efron et al 2001).

Many microarray studies have demonstrated biologically plausible results with only a few arrays (Yoon et al 2002), leading to a mis-perception that a handful of hybridized arrays can always find something meaningful. It is plausible that in some cases surprisingly few arrays may be sufficient, and the current dominance of microarray research in cancer has produced an optimistic view of microarray studies that may be unwarranted in other settings, such as examining expression differences among closely related rodent strains. From a statistical point of view, it is important to prospectively estimate required sample sizes prior to undertaking an experiment. Using ANOVA analysis of gene expression, Black and Doerge (2002) propose a parametric approach assuming lognormal or gamma distributions for gene expression intensities. Lee and Whitmore (2002) also describe sample size calculations for the ANOVA model, under several different experimental designs. Pan et al (2002) discuss

sample size calculations by applying a nonparametric statistical method, the normal mixture model approach. All these methods relate power to sample size while controlling for the Type I error. Hu et al (2005) relate the sample size directly to the false discovery rate (FDR), as this criterion is often used for controlling for multiple comparisons.

All sample size calculations need to directly or indirectly estimate the unknown distribution of the effect sizes of the gene expression intensities. In the statistical literature, there are many proposed methods for estimating a distribution parametrically or non-parametrically. Hu et al (2005) employ three-component parametric mixture models to estimate the effect size distribution. Alternatively, the non-parametric methods of Robbins (1955), Laird (1978; 1981), Lindsay (1983), Wu et al (2006), Ruppert et al (2007) and Guan et al (2008) can be used for estimating the effect size distribution. However, these frequentist nonparametric methods assume a (known) fixed number of mixture components. Pan et al (2002) fit a series of normal mixture models with a varying number of mixture components and then use AIC or BIC to pick up the number of mixture components corresponding to the first local minimum of AIC or BIC. In the Bayesian literature, the mixture of Dirichlet process (MDP) model (Ferguson 1973; Blackwell and MacQueen 1973; Berry and Christensen 1979; Escobar and West 1995; Liu 1996; MacEachern 1994; MacEachern and Müller 1998; Kleinman and Ibrahim 1998; Neal 2000; Rasmussen 2000 and Ishwaran and James 2002) have been widely used in nonparametric Bayesian models. The Dirichlet process (DP) provides a nonparametric prior specifications for the parameters of a mixture model and allows the number of mixture components to grow as the size of the data grows. While classical parametric modeling techniques assume that a fixed number of components, the DP approach makes no such assumptions. Tang et al (2007) apply Dirichlet mixtures of beta densities to the distribution of p-values in estimating positive FDR.

This paper extends the three-component parametric mixture models in Hu et al (2005) to the DP mixture model in estimating the effect size distribution of the gene expressions. Our method is described in the next section with details of the Markov chain Monte Carlo (MCMC) steps for fitting the model. Simulation studies and the analysis of real datasets are presented in Section 3. We conclude the paper with some remarks and discussion.

2 Methods

2.1 Notation and assumptions

Here we largely follow the notation in Hu et al (2005). Let $x_{1j,i}$ ($j = 0, \dots, n_1$) and $x_{2j,i}$ ($j = 1, \dots, n_2$) denote the expression levels for a single gene i ($i = 1, \dots, m$) under two conditions. For simplicity, we assume $n_1 = n_2 = n$. For a single gene, we assume that the gene expression estimates are normally distributed within each condition. Let $\mu_{1,i}$ and $\mu_{2,i}$ denote the true mean expressions for gene i under conditions 1 and 2, and $\sigma_{1,i}^2$ and $\sigma_{2,i}^2$ denote the corresponding variances. For testing purposes, the hypotheses are $H_0 : \mu_{1,i} = \mu_{2,i}$ versus $H_1 :$

$\mu_{1,i} \neq \mu_{2,i}$, and we use the statistic
$$t_i = \frac{\sqrt{n}(\bar{x}_{1,i} - \bar{x}_{2,i})}{\sqrt{s_{1,i}^2 + s_{2,i}^2}}, \text{ where}$$

$$\bar{x}_{1,i} = \sum_{j=1}^n \frac{x_{1j,i}}{n}, \bar{x}_{2,i} = \sum_{j=1}^n \frac{x_{2j,i}}{n}, s_{1,i}^2 = \sum_{j=1}^n \frac{(x_{1j,i} - \bar{x}_{1,i})^2}{(n-1)} \text{ and } s_{2,i}^2 = \sum_{j=1}^n \frac{(x_{2j,i} - \bar{x}_{2,i})^2}{(n-1)}.$$
 Then t_i follows the student t -distribution $2n - 2$ degrees of freedom and non-centrality parameter $\sqrt{n}\delta_i$. The effect size

$$\delta_i = \frac{\mu_{1,i} - \mu_{2,i}}{\sqrt{\sigma_{1,i}^2 + \sigma_{2,i}^2}} \quad (1)$$

governs the power to detect departures from H_0 , and has a biological interpretation as a scaled expression difference. We can now restate the hypotheses as $H_0 : \delta_i = 0$ versus $H_1 : \delta_i \neq 0$.

With thousands of genes assayed in a typical microarray experiment, we can treat the δ_i s as realizations from a common CDF, F , represented by a random variable Δ . The distributions of the t_i s and derived p-values depend entirely on F . Because the sign of δ_i has biological importance, F is typically decomposed as the following three component mixture:

$$F(\delta) \sim \pi_0 F_0(\delta) + \pi_1 F_1(\delta) + \pi_2 F_2(\delta). \quad (2)$$

Here π_0 , π_1 and π_2 are the probabilities that Δ is 0, positive or negative respectively, and $\pi_0 + \pi_1 + \pi_2 = 1$. $F_0(\delta) = I_0(\delta)$ is the cumulative distribution function (CDF) of the random variable who has a point mass at 0, whereas F_1 and F_2 are the conditional CDFs for a positive and negative Δ . Equation (2) encompasses a large number of practical situations, where F_1 and F_2 can be discrete or continuous.

Given δ , the conditional distribution of the p -value of a gene is

$$P(P < p_0 | \delta) = 1 - G(t_{p_0/2}; \sqrt{n}\delta) + G(-t_{p_0/2}; \sqrt{n}\delta), \quad (3)$$

where $G(t; \sqrt{n}\delta)$ is the CDF of the non-central student t -distribution with $2n - 2$ degrees of freedom and noncentrality parameter $\sqrt{n}\delta$ and t_α is the $1 - \alpha$ quantile of the central t distribution with $2n - 2$ degrees of freedom. The marginal distribution of P can be written as

$$\begin{aligned} P(P < p_0) &= \int_{\delta} P(P < p_0 | \delta) dF(\delta) \\ &= \pi_0 p_0 + \pi_1 \int_{\delta > 0} P(P < p_0 | \delta) dF_1(\delta) + \pi_2 \int_{\delta < 0} P(P < p_0 | \delta) dF_2(\delta). \end{aligned} \quad (4)$$

Thus, the relationship between the FDR and p_0 is

$$FDR(p_0) = \frac{P(P < p_0, \delta = 0)}{P(P < p_0)} = \frac{\pi_0 p_0}{P(P < p_0)}, \quad (5)$$

which depends on n through (3). The relationship (5) depends on the underlying distribution of δ . For known π_0 , π_1 , π_2 , F_1 and F_2 , through (5), we can estimate the required sample size n in order to achieve a desired FDR value for a given p-value threshold p_0 . Therefore, the sample size calculation reduces to the estimation of π_0 , π_1 , π_2 , F_1 and F_2 . For such purposes, Hu et al (2005) propose a parametric method by considering three situations: (1) F_1 (and F_2) has a point mass at (positive) a_1 (and (negative) a_2); (2) the exponential mixture model where F_1 and F_2 follow positive and negative exponential distributions, and (3) the normal mixture model where F_1 and F_2 are two normal densities truncated at 0. However, for real datasets, the above parametric assumptions are likely to be violated. Below, we relax the parametric assumptions and estimate F nonparametrically via the MDP model.

2.2 Mixture Dirichlet Process Model

In this section, we propose the MDP model for δ , derive the full conditional distributions, and demonstrate how to implement the Gibbs sampling. Parametric Bayesian modeling typically assumes that δ follows a prior parametric distribution with known hyperparameters Ψ_0 . That is,

$$\begin{aligned} t_i|\delta_i &\sim t(\delta_i, d), \\ \delta_i|\Psi_0 &\sim h(\delta_i|\Psi_0), \end{aligned} \quad (6)$$

where $t(\delta, d)$ is the non-central student t -distribution with d degrees of freedom and non-centrality parameter δ and $h(\delta|\Psi_0)$ is a known distribution. The MDP model removes the distributional assumption on the δ_i s and assumes that they follow an unspecified distribution G , which itself follows a Dirichlet process prior:

$$\begin{aligned} t_i|\delta_i &\sim t(\delta_i, d), \\ \delta_i &\sim G, \\ G|M &\sim DP(M \cdot G_0). \end{aligned} \quad (7)$$

The parameters of the Dirichlet process, $DP(M \cdot G_0)$ are G_0 and a positive scalar M . In practice, the base measure G_0 is set to a parametric distribution that (hopefully) approximates the true nonparametric shape of G . The scalar M reflects our prior belief about how similar the nonparametric distribution G is to the base measure G_0 . As $M \rightarrow \infty$, $G \rightarrow G_0$, the MDP model leads to the fully parametric case and the base measure is the prior distribution for δ .

Note that the t -distribution can be obtained as a scale mixture of normal distributions as follows:

$$\begin{aligned} t_i|\delta_i, \tau_i &\sim N(\delta_i, \tau_i^{-1}) \\ \tau_i &\sim \text{Gamma}\left(\frac{d}{2}, \frac{d}{2}\right), \end{aligned} \quad (8)$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 and $\text{Gamma}(a, b)$ denotes the Gamma distribution with shape parameter a and scale parameter b . The marginal distribution of t_i given δ_i is therefore $t(\delta_i, d)$. Thus the MDP model (7) can be re-expressed as the following,

$$\begin{aligned} t_i|\delta_i, \tau_i &\sim N(\delta_i, \tau_i^{-1}), \\ \tau_i &\sim \text{Gamma}\left(\frac{d}{2}, \frac{d}{2}\right), \\ \delta_i &\sim G, \\ G &\sim DP(M \cdot G_0). \end{aligned} \quad (9)$$

This model, as will be shown below, allows direct Gibbs sampling with closed-form full conditional distributions.

For the base measure G_0 , we set it to a three-component normal mixture

$$G_0(\delta) \sim p_0 I_0(\delta) + p_1 N(\mu_{01}, \tau_{01}^{-1}) + p_2 N(\mu_{02}, \tau_{02}^{-1}), \quad (10)$$

where the indicator function $I_x(y) = 1$ (or 0) if $y = x$ (or $y \neq x$); N_+ , N_- are positively and negatively truncated normal distributions, respectively, and $p_0, p_1, p_2, \mu_{01}, \mu_{02}, \tau_{01}, \tau_{02}$ are unspecified hyperparameters with hyper priors

$$(p_0, p_1, p_2) \sim \text{Dirichlet}(\bar{p}_0, \bar{p}_1, \bar{p}_2), \quad (11)$$

$$\mu_{01} \sim N\left(\bar{\mu}_{01}, \bar{\tau}_{01}^{-1}\right), \quad (12)$$

$$\mu_{02} \sim N\left(\bar{\mu}_{02}, \bar{\tau}_{02}^{-1}\right), \quad (13)$$

$$\tau_{01} \sim \text{Gamma}\left(\frac{\bar{\alpha}_{10}}{2}, \frac{\bar{\beta}_{10}}{2}\right), \quad (14)$$

$$\tau_{02} \sim \text{Gamma}\left(\frac{\bar{\alpha}_{20}}{2}, \frac{\bar{\beta}_{20}}{2}\right), \quad (15)$$

respectively. Here $\text{Dirichlet}(\bar{p}_0, \bar{p}_1, \bar{p}_2)$ denotes the Dirichlet distribution of order 3 with parameters $\bar{p}_0, \bar{p}_1, \bar{p}_2$. Last, we set the prior of M to $\text{Gamma}(M_a, M_b)$.

The MDP model (9) allows us to get the full conditional for each of the unknown parameters via Markov chain Monte Carlo, which we describe in detail below. Let θ be the vector of all unknown parameters and let θ_{-v} be the remaining vector of θ after removing element v from θ , and $t = (t_1, \dots, t_m)$.

Step 0: Assign starting values to $\tau_i, \delta_i, i = 1, \dots, m$, and $\mathbf{P} = (p_0, p_1, p_2), \mu_{01}, \mu_{02}, \tau_{01}, \tau_{02}$.

Step 1: Sample $\tau_i, i = 1, \dots, m$ according to the following conditional distribution:

$$\tau_i | \theta_{-\tau_i}, t \sim \text{Gamma}\left(\frac{d+1}{2}, \frac{d+(t_i - \delta_i)^2}{2}\right). \quad (16)$$

Step 2: Sample $\delta_i, i = 0, \dots, m$. As derived in Escobar (1994), the full conditional of δ_i is proportional to

$$\sum_{k \neq i} q_{ik} I_{\delta_k}(\delta_i) + M q_{i0} G_0(\delta_i) N(t_i | \delta_i, \tau_i), \quad (17)$$

where $\delta_{-i} = (\delta_1, \dots, \delta_{i-1}, \delta_{i+1}, \dots, \delta_m)$, and $N(\cdot|\delta_i, \tau_i^{-1})$ is the density function of $N(\delta_i, \tau_i^{-1})$ at t_i . Let $q_{ik} = N(t_i|\delta_k, \tau_i^{-1})$, $k=1, \dots, i-1, i+1, \dots, m$, and $q_{i0} = \int N(t_i|\delta, \tau_i) G_0(\delta) d\delta$, which equals

$$\begin{aligned} q_{i0} &= \int N(t_i|\delta, \tau_i^{-1}) G_0(\delta) d\delta \\ &= p_0 N(t_i|0, \tau_i^{-1}) + p_1 \int_0^\infty N_+(t_i|\delta, \tau_i^{-1}) N(\delta|\mu_{01}, \tau_{01}^{-1}) d\delta + p_2 \int_{-\infty}^0 N_-(t_i|\delta, \tau_i^{-1}) N(\delta|\mu_{02}, \tau_{02}^{-1}) d\delta. \end{aligned} \quad (18)$$

The q_{ik} 's and Mq_{i0} need to be further normalized as below to find the selection probabilities. Let $C^{-1} = \sum_{k \neq i} q_{ik} + Mq_{i0}$. With probability Cq_{ik} , δ_i is drawn from $I_{\delta_k}(\delta_i)$ and with probability CMq_{i0} , δ_i is drawn from the distribution

$$aI_0(\delta_i) + bN_+\left(\delta_i \left| \frac{\tau_i t_i + \tau_{01} \mu_{01}}{\tau_i + \tau_{01}}, \frac{1}{\tau_i + \tau_{01}} \right.\right) + cN_-\left(\delta_i \left| \frac{\tau_i t_i + \tau_{02} \mu_{02}}{\tau_i + \tau_{02}}, \frac{1}{\tau_i + \tau_{02}} \right.\right),$$

where

$$\begin{aligned} a &= sp_0, \\ b &= sp_1 \left(\frac{\tau_{01}}{\tau_i + \tau_{01}} \right)^{\frac{1}{2}} \frac{N(-\mu_1^*|0, 1)}{N(-\tau_{01}^{\frac{1}{2}} \mu_{01}|0, 1)} \exp \left[-\frac{1}{2} (\tau_{01} \mu_{01}^2 - \mu_1^{*2}) \right], \end{aligned} \quad (19)$$

$$c = sp_2 \left(\frac{\tau_{02}}{\tau_i + \tau_{02}} \right)^{\frac{1}{2}} \frac{N(-\mu_2^*|0, 1)}{N(-\tau_{02}^{\frac{1}{2}} \mu_{02}|0, 1)} \exp \left[-\frac{1}{2} (\tau_{02} \mu_{02}^2 - \mu_2^{*2}) \right], \quad (20)$$

with s being the normalizing constant such that $a + b + c = 1$, and

$$\mu_1^* = \frac{\tau_i t_i + \tau_{01} \mu_{01}}{(\tau_i + \tau_{01})^{\frac{1}{2}}}, \quad \mu_2^* = \frac{\tau_i t_i + \tau_{02} \mu_{02}}{(\tau_i + \tau_{02})^{\frac{1}{2}}}. \quad (21)$$

We therefore sample $\delta_i^{(j+1)}$, $i=0, \dots, m$ according to (17). After sampling the δ_i 's for all subjects, we group them into clusters such that subjects within each cluster have the same δ value. Let L ($0 < L \leq m$) be the number of unique δ values. Denote these unique values as z_l , $l = 1, \dots, L$. Note that knowing δ_i is equivalent to knowing the z_l 's and the cluster that subject i belongs to.

One additional step in the sampling is recommended by Bush and MacEachern (1996) in order to improve the convergence for the Gibbs sampler. To speed the mixing over the entire parameter space, they suggest sampling the z_l 's according to the conditional density of z_l after assigning subjects to their clusters:

$$p(z_l|\theta, t) \propto \prod_{i \in \text{cluster } l} p(t_i|z_l, \tau_i) G_0(z_l)$$

which implies that

$$p(z_1|\theta, t) = \tilde{a} I_0(z_1) + \tilde{b} N_+\left(z_1 \middle| \tilde{\mu}_1^*, \frac{1}{\sum_{i \in l} \tau_i + \tau_{01}}\right) + \tilde{c} N_-\left(z_1 \middle| \tilde{\mu}_2^*, \frac{1}{\sum_{i \in l} \tau_i + \tau_{02}}\right), \quad (22)$$

where

$$\begin{aligned} \tilde{a} &= \tilde{s} p_0, \\ \tilde{b} &= \tilde{s} p_1 \left(\frac{\tau_{01}}{\sum_{i \in l} \tau_i + \tau_{01}} \right)^{\frac{1}{2}} \frac{N(-\tilde{\mu}_1^*|0, 1)}{N(-\tau_{01}^{\frac{1}{2}} \mu_{01}|0, 1)} \exp \left[-\frac{1}{2} \left(\tau_{01} \mu_{01}^2 - \tilde{\mu}_1^{*2} \right) \right], \end{aligned} \quad (23)$$

$$\tilde{c} = \tilde{s} p_2 \left(\frac{\tau_{02}}{\sum_{i \in l} \tau_i + \tau_{02}} \right)^{\frac{1}{2}} \frac{N(-\tilde{\mu}_2^*|0, 1)}{N(-\tau_{02}^{\frac{1}{2}} \mu_{02}|0, 1)} \exp \left[-\frac{1}{2} \left(\tau_{02} \mu_{02}^2 - \tilde{\mu}_2^{*2} \right) \right], \quad (24)$$

with $\tilde{s}$ being the normalizing constant to ensure that $\tilde{a} + \tilde{b} + \tilde{c} = 1$, and

$$\begin{aligned} \tilde{\mu}_1^* &= \frac{\sum_{i \in l} \tau_i t_i + \tau_{01} \mu_{01}}{\left(\sum_{i \in l} \tau_i + \tau_{01} \right)^{\frac{1}{2}}}, & \tilde{\mu}_2^* &= \frac{\sum_{i \in l} \tau_i t_i + \tau_{02} \mu_{02}}{\left(\sum_{i \in l} \tau_i + \tau_{02} \right)^{\frac{1}{2}}}. \end{aligned} \quad (25)$$

Step 3: Sample (p_0, p_1, p_2) : The full conditional of $\mathbf{p} = (p_0, p_1, p_2)$ is

$$\left[(p_0, p_1, p_2) | \theta_{-(p_0, p_1, p_2)}, t \right] \sim \text{Dirichlet}(\tilde{p}_0 + \tilde{n}_0, \tilde{p}_1 + \tilde{n}_1, \tilde{p}_2 + \tilde{n}_2), \quad (26)$$

where $\tilde{n}_0, \tilde{n}_1, \tilde{n}_2$ are the number of z_i s belonging to the clusters whose δ_i' are equal to, greater than, or less than 0 respectively, that is $\tilde{n}_0 + \tilde{n}_1 + \tilde{n}_2 = L$.

Step 4: Sample the hyperparameters $\mu_{01}, \mu_{02}, \tau_{01}, \tau_{02}$ from their full conditionals:

$$\mu_{01} | \theta_{-\mu_{01}}, t \sim N \left(\frac{\tau_{01} \sum_{l=1}^L I(\delta_l = 0) \delta_l + \tilde{\tau}_{01} \tilde{\mu}_{01}}{\tilde{n}_1 \tau_{01} + \tilde{\tau}_{01}}, \frac{1}{\tilde{n}_1 \tau_{01} + \tilde{\tau}_{01}} \right), \quad (27)$$

$$\mu_{02} | \theta_{-\mu_{02}}, t \sim N \left(\frac{\tau_{02} \sum_{l=1}^L I(\delta_l < 0) \delta_l + \tilde{\tau}_{02} \tilde{\mu}_{02}}{\tilde{n}_2 \tau_{02} + \tilde{\tau}_{02}}, \frac{1}{\tilde{n}_2 \tau_{02} + \tilde{\tau}_{02}} \right), \quad (28)$$

$$\tau_{01}|\theta_{-\tau_{01}}, t \sim \text{Gamma}\left(\frac{\tilde{n}_1 + \tilde{\alpha}_{10}}{2}, \frac{\sum_{l=1}^L I(\delta_l = 0) (\delta_l - \mu_{01})^2 + \tilde{\beta}_{10}}{2}\right), \quad (29)$$

$$\tau_{02}|\theta_{-\tau_{02}}, t \sim \text{Gamma}\left(\frac{\tilde{n}_2 + \tilde{\alpha}_{20}}{2}, \frac{\sum_{l=1}^L I(\delta_l < 0) (\delta_l - \mu_{02})^2 + \tilde{\beta}_{20}}{2}\right), \quad (30)$$

respectively.

Step 5. Sample M in two steps: we first sample a latent variable η from

$$\eta \sim \text{Beta}(M+1, N), \quad (31)$$

then we sample M from

$$M|\eta \sim \pi_\eta \Gamma(M_a + L, M_b - \log(\eta)) + (1 - \pi_\eta) \Gamma(M_a + L - 1, M_b - \log(\eta)), \quad (32)$$

where

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{M_a + L - 1}{N(M_b - \log(\eta))}. \quad (33)$$

Steps 1-5 complete one cycle of the Gibbs sampler. From the Gibbs samples, one can easily obtain Bayesian estimates of the parameters of primary interest, such as the FDR shown in Equation (5).

3 Numerical Studies

In this section, we test the performance of our method using both simulated and real datasets. Similar to Hu et al (2005), we first simulated two sets of data from mixture exponential and mixture normal distributions for F . The range of π_0 is from 0.3 to 0.9, which covers real biologic scenarios. In the exponential setup, we chose $a_1 = a_2 = 1/\lambda = 0.5, 1, 2.5, 5$, which represent an F with comparatively short, medium and heavy tails, respectively. In the normal setup, we chose $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 0.5, 1, 2.5, 5$ which again represent an F with comparatively short, medium and heavy tails. We set the hyperparameters $\bar{p}_0 = \bar{p}_1 = \bar{p}_2 = 1, \bar{\mu}_{01} = \bar{\mu}_{02} = 0, \bar{\tau}_{01} = \bar{\tau}_{02} = 1, \bar{\alpha}_{10} = \bar{\alpha}_{20} = \bar{\beta}_{10} = \bar{\beta}_{20} = 1$, and $M_a = M_b = 1$.

The length of the chain consisted of 23000 Gibbs samples. The first 3000 samples (the burn-in) were discarded. The remaining samples were saved for every 20 samples to reduce serial correlations. The final posterior samples contained 1000 samples and were used for the subsequent Bayesian analysis. Table 1 shows the posterior mean and the 90% credible interval for one of the primary parameters, π_0 . The upper part is for the exponential mixture

settings and the lower part is for the normal mixture settings. Figure 1 shows one example of the posterior distribution for $\pi_0 = 0.7$. For all the π_0 values simulated, the farther the two non-zero components of F are from zero, the better our method performs. This is intuitive since for those distributions with two non-zero components of F close to 0, the resulting distributions of the test statistics t_i s are centered around 0 and thus it becomes harder to distinguish the two non-zero components from 0. This statement can be further illustrated by Figure 2, in which the Q-Q plots of the estimated F distributions against the simulated ones are provided. Comparing the exponential mixture results (upper panel) to the normal mixture ones (lower panel), we find that our approach fits the exponential mixture model better. This is consistent with the above conclusion since the two non-zero exponential distributions have longer tails than its two counterpart normal distributions when $a = \sigma$.

We also reconstruct the distributions of the test statistics based on the estimated underlying distributions of δ and compare them with the true distributions of the test statistics (Figure 3). It is clear that the reconstructed distributions of the test statistics match their corresponding true distributions well, regardless of how heavy the tail for δ is. Further, our method always performs better than the parametric estimations in Hu et al (2005) unless the parametric model assumptions hold exactly.

Next, we demonstrate the effectiveness of our method on a real data set. we consider a murine lung microarray dataset submitted to Gene Expression Omnibus (E. Hoffman, Children's National Medical Center, GDS251). The study used the Affymetrix U74Av2 array (12488 probesets) to compare samples from the two mouse strains: C57BL/6/J (n=12) and Balb/c (n=12) in their sensitivity with pulmonary fibrosis. This data has been analyzed by Hu et al (2005) and we compare our results to theirs.

The 95% credible intervals of the π_i s are (0.27, 0.47), (0.33, 0.55) and (0.09, 0.29) for π_0 , π_1 and π_2 , respectively. These results are somewhat different from the estimations in Hu et al (2005). The difference between π_1 and π_2 indicates that the distribution of δ is not symmetric in the positive and the negative components. The fitted distribution for δ is shown in the right panel of Figure 4. The Q-Q plot of 12488 observed t-statistics versus the estimated ones is given in the left panel of Figure 4. Compared to the Q-Q plot (Figure 3) in Hu et al (2005), our nonparametric model clearly provides a better fit.

4 Discussion

We have described a Bayesian semiparametric approach to estimate the underlying gene expression distribution F . Our method involves specifying a nonparametric prior for the distribution of the scaled gene expression difference via a mixture of Dirichlet processes. The resulting model is estimated via the Gibbs sampler. Compared to the parametric mixture model, our nonparametric Bayesian method is more flexible and robust in modeling the gene expression distribution. Once the distribution of the effect size is estimated, we can easily apply this method to the FDR-based sample size calculations in Hu et al (2005). The biases in estimating F with short tails is not unique to our method. All parametric and non-parametric methods share similar (identification) problems. For such situations, increasing the sample size may be necessary in improving estimation accuracy. In this paper, only a two-sample design is considered. One future extension is to extend the method to other experimental designs. More general multi-sample designs or general linear regression modeling require a more careful consideration regarding δ , but the essence of the approach remains the same.

Acknowledgments

This research is supported in part by NIH grant R01 GM074175.

References

- [1]. Berry DA, Christensen R. Empirical bayes estimation of a Binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics* 1979;7:558–568.
- [2]. Black MA, Doerge RW. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* 2002;18:1609–1616. [PubMed: 12490445]
- [3]. Blackwell D, MacQueen JB. Ferguson distributions via Polya urn schemes. *Annals of Statistics* 1973;1:353–355.
- [4]. Bush CA, MacEachern SN. A semi-parametric Bayesian model for randomized block designs. *Biometrika* 1996;83:275–285.
- [5]. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 1997;2:364–367.
- [6]. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12:111–139.
- [7]. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001;96:1151–1160.
- [8]. Escobar MD. Estimating normal means with a Dirichlet process prior. *Journal of American Statistical Association* 1994;89:268–277.
- [9]. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* 1995;90:577–588.
- [10]. Ferguson TS. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1973;1:209–230.
- [11]. Guan Z, Wu B, Zhao H. Application of Bernstein polynomial in the estimation of false-discovery-rate. *Statistica Sinica* 2008;18:905–923.
- [12]. Hu J, Wright FA, Zou F. Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* 2005;21:3264–3272. [PubMed: 15932903]
- [13]. Ishwaran H, James LF. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* 2002;11:508–532.
- [14]. Kleinman KP, Ibrahim JG. A Semi-parametric bayesian approach to the random effects model. *Biometrics* 1998;54:921–938. [PubMed: 9750242]
- [15]. Lee MT, Whitmore GA. Power and sample size for dna microarray studies. *Statistics in Medicine* 2002;21:3543–3570. [PubMed: 12436455]
- [16]. Laird NM. Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc* 1978;73:805–811.
- [17]. Laird NM. Empirical Bayes estimates using nonparametric maximum likelihood estimate for the prior. *J. Statist. Comput. Simulation* 1981;15:211–220.
- [18]. Lindsay BG. The geometry of mixture likelihoods: a general theory. *Ann Stat* 1983;11:86–94.
- [19]. Liu JS. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* 1996;24:911–930.
- [20]. MacEachern SN. Estimating normal means with conjugate style Dirichlet process prior. *Communication in Statistics Simulation and Computation* 1994;23:727–741.
- [21]. MacEachern SN, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* 1998;7:223–238.
- [22]. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000;9:249–265.

- [23]. Pan W, Lin J, Le C. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* 2002;3:research0022.1-0022.10.
- [24]. Rasmussen, CE. The Infinite Gaussian Mixture Model. In: Solla, SA.; Leen, TK.; Muller, K-R., editors. *Adv. Neur. Inf. Proc. Sys.* Vol. 12. MIT Press; 2000. p. 554-560.
- [25]. Robbins, H. An Empirical Bayes approach to statistics; *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; 1955; p. 157-163.
- [26]. Ruppert D, Nettleton D, Hwang J. T. Gene. Exploring the information in P-values for the analysis and planning of multiple-test experiments. *Biometrics* 2007;63:483–495. [PubMed: 17715492]
- [27]. Tang Y, Ghosal S, Roy A. Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics* 2007;63:1126–1134. [PubMed: 17501943]
- [28]. Troyanskaya OG, Garber ME, Brown P, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* 2002;18:1454–1461. [PubMed: 12424116]
- [29]. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* 2001;98:5116–5121.
- [30]. Wu B, Guan Z, Zhao H. Parametric and nonparametric FDR estimation revisited. *Biometrics* 2006;62:735–744. [PubMed: 16984315]
- [31]. Yoon H, Liyanarachchi S, Wright FA, Davuluri R, Lockman JC, de la Chapelle A, Pellegata NS. Gene expression profiling of isogenic cells with different tp53 gene dosage reveals numerous genes that are affected by tp53 dosage and identifies cspg2 as a direct target of p53. *Proceedings of the National Academy of Sciences USA* 2002;99:15632–15637.

Figure 1. Histograms for the posterior distributions of π_0 for the four simulation studies for the mixture exponential model (upper) and mixture normal distribution (lower). $\pi_0 = 0.7$.

Figure 2.

Q-Q plots for the underlying F distribution in a simulation study ($n=50$, $m=10000$). The lower panel is for the normal mixture model, the upper panel is for the exponential mixture model, and $\pi_0 = 0.7$.

Figure 3. Q-Q plots for the observed t-statistics in a simulation study ($n=50$, $m=10000$). The lower panel is for the normal mixture model, the upper panel is for the exponential mixture model, and $\pi_0 = 0.7$.

Figure 4. Real murine lung study: the left panel is the Q-Q plot of the observed t-statistics vs the estimated ones; the right panel displays the histogram of the posterior distribution of F .

Table 1 π_0 and their estimates obtained from the simulated data

Exponential Mixture				
π_0	a=5	a=2.5	a=1	a=0.5
0.3	0.31(0.1,0.43)	0.26(0.002,0.48)	0.49(0.06,0.72)	0.42(0,0.88)
0.5	0.52(0.29,0.64)	0.55(0.25,0.72)	0.49(0,0.81)	0.60(0.06,0.92)
0.7	0.69(0.54,0.75)	0.68(0.46,0.78)	0.74(0.11,0.87)	0.75(0.05,0.92)
0.9	0.92(0.82,0.96)	0.91(0.74,0.97)	0.89(0.12,0.99)	0.97(0.93,1)

Normal Mixture				
π_0	$\sigma = 5$	$\sigma = 2.5$	$\sigma = 1$	$\sigma = 0.5$
0.3	0.28(0.001,0.45)	0.43(0.004,0.61)	0.50(0,0.81)	0.86(0.18,1)
0.5	0.42(0.02,0.57)	0.45(0.009,0.67)	0.32(0,0.72)	0.92(0.46,1)
0.7	0.66(0.38,0.75)	0.68(0.45,0.80)	0.64(0.01,0.87)	0.59(0,0.94)
0.9	0.89(0.66,0.97)	0.90(0.66,0.98)	0.97(0.81,1)	0.97(0.88,1)

The fifth and ninety-fifth percentiles of the posterior distribution are given in parentheses.