

# Nonparametric Bayesian Estimation of Positive False Discovery Rates

Yongqiang Tang,<sup>1</sup> Subhashis Ghosal,<sup>2</sup> and Anindya Roy<sup>3,\*</sup>

Department of Psychiatry, SUNY Health Science Center, Brooklyn, New York 11203, U.S.A.  
Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

Department of Mathematics and Statistics, University of Maryland, Baltimore County,  
1000 Hilltop Circle, Baltimore, Maryland 21250, U.S.A.

\**email:* anindya@math.umbc.edu

**SUMMARY.** We propose a Dirichlet process mixture model (DPMM) for the  $P$ -value distribution in a multiple testing problem. The DPMM allows us to obtain posterior estimates of quantities such as the proportion of true null hypothesis and the probability of rejection of a single hypothesis. We describe a Markov chain Monte Carlo algorithm for computing the posterior and the posterior estimates. We propose an estimator of the positive false discovery rate based on these posterior estimates and investigate the performance of the proposed estimator via simulation. We also apply our methodology to analyze a leukemia data set.

**KEY WORDS:** Dirichlet mixture; Dirichlet process; Markov chain Monte Carlo; Multiple testing; Positive false discovery rate; Posterior estimates.

## 1. Introduction

Recent advances in data-collection techniques in biological and related sciences have made it possible for scientists to study very complex problems and have also challenged statisticians to develop tools for analyzing the resulting complex data structures. In doing so, multiple testing has come to the forefront of statistical research. Multiple testing considers the problem of simultaneously testing  $m$  null hypotheses  $H_{0,1}, \dots, H_{0,m}$ , where  $m$  can be considerably large. In such situations, false discoveries (true null hypothesis declared significant) are inevitable. Thus, it is important in any multiple testing problem to control the error rate of false discoveries. When  $m$  is small, the family-wise error rate (FWER) defined as probability of making at least one false discovery is the obvious analogue of Type I error probability. Thus, traditionally in multiple testing problems with small  $m$ , FWER has been the natural measure to control. However, in the analysis of modern biological data, such as microarray data, proteomics data, and functional Magnetic Resonance Imaging (fMRI) data, FWER is too stringent and counterproductive for the scientific goals of finding significant number of discoveries. To this end, Benjamini and Hochberg (1995) introduced the false discovery rate (FDR) as a relevant measure to be controlled in a multiple testing problem. The FDR is defined as the expected proportion of false rejections among the rejected hypotheses. Specifically, if  $R$  is the number of significances declared, and  $V$  represent the number of false discoveries made, then the FDR is defined as

$$\text{FDR} = E \left\{ \frac{V}{\max(R, 1)} \right\} = E \left( \frac{V}{R} \mid R > 0 \right) \Pr(R > 0). \quad (1)$$

Since its introduction, FDR has become the most popular measure to control in large multiple testing problems. Many related measures such as the positive FDR (pFDR) made popular by Storey (2002), marginal FDR (Benjamini and Hochberg, 1995), local FDR (Efron et al., 2001; Efron and Tibshirani, 2002), and conditional FDR (Tsai, Hsueh, and Chen, 2003) have been suggested in the literature. Storey (2002, 2003) argue that the positive pFDR defined as  $\text{pFDR} = E(V/R \mid R > 0)$ , may be conceptually more sound than the FDR.

There are two key aspects in the estimation of FDR or pFDR—the estimation of the unknown number of true null hypothesis and the estimation of the  $P$ -value distribution when the alternative hypothesis is true. Storey (2002) proposed a mixture model setup for evaluating or estimating pFDR (or FDR) that seems convenient and appropriate in many multiple testing situations. In his framework, the test statistics are supposed to be independent and identically distributed. Each null hypothesis has a fixed probability,  $\pi_0$ , of being true. Thus, the number of true null hypotheses,  $m_0$ , is taken to be a random variable distributed as binomial ( $m$ ,  $\pi_0$ ). Also, marginal  $P$ -value distribution,  $F$ , is then a mixture of the uniform distribution (distribution when the null hypothesis is true) and an alternative distribution  $F_1$  (distribution when the alternative hypothesis is true). If individual hypotheses are tested at nominal level  $\alpha$ , then Storey (2002) showed that the pFDR is given by

$$\text{pFDR} = P(\text{A null hypothesis is true} \mid \text{it is rejected}) = \frac{\pi_0 \alpha}{F(\alpha)}. \quad (2)$$

Based on this relationship, Storey (2002) proposed estimators for pFDR and consequently that of FDR. Tsai et al. (2003), Pounds and Morris (2003), Pounds and Cheng (2004), and Dalmaso, Broët, and Moreau (2005), among others have also utilized the relationship (2) to propose estimators of pFDR and FDR. Although the above framework assumes a simple versus simple testing situation with a fixed alternative value, as Storey (2003) mentioned, the relationship (2) holds even for simple versus composite testing situation as long as one models the alternative parameter values as a random variable. Then  $F_1$  is the mixture of the alternative distributions. Thus, even though the problem of multiple testing belongs to the classical frequentist paradigm, the probabilities that one would like to estimate seems more natural to arise in a Bayesian framework. The pFDR is written in the form of a posterior probability and in case of simple versus composite testing one needs to resort to mixed effect models (cf., Genovese and Wasserman, 2002) or a Bayesian model to incorporate the added variability of the alternative parameter into the analysis of pFDR.

In this article, we formulate a mixture model framework for the alternative  $P$ -value distribution under a simple versus composite testing situation and estimate  $\pi_0$  and  $F_1$  using a nonparametric Bayesian technique. Thus, in turn we propose a nonparametric Bayesian estimator for pFDR and FDR. In Section 2, we formulate our model and propose a prior for the model. In Section 3, we describe our algorithm for computing the posterior distribution and define the proposed estimator of pFDR. In Section 4, we illustrate our methodology with two simulation experiments and analyze a leukemia data set. Section 5 summarizes our findings.

**2. Dirichlet Process Mixture Model**

Consider the problem of testing  $m$  independent hypotheses. Let  $X = (X_1, \dots, X_m)$  be the  $P$ -values for the  $m$  tests. Let  $H_i = 1$  if the  $i$ th null hypothesis is false and  $H_i = 0$  otherwise. Suppose that the density of  $X_i$  under the null hypothesis is  $f_0$ , and the distribution is  $f_1$  under the alternative hypothesis. We model  $H_i, i = 1, \dots, m$ , as independent Bernoulli trials with success probability  $\pi_1 = 1 - \pi_0$  and assume that  $f_0$  is the uniform distribution on  $[0, 1]$ . Then the  $P$ -values could be viewed as independent and identically distributed (i.i.d.) samples from the two-component mixture model:

$$f(x) = 1 - \pi_1 + \pi_1 f_1(x). \tag{3}$$

As pointed out by Parker and Rothenberg (1988), many distributions on the interval  $[0, 1]$  can be modeled as a mixture of beta distributions. A nonparametric mixture model (NMM) is extremely useful for modeling  $P$ -values in this sense. In NMM,  $f_1$  is modeled as a nonparametric mixture of beta distributions:

$$f_1(x) = \int g(x | a, b) dG(a, b), \tag{4}$$

where  $G$  is a mixing distribution function,  $g(x | a, b) = x^{a-1}(1-x)^{b-1}/B(a, b)$  is the probability density function (PDF) of the beta distribution, and  $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$  stands for the beta function. The  $P$ -value density under the alternative hypothesis tends to concentrate around

zero. Thus, it is reasonable to model the  $P$ -value density under the alternative to have a “ $J$ ” shaped form that decreases monotonically over the interval  $[0, 1]$ . To build this additional feature in our model, we impose a restriction that  $G$  put all mass on  $\{(a, b) : a < 1, b \geq 1\}$ . Under this assumption the beta kernel and hence  $f_1$  have a “ $J$ ” shape and decreases monotonically over  $[0, 1]$ . The case  $b = 1$  implies that the  $P$ -value density under the alternative is nonzero at one, i.e.,  $f_1(1) > 0$  and that results in nonidentifiability in terms of the pair  $(\pi_0, f_1)$ . We will only consider the cases  $f_1(1) = 0$  in this article. However, extensions can be easily made by letting  $G$  put positive mass on  $(0, 1) \times \{1\}$ . Allison et al. (2002) proposed a finite mixture of beta density model that is a special case when  $G$  is a finitely supported distribution. However, it seems that much more flexibility can be built into the model by recognizing the uncertainty in the mixing distribution. Moreover, if the alternative values arise randomly from a population distribution, then also the mixture model (4) continues to hold. Thus, our method covers much wider models and this is a distinct advantage over other methods proposed in the literature. We thus propose a Dirichlet process mixture model (DPMM) where the mixing distribution  $G$  is random and assumed to be drawn from a Dirichlet process (cf., Ferguson, 1973). The parameters of a Dirichlet process  $DP(G_0, \tau)$  are a base measure  $G_0$ , and a scalar precision parameter  $\tau > 0$ . The base distribution  $G_0$  is the best guess of what the true  $G$  is believed to be and it is restricted to be supported on  $(0, 1) \times (1, \infty)$ . The realization of the process concentrates around the base measure  $G_0$  as  $\tau$  becomes larger.

The DPMM is equivalent to the following hierarchical model, where associated with each  $X_i$  is a latent variable  $\theta_i = (a_i, b_i)$ ,

$$X_i | \theta_i \sim 1 - \pi_1 + \pi_1 g(x_i | \theta_i), \quad \theta_1, \dots, \theta_n | G \stackrel{\text{i.i.d.}}{\sim} G \quad \text{and} \\ G \sim DP(G_0, \tau).$$

Such a hierarchical representation in terms of the latent variables  $\theta_i$ 's is useful for developing the MCMC algorithm for posterior computation described in the next section. The random measure  $G$  could be further integrated out from the prior distribution, and the joint distribution of  $\theta_i$ 's is given by the generalized Polya urn scheme (cf., Blackwell and MacQueen, 1973):

$$\theta_1 \sim G_0(\theta_1), \quad \theta_{i+1} | \theta_1, \dots, \theta_i \\ \sim \frac{\tau}{\tau + i} G_0(\cdot) + \sum_{j=1}^i \frac{1}{\tau + i} \delta_{\theta_j}(\cdot) \quad \text{for } i \geq 1, \tag{5}$$

where  $\delta_{\theta_j}(\cdot)$  denotes a unit point mass at  $\theta_j$ . The distribution (5) implies that  $\theta_i$ 's tend to share values in common. In a typical problem, there may be only few distinct  $\theta_i$  values (or there are only few distinct alternative values) but the number of such values is not fixed beforehand. The DPMM model has the flexibility to let the data update the shape of  $G$ , and determine clusters of  $P$ -values according to their association with the distinct latent components. This property has been used by McLachlan, Bean, and Peel (2002) for gene clustering using microarray data.

In the model, we reparameterize  $a$  as  $L_a$  and  $b$  as  $L_b$ , where  $a = \exp(-|L_a|)$ ,  $b = \exp(|L_b|)$ , and specify  $G_0$  as  $G_0(a, b) = N(L_a | 0, \sigma_a^2)N(L_b | 0, \sigma_b^2)$ . To avoid extra notation, we use  $N(x | u, \sigma^2)$  to denote both a normal distribution with mean  $u$  and standard deviation  $\sigma$ , and the corresponding PDF. The prior for  $\pi_1 = \exp(-|L_\pi|)$  is obtained from  $L_\pi \sim N(0, \sigma_\pi^2)$  and a priori  $\pi_1$  is assumed to be independent of  $f_1$ . Our specific formulation of the model and the prior makes the task of designing a random walk chain in the MCMC simulation straightforward. It is well known (cf., Tierney, 1994) that random walk chains are more efficient than independence chains. We also tested the sensitivity of the posterior to the particular prior formulation, by using a  $G_0$  that is the product of a beta and a truncated gamma distribution and choosing the prior for  $\pi$  to be a beta, and using the independence chain for posterior computation. The results from the two models were very similar, indicating that the inference in the DPMM model is not sensitive to the particular functional form of  $G_0$ . The independence chain moved very slowly and thus increased the computational time substantially.

**3. Posterior Computation**

In this section, we describe an MCMC algorithm that could be used to sample from the posterior distribution of  $(\pi_1, f_1)$ . Let  $\phi = \{\phi_1, \dots, \phi_k\}$  denote the set of distinct  $\theta_i$ 's, where  $k$  is the number of distinct elements of  $\theta_1, \dots, \theta_m$ . Let  $s = (s_1, \dots, s_m)$  denote the configuration vector, that is,  $s_i = j$  if and only if  $\theta_i = \phi_j$ . Thus  $\theta = \{\theta_i : i = 1, \dots, m\}$  is reparameterized as  $\{\phi_1, \dots, \phi_k, s_1, \dots, s_m\}$ . Let  $m_i, i = 1, \dots, k$ , be the number of elements  $s_j$  that are equal to  $i$ . Let the subscript “ $-i$ ” stand for all the variables except the  $i$ th one. Thus,  $m_{-i,j}$  will be the number elements  $s_r = j$ , where  $s_r$  are computed based on all  $\theta$ 's except the  $i$ th one and let  $k_{-i}$  denote the set  $\{1, 2, \dots, k\}$  except  $s_i$ .

In the standard MCMC scheme based on the generalized Polya urn scheme (cf., Escobar and West, 1995), a Gibbs step is used to sample  $\theta_i$  (equivalently,  $s_i$ ) from its posterior distribution

$$p(\theta_i | X, s_{-i}, \phi_{-i}) \propto \sum_{j \in k_{-i}} q_{i,j} \delta_{\phi_j} + q_{i,0} G_i(\theta_i),$$

$$dG_i(\theta) \propto f_\theta(X_i) dG_0(\theta),$$

$$q_{i,j} \propto m_{-i,j} f_{\phi_j}(X_i), \quad q_{i,0} \propto \tau \int f_\theta(X_i) dG_0(\theta).$$

(6)

When  $G_0$  is not conjugate with  $f$ , the integral  $q_{i,0}$  will be difficult to evaluate and drawing sample from  $G_i$  will be extremely challenging. To overcome this difficulty, MacEachern and Muller (1998) developed an extremely innovative algorithm, called the no-gaps algorithm, which can bypass the problems of drawing samples from (6) and evaluating  $q_{i,0}$ . Alternative MCMC schemes for handling the nonconjugate prior may be found in Neal (1998) and MacEachern and Muller (2000).

In the no-gaps algorithm, the vector of distinct  $\phi$ 's is augmented with an additional set of variables as follows:

$$\underbrace{\{\phi_1, \dots, \phi_k\}}_{\phi_F}, \underbrace{\{\phi_{k+1}, \dots, \phi_m\}}_{\phi_E},$$

with the same independent prior  $\phi_j \sim G_0$  and the same definition of configuration  $s$ . The vectors  $\phi_F$  and  $\phi_E$  are, respectively, referred to as the full and the potential clusters. The augmentation relies upon the constraint that there be no gaps in the values of the  $s_i$ , that is,  $m_j > 0$  for  $j = 1, \dots, k$  and  $m_j = 0$  for  $j = k + 1, \dots, m$ .

To implement the algorithm, initialize  $\pi$ ,  $s$  and  $\phi_F$  and repeat the following steps until the algorithm converges:

- (1) For  $i = 1, \dots, m$ , repeat (ia) and (ib).
  - (ia) If  $m_{s_i} > 1$ , resample  $s_i$  from the discrete distribution
 
$$\Pr(s_i = j | \phi, s_{-i}, \pi, X) \propto \begin{cases} m_{-i,j} f(X_i, \phi_j, \pi) & j \in k_{-i} \\ \frac{\tau}{k_{-i} + 1} f(X_i, \phi_j, \pi) & j = s_i. \end{cases}$$

(7)
  - (ib) If  $m_{s_i} = 1$ , with probability  $1 - k^{-1}$ , leave  $s_i$  unchanged. Otherwise rearrange the indices of  $\phi_j$  and correspondingly  $s$  and  $m_j$ 's such that  $s_i = k$ , and then resample  $s_i$  according to (7).

Note that the integral expression for  $q_{i,0}$  in (6) is replaced by simple density evaluations required by (7).

- (2) Both prior and posterior distributions for  $\phi_i \in \phi_E$  are  $G_0$ .

Because the number of distinct components  $k$  is typically small,  $\phi_i \in \phi_E$  is generated only if new distinct components are needed. The posterior distribution for  $\phi_i \in \phi_F$ , is

$$((L_{a_i}, L_{b_i}) | \phi_{-i}, s, X, \pi_1) \propto N(L_{a_i} | 0, \sigma_a^2) N(L_{b_i} | 0, \sigma_b^2) \times \prod_{j: s_j=i} \{1 - \pi_1 + \pi_1 g(X_j | a_i, b_i)\}.$$

We use a random walk Metropolis step to update  $\phi_i \in \phi_F$ . The “candidate”  $\theta_i^* = (\exp(-|L_{a_i}^*|), \exp(|L_{b_i}^*|))$  is generated from  $L_{a_i}^* \sim N(\cdot | L_{a_i}, V_a)$  and  $L_{b_i}^* \sim N(\cdot | L_{b_i}, V_b)$ . The values of  $V_a$  and  $V_b$  are determined automatically by the program such that the average acceptance probability of  $\phi_i$ 's is roughly between 0.20 and 0.65. The choice of  $V_a$  and  $V_b$  makes a compromise between the jump distance in the parameter space and the acceptance frequency, both of them ensure the efficiency of the MCMC algorithm.

- (3) The posterior distribution for  $\pi_1$  is

$$(L_\pi | \phi, s, X) \propto N(L_\pi | 0, \sigma_\pi^2) \prod_{i=1}^n \{1 - \pi_1 + \pi_1 g(X_i | a_i, b_i)\}.$$

We update it with a random walk Metropolis step. The “candidate”  $\pi^* = \exp(-|L_\pi^*|)$  was sampled from  $L_{\pi,i}^* \sim N(L_{\pi,i}^* | L_{\pi,i}, V_\pi)$ . Again,  $V_\pi$  was automatically determined by the program so that the average acceptance probability of  $\pi$  is roughly between 0.20 and 0.65.

MacEachern and Muller (1998) showed that the no-gaps algorithm converges almost surely under a very mild sufficient condition. In our empirical studies, convergence of the no-gaps algorithm, assessed through multiple chain together with informal graphical methods, was found to be quick. A burn-in period of 15,000 was adequate for all examples.

#### 4. Estimators

Posterior inference using Monte Carlo approximation is based on the general formula in the Web Appendix. Suppose  $M$  posterior observations were collected from the MCMC simulation after a burn-in period to ensure that the chain has reached stationarity. Also sufficient MCMC iterations were allowed between consecutive sample observations to reduce dependence. Let us use an additional superscript  $(j)$ ,  $j = 1, \dots, M$  to denote the sample index. Based on the posterior sample of  $\pi_1$  and  $\theta_1, \dots, \theta_m$ , define the  $j$ th posterior sample of the  $P$ -value density by  $f^{(j)} = 1 - \pi_1^{(j)} + \pi_1^{(j)} f_1^{(j)}$ , where

$$f_1^{(j)}(x) = \frac{\tau \int g(x|\theta) dG_0(\theta) + \sum_{i=1}^m g(x|\theta_i^{(j)})}{\tau + m}, \quad (8)$$

is the  $j$ th posterior sample of the  $P$ -value density under the alternative. Let  $F^{(j)}$  and  $F_1^{(j)}$  be the  $j$ th posterior sample of the corresponding distribution functions. The proportion of true null hypotheses  $\pi_0$  can be estimated as

$$\hat{\pi}_{0,\text{Bayes}} = E(\pi_0 | X) \cong 1 - \frac{1}{M} \sum_{j=1}^M \pi_1^{(j)}. \quad (9)$$

Note that this estimate is also the minimum value of the predictive density. Also for any function  $\psi = f, f_1, F, F_1$ , the Bayes estimate  $\hat{\psi}(x)$  of  $\psi(x)$  is computed from the corresponding Monte Carlo average  $M^{-1} \sum_{j=1}^M \psi^{(j)}(x)$ . This leads to the computational formula for the Bayes estimator of pFDR as

$$\widehat{\text{pFDR}}_{\text{Bayes}}(\alpha) = E \left\{ \frac{\pi_0 \alpha}{F(\alpha)} \mid X \right\} \cong M^{-1} \sum_{j=1}^M \frac{\pi_0^{(j)} \alpha}{F^{(j)}(\alpha)}. \quad (10)$$

Storey's nonparametric estimators of pFDR (defined in the next section) have a desirable property that they remain conservative even when the null distribution is stochastically larger than uniform. This is a consequence of the monotonicity of the empirical cdf in the observed values. If  $\mathbf{X} = (X_1, \dots, X_n) \geq \mathbf{Y} = (Y_1, \dots, Y_n)$  in the natural partial ordering, then  $F_n(\cdot | \mathbf{X})$  is stochastically larger than  $F_n(\cdot | \mathbf{Y})$ . Also,  $\hat{\pi}_0(\mathbf{X}) \geq \hat{\pi}_0(\mathbf{Y})$  because the proportion of  $P$ -values greater than a particular threshold will be bigger for the stochastically larger distribution. A similar property holds for Bayes estimators obtained from Dirichlet mixtures. Here is a sketch of the proof. Let  $\pi = \pi_0 = 1 - \pi_1$ . From (8), the  $P$ -value cumulative distribution function (CDF) obtained from the  $j$ th posterior sample is  $F^{(j)}(x) = \pi^{(j)} + (1 - \pi^{(j)}) F_1^{(j)}(x)$ , where

$$F_1^{(j)}(x) = F_1^{(j)}(x) = \frac{\tau \int \tilde{G}(x|\theta) dG_0(\theta) + \sum_{i=1}^m \tilde{G}(x|\theta_i^{(j)})}{\tau + m}, \quad (11)$$

and  $\tilde{G}$  is the beta CDF. Then the expression for pFDR from the  $j$ th posterior sample is  $\alpha \pi^{(j)} / F^{(j)}(\alpha)$ . Thus, it is enough to show that after reaching the stationary distribution, the MCMC chains pick larger values of  $\pi$  and stochastically larger values of  $F$  with greater probability for the  $\mathbf{X}$  sample than compared to the  $\mathbf{Y}$  sample. From (11), the second condition is satisfied if the MCMC chains for  $\theta_i = (a_i, b_i)$ ,

$i = 1, \dots, m$ , pick larger values with greater probability for the  $\mathbf{X}$  sample than for the  $\mathbf{Y}$  sample. Here by larger value of  $\theta$  we mean with respect to the ordering  $\theta < \theta'$  iff  $a(\theta) > a(\theta')$  and  $b(\theta) < b(\theta')$ . In order to prove this, it suffices to claim the stochastic ordering property for transition densities component by component. For the transition densities of  $\pi$ , the fact that beta densities in the mixture are decreasing implies that the updating factor  $\prod_{i=1}^n [\pi + (1 - \pi)g\{X_i | \theta_i^{(j)}(\mathbf{X})\}]$  favor larger  $\pi$  more than the updating factor  $\prod_{i=1}^n [\pi + (1 - \pi)g\{Y_i | \theta_i^{(j)}(\mathbf{Y})\}]$ . This is because if the  $\theta_i^{(j)}(\mathbf{X})$  are larger than  $\theta_i^{(j)}(\mathbf{Y})$  then the ratio of the two updating factors in that order is increasing in  $\pi$ . The quantity  $\theta$  given the hidden  $(\theta_1, \dots, \theta_n)$  has a distribution that is a mixture of point masses at those points and a new draw from the baseline posterior with certain weights. By the fact that for larger  $\mathbf{X}$  more relative weight is given to higher values, it remains only to show that the stationary distribution of the  $n$ -vector  $(\theta_1, \dots, \theta_n)$  is stochastically larger (in terms of partial ordering) for data  $\mathbf{X}$  than for data  $\mathbf{Y}$ . By the Polya urn scheme representation of Dirichlet mixture Gibbs sampling (5) and the fact that given  $\pi^{(j)}(\mathbf{X}) \geq \pi^{(j)}(\mathbf{Y})$  the CDF  $F^{(j)}$  are stochastically increasing in  $(\theta_1, \dots, \theta_m)$  (where the ordering is with respect to each component), we have that for each component  $\mathbf{X}$  induces higher  $a$  values and lower  $b$  values than what  $\mathbf{Y}$  does. But this in turn leads to smaller  $F(\alpha)$  values for each  $\alpha$ . This completes the argument.

Consistency of an estimator is one of the most fundamental requirements in the frequentist set up. Recently, Ghosal, Roy, and Tang (2006) have shown that the nonparametric Bayesian estimators, (9) and (10), of  $\pi_0$  and pFDR, respectively, are consistent in the frequentist sense. Specifically, they have demonstrated that the posteriors of  $\pi_0$  and pFDR arising from the proposed model are consistent. Other frequentist properties of the proposed estimators should be investigated as well.

#### 5. Numerical Examples

In this section, we will provide empirical illustration of the properties of the proposed Bayesian estimator. As a benchmark we will compare the Bayesian estimator with the three estimators proposed by Storey and a parametric maximum likelihood estimator (MLE) based on a single beta density model for the alternative distribution; see Pounds and Morris (2003). Throughout we will refer to the proposed estimators in (9) and (10) as "Bayes." We will refer to the estimators proposed by Storey as "p-HALF," "SPLINE," and "BOOT" and we will refer to the parametric estimator as "MLE." The main formula for frequentist estimators of pFDR is

$$\widehat{\text{pFDR}} = \frac{\hat{\pi}_0 \alpha}{\hat{F}(\alpha)},$$

where  $\hat{F}(\alpha)$  is based on the proportion of rejections among all hypotheses and the proportion of true null hypothesis,  $\pi_0$  is usually estimated as

$$\hat{\pi}_0(\lambda) = \frac{\#\{X_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)},$$

where  $\lambda$  is a tuning parameter. The three estimators of Storey are based on three different estimators of the tuning parameter. In p-HALF,  $\lambda$  is chosen to be 0.5. In "Boot"  $\lambda$  is chosen

by minimizing the bootstrap version of the mean square error of the pFDR estimator as a function of  $\lambda$ . We used 1000 bootstrap samples to compute the MSE of pFDR. In ‘‘SPLINE,’’ a natural cubic spline is fitted to  $\hat{\pi}_0(\lambda)$  for the entire range of  $\lambda$  and the final estimate of  $\pi_0$  is the limiting value of the cubic spline fit as  $\lambda$  approaches one, i.e., the cubic spline fit evaluated at  $\lambda = 1$ ; see Storey and Tibshirani (2003). For the parametric estimator the  $P$ -value density model is

$$f_1(x) = \pi_0 + (1 - \pi_0)g(x; a, b),$$

where  $g(x; a, b)$  is the beta density with parameters  $(a, b)$ . The MLE of  $F(\alpha)$  is then  $\hat{\pi}_{0,M} + (1 - \hat{\pi}_{0,M})G(\alpha; \hat{a}_M, \hat{b}_M)$ , where  $G$  is the beta CDF. The estimator for pFDR is found by substituting  $(\pi_0, F(\alpha))$  by their corresponding MLE in the expression for pFDR.

For the simulation experiment we consider two different models. The first one is a two-way mixed model where the test of interest is the equality of the fixed effect across two groups and the alternative value is unknown but fixed for all the tests. Specifically, the model for the raw data (and hence the corresponding  $P$ -values) for the  $r$ th test is

$$\begin{aligned} \text{Model I: } & y_{ijk} = \mu + \tau_i + g_{ij} + \varepsilon_{ijk}, \\ & \text{for } i = 1, 2, j = 1, \dots, 5, k = 1, \dots, 5, \end{aligned}$$

where  $g_{ij}$  are random effects and assumed to be i.i.d.  $N(0, 0.5)$  and the model errors  $\varepsilon_{ijk}$  are i.i.d.  $N(0, 1)$ . The  $\tau_i$  are fixed effects (e.g., treatment effect) and the  $r$ th hypothesis is

$$H_{0,r} : \tau_1 = \tau_2 \text{ vs. } H_{1,r} : \tau_1 \neq \tau_2, \quad r = 1, 2, \dots, m.$$

Then a test for  $H_{0,r}$  will be the usual level  $\alpha$   $F$ -test based on the averages  $z_{ij} = \frac{1}{5} \sum_k y_{ijk}$ . To study the empirical properties of the estimators, we set the following values for the parameters involved: The total number of hypothesis is  $m = 100, 1000$ , the proportion of true null hypotheses is  $\pi_0 = 0.8$ ,

0.9, 0.95, the alternative value for the false null hypotheses is fixed at  $\tau_1 - \tau_2 = 1.5$ , and the size of each test is  $\alpha = 0.05, 0.01$ .

The second simulation model is a one-way model with mixture alternatives. Specifically, the data (test statistics) for the  $r$ th hypothesis are generated from

$$\text{Model II: } T_i \sim N(\mu, 1),$$

and the test of interest are  $H_{0,r} : \mu = 0$  vs  $H_{1,r} : \mu > 0$ . However, while generating the data for the false null hypotheses, the alternative value of  $\mu$  is taken from a mixture distribution. The test statistic under the alternative model is simulated according to

$$\begin{aligned} T_i \sim & 0.25N(1, 1) + 0.4N(1.5, 1) + 0.2N(2, 1) \\ & + 0.1N(2.5, 1) + 0.05N(3, 1). \end{aligned}$$

The mean, standard error (STD), and the root mean square error (rmSE) of the estimators for  $\pi_0$  for the two models are given in Table 1. The biases of all of the nonparametric estimators except for the bootstrap estimator are comparable for the parameter values considered. The BOOT estimator seems to have a severe downward bias while estimating the true null proportion. The parametric estimator also has a severe bias in model I but the bias seems to be comparable with those of nonparametric estimators in model II. In terms of rmSE, the Bayes estimator outperforms the frequentist estimators, especially for the second model. In most practical examples, the proportion of false null hypotheses will be small compared to the total number of hypotheses that on the other hand will be large. The Bayes estimator of  $\pi_0$  compares favorably with the frequentist estimators for small  $\pi_1$  and large  $m$  cases. The MLE performs poorly compared to the nonparametric estimators. The relative performance of the Bayesian estimator and the frequentist estimators is more mixed in estimating

**Table 1**  
Summary of estimators of  $\pi_0$ : superscripts <sup>a,b,c</sup> represent mean, STD, and rmSE evaluated over 100 simulated samples, respectively

$m/\pi_0$	Model I					Model II				
	Bayes	p-Half	Spline	Boot	MLE	Bayes	p-Half	Spline	Boot	MLE
100/0.80	0.780 <sup>a</sup>	0.827	0.795	0.724	0.771	0.793	0.824	0.781	0.712	0.753
	0.064 <sup>b</sup>	0.098	0.168	0.145	0.169	0.053	0.095	0.169	0.149	0.185
	0.067 <sup>c</sup>	0.101	0.167	0.163	0.171	0.053	0.098	0.169	0.172	0.190
100/0.90	0.864	0.907	0.863	0.802	0.828	0.865	0.899	0.847	0.786	0.829
	0.051	0.086	0.145	0.146	0.209	0.043	0.086	0.147	0.150	0.211
	0.062	0.086	0.148	0.175	0.220	0.056	0.085	0.155	0.188	0.221
100/0.95	0.897	0.934	0.890	0.835	0.808	0.894	0.924	0.870	0.826	0.816
	0.042	0.076	0.137	0.147	0.293	0.038	0.076	0.136	0.148	0.288
	0.068	0.077	0.149	0.186	0.324	0.067	0.080	0.158	0.192	0.317
1000/0.80	0.802	0.826	0.810	0.781	0.810	0.803	0.832	0.803	0.784	0.786
	0.029	0.033	0.067	0.058	0.123	0.028	0.031	0.067	0.057	0.091
	0.029	0.042	0.068	0.060	0.123	0.028	0.044	0.066	0.058	0.092
1000/0.90	0.882	0.915	0.902	0.872	0.884	0.888	0.915	0.895	0.870	0.894
	0.028	0.034	0.065	0.059	0.104	0.023	0.033	0.065	0.057	0.063
	0.033	0.037	0.065	0.065	0.105	0.026	0.036	0.065	0.064	0.063
1000/0.95	0.930	0.956	0.939	0.917	0.909	0.935	0.956	0.936	0.914	0.944
	0.025	0.030	0.054	0.055	0.152	0.018	0.031	0.059	0.057	0.059
	0.032	0.030	0.055	0.064	0.156	0.024	0.031	0.060	0.067	0.059

**Table 2**

Summary of estimators of pFDR for model I: superscripts <sup>a,b,c</sup> represent mean, STD, and rMSE evaluated over 100 simulated samples, respectively

$m/\pi_0/\alpha$	True pFDR	Bayes	p-HALF	SPLINE	BOOT	MLE
100/0.80/0.05	0.275	0.310 <sup>a</sup>	0.317	0.303	0.278	0.277
		0.088 <sup>b</sup>	0.134	0.133	0.123	0.101
		0.094 <sup>c</sup>	0.140	0.135	0.122	0.101
100/0.80/0.01	0.146	0.158	0.210	0.199	0.182	0.165
		0.057	0.184	0.173	0.156	0.082
		0.058	0.194	0.181	0.160	0.084
100/0.90/0.05	0.456	0.468	0.532	0.506	0.473	0.449
		0.105	0.200	0.203	0.202	0.170
		0.105	0.213	0.208	0.202	0.170
100/0.90/0.01	0.266	0.271	0.365	0.346	0.321	0.315
		0.083	0.271	0.259	0.239	0.160
		0.083	0.287	0.270	0.244	0.167
100/0.95/0.05	0.635	0.553	0.687	0.653	0.619	0.543
		0.101	0.230	0.230	0.236	0.262
		0.129	0.235	0.229	0.235	0.276
100/0.95/0.01	0.392	0.343	0.412	0.390	0.364	0.428
		0.086	0.313	0.298	0.278	0.259
		0.099	0.312	0.296	0.278	0.260
1000/0.80/0.05	0.281	0.298	0.295	0.289	0.279	0.288
		0.031	0.033	0.040	0.037	0.050
		0.035	0.036	0.041	0.037	0.050
1000/0.80/0.01	0.151	0.163	0.159	0.155	0.150	0.163
		0.022	0.023	0.024	0.022	0.033
		0.025	0.024	0.024	0.022	0.035
1000/0.90/0.05	0.467	0.491	0.492	0.485	0.470	0.469
		0.055	0.061	0.068	0.065	0.071
		0.060	0.066	0.070	0.065	0.071
1000/0.90/0.01	0.285	0.306	0.304	0.300	0.290	0.299
		0.050	0.061	0.063	0.061	0.061
		0.053	0.064	0.065	0.060	0.062
1000/0.95/0.05	0.649	0.668	0.679	0.667	0.652	0.631
		0.079	0.100	0.102	0.103	0.131
		0.081	0.104	0.103	0.103	0.132
1000/0.95/0.01	0.457	0.482	0.496	0.486	0.475	0.466
		0.089	0.115	0.114	0.112	0.122
		0.092	0.121	0.117	0.113	0.122

the pFDR. The results of the simulation for model I are given in Table 2 and those for model II are given in Table 3, respectively. However, the nonparametric estimators are still much better than the parametric estimator. This is mainly because the parametric model is a misspecified model. The first columns in Table 2 and Table 3 give the values of  $m$ ,  $\pi_0$ , and  $\alpha$ . The true pFDR values are given in the second columns of the tables. The three frequentist nonparametric estimators of pFDR have very similar performance in terms of their rMSE. When the size of individual tests is small (and hence the true pFDR is small as well), the Bayes estimator is slightly worse than the other three estimators in certain cases. However, in most cases for higher (and more commonly used)  $\alpha$ -values, the Bayes estimator of the pFDR is significantly better (in terms of rMSE) than Storey's estimators.

5.1 Application: Leukemia Data

The leukemia data set was published by Golub et al. (1999) for the classification of two leukemia, acute myeloid leukemia

(ALL) and acute lymphoblastic leukemia (AML). The two cancer types were identified based on their origins, lymphoid (lymph or lymphatic tissue related), and myeloid (bone marrow related), respectively. ALL could be further divided into B-cell and T-cell ALLs. There are 38 training samples (ALL B-cell: 19, ALL T-cell 8 and AML: 11) in the training set and 34 samples (ALL B-cell: 19, ALL T-cell 1 and AML: 14) in the testing set. For each sample, the expression values of 7129 genes were available. The leukemia data set has been widely studied in the literature for cancer classifications; see Tang and Zhang (2006) for a survey. Pan (2002) used this data set to compare various statistical procedures for the detection of significantly expressed genes. In this article, the leukemia data set was used to illustrate the Bayes nonparametric model by assessing the FDRs among statistically significantly expressed genes between two cancer types (ALL/AML) and between three classes (AML/ALL B-cell/ALL T-cell).

All 72 samples were preprocessed according to the procedure described in Tang and Zhang (2006): (i) thresholding

**Table 3**  
 Numerical summary of pFDR for model II: superscripts <sup>a,b,c</sup> represent mean, STD, and rMSE evaluated over 100 simulated samples, respectively

$m/\pi_0/\alpha$	True pFDR	Bayes	p-HALF	SPLINE	BOOT	MLE
100/0.80/0.05	0.281	0.316 <sup>a</sup>	0.336	0.318	0.292	0.284
		0.078 <sup>b</sup>	0.136	0.143	0.138	0.106
		0.085 <sup>c</sup>	0.146	0.147	0.138	0.105
100/0.80/0.01	0.123	0.154	0.169	0.161	0.148	0.146
		0.052	0.104	0.106	0.103	0.070
		0.061	0.114	0.112	0.105	0.073
100/0.90/0.05	0.464	0.453	0.539	0.510	0.475	0.444
		0.091	0.185	0.195	0.194	0.167
		0.091	0.199	0.200	0.193	0.167
100/0.90/0.01	0.231	0.249	0.310	0.289	0.270	0.271
		0.074	0.187	0.184	0.172	0.144
		0.075	0.202	0.192	0.176	0.148
100/0.95/0.05	0.642	0.533	0.684	0.652	0.624	0.538
		0.090	0.208	0.226	0.237	0.243
		0.141	0.212	0.226	0.236	0.263
100/0.95/0.01	0.358	0.317	0.463	0.435	0.413	0.369
		0.077	0.287	0.271	0.263	0.213
		0.087	0.305	0.280	0.268	0.212
1000/0.80/0.05	0.287	0.294	0.301	0.291	0.284	0.283
		0.022	0.024	0.034	0.030	0.036
		0.023	0.028	0.034	0.030	0.036
1000/0.80/0.01	0.127	0.135	0.135	0.130	0.127	0.129
		0.013	0.016	0.018	0.017	0.018
		0.015	0.018	0.019	0.017	0.018
1000/0.90/0.05	0.475	0.472	0.484	0.474	0.461	0.468
		0.040	0.047	0.054	0.052	0.051
		0.040	0.047	0.054	0.053	0.051
1000/0.90/0.01	0.246	0.257	0.252	0.247	0.240	0.246
		0.031	0.036	0.038	0.037	0.031
		0.033	0.036	0.038	0.037	0.031
1000/0.95/0.05	0.656	0.646	0.672	0.658	0.642	0.648
		0.054	0.076	0.083	0.084	0.073
		0.055	0.077	0.083	0.085	0.073
1000/0.95/0.01	0.407	0.425	0.434	0.425	0.415	0.417
		0.062	0.099	0.101	0.098	0.076
		0.064	0.102	0.102	0.098	0.076

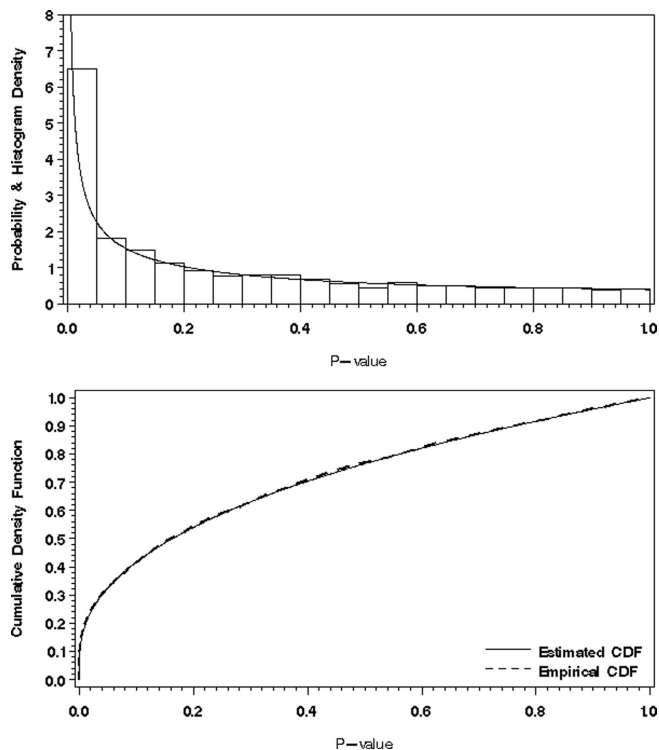
(floor of 100 and ceiling of 16,000), (ii) filtering (exclusion of genes with  $\max/\min \leq 5$  and  $\max - \min \leq 500$  across the samples), and (iii) base 10 logarithmic transformation. We further removed eight genes that had no variation across the training samples. The preprocessing resulted in 3563 remaining genes. The  $P$ -values were evaluated using all the training samples by the standard ANOVA  $t$ -tests (two classes) and  $F$ -tests (three classes) implemented in SAS proc glm procedures, assuming homoscedasticity of errors. We also computed the  $P$ -values using the Welch's (1951)  $t$ -tests and  $F$ -tests (SAS proc glm), that are robust to the assumptions of homogeneous within-group variance. As the overall performance was very similar under both assumptions, we report mainly the results based on the standard ANOVA models.

Table 4 lists the estimates of  $\pi_0$  obtained by the Bayesian nonparametric models, Storey's smoothing spline and bootstrap methods and the parametric MLE method. The estimated  $\pi_0$  was quite small after gene screen procedure. The  $\pi_0$

**Table 4**  
 Estimates of  $\pi_0$  for the leukemia data

	$t$ -tests		$F$ -tests	
	(two classes)		(three classes)	
	Standard	Welch	Standard	Welch
Bayes	0.485	0.479	0.394	0.390
Smoothing spline	0.489	0.476	0.398	0.405
Bootstrap	0.496	0.477	0.417	0.407
$P$ -value:0.5	0.540	0.526	0.461	0.450
MLE	0.510	0.505	0.395	0.414

estimated by Bayes procedure seemed to be a little smaller than that by Storey's methods. The estimated  $\pi_0$  for three classes was smaller than that for two classes, which may indicate that there is difference between ALL-B and ALL-T samples.



**Figure 1.** Estimated PDF and CDF of  $P$ -values from ANOVA  $F$ -tests.

Figure 1 displays the PDF and CDF estimated by the Bayes procedure. The estimated PDF matches well with the histogram density, and CDF matches very well with the empirical CDF, both indicating excellent model fitting.

**6. Discussion**

We introduced a flexible yet tractable model for the multiple testing problem. The proposed model allows us to compute posterior estimates of key components of the pFDR formula. Thus, we replaced the proportion of true null hypothesis and the probability of rejection by their posterior estimates to estimate the pFDR. Storey (2003) notes that the pFDR has a distinct Bayesian flavor even though inherently it is a frequentist quantity. Thus, it is natural to estimate the error rate under a Bayesian scheme that also allows the flexibility of viewing the  $P$ -value distribution under the alternative as a mixture of  $P$ -value distribution arising from different alternatives. Thus, the framework is broad enough to incorporate added uncertainty in the alternative values. Even though we have focused on the pFDR for the present article, the Bayesian scheme provides posterior estimates of other quantities such as the false negative rate, the probability of rejection, and the aggregate error rate as well. Thus, multiple aspects of the testing problem can be simultaneously analyzed in the proposed setup. The full potential of the Bayesian approach is apparent in the breadth of observable quantities. For example, the posterior probability that the  $i$ th null hypothesis is false is given by

$$\begin{aligned} \Pr(H_i = 1 | X) &= E \left[ \frac{f_1(X_i)}{f(X_i)} \middle| X \right] \\ &= 1 - E \left[ E \left( \frac{\pi_0}{1 - \pi_1 + \pi_1 \int g(X_i | a, b) dG(a, b)} \middle| \theta, \pi_1 \right) \middle| X \right] \\ &\cong 1 - M^{-1} \sum_{j=1}^M \frac{\pi_1^{(j)}}{f^{(j)}(X_i)}. \end{aligned} \tag{22}$$

Investigation of properties of such measures is part of ongoing research.

**7. Supplementary Materials**

A Web Appendix detailing how to perform posterior inference on functions such as the pFDR and the SAS Macros used to perform the simulation are available as supplementary materials under the Paper Information link at the *Biometrics* website <http://www.tibs.org/biometrics>.

ACKNOWLEDGEMENTS

The research of SG was partially supported by NSF grant number DMS-0349111. The research of AR was partially supported by NIH grant number 1R01GM075298-01. The authors are grateful to the editor and an anonymous referee for their helpful comments and suggestions.

REFERENCES

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics* **1**, 353–355.

Dalmasso, C., Broët, P., and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics* **21**(5), 660–668.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

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.

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

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

- Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B* **64**, 499–517.
- Ghosal, S., Roy, A., and Tang, Y. (2006). Posterior consistency of Dirichlet mixtures of beta densities in estimating positive false discovery rates. In *Invited paper in IMS collection: Festschrift in honor of P. K. Sen*, N. E. P. Balakrishnan and M. Silvapulle (eds), Forthcoming. Bethesda: Institute of Mathematical Statistics. [www4.stat.ncsu.edu/~sghosal/papers/FDR\\_IMS.pdf](http://www4.stat.ncsu.edu/~sghosal/papers/FDR_IMS.pdf).
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
- MacEachern, S. N. and Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–228.
- MacEachern, S. N. and Muller, P. (2000). Efficient MCMC schemes for robust model extensions using ecompassing Dirichlet process mixture models. In *Robust Bayesian Analysis*, D. R. Insua and F. Ruggeri (eds), 295–315. New York: Springer-Verlag.
- McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**(3), 413–422.
- Neal, R. M. (1998). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.
- Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**(4), 546–554.
- Parker, R. A. and Rothenberg, R. B. (1988). Identifying important results from multiple statistical tests. *Statistics in Medicine* **7**, 1031–1043.
- Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics* **20**(11), 1737–1745.
- Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236–1242.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**, 479–498.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA* **100**(16), 9440–9445.
- Tang, Y. and Zhang, H. (2006). Multiclass proximal support vector machines. *Journal of Computational and Graphical Statistics* **15**(2), 339–355.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701–1762.
- Tsai, C., Hsueh, H., and Chen, J. (2003). Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics* **59**(4), 1071–1081.
- Welch, B. L. (1951). The comparison of several mean values: An alternative approach. *Biometrika* **38**, 330–336.

Received December 2005. Revised November 2006.

Accepted February 2007.

Copyright of Biometrics is the property of Blackwell Publishing Limited and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.