Measuring Statistical Significance for Full Bayesian Methods in Microarray Analysis

Jing Cao[1] **and Song Zhang**[2]

[1] Southern Methodist University, USA

[2] U.T. Southwestern Medical Center, USA

June 25, 2009

Summary. Full Bayesian methods are useful tools to incorporate complex data structure in high-throughput data analysis. The Bayesian FDR, which is the posterior proportion of false positives relative to the total number of rejections, is widely used to measure statistical significance for full Bayesian methods in multiple comparisons. However, the Bayesian FDR is sensitive to prior specification and it is incomparable to the resampling-based FDR estimates employed by most frequentist and empirical Bayesian methods. In this paper, we propose an approach to objectively evaluating the statistical significance for full Bayesian methods in a resampling-based framework. The resulting predictive Bayesian FDR is shown to be robust to prior specification and it can produce a more accurate estimate of the true FDR. In addition, the approach provides a platform for the comparison of performance between full Bayesian methods and other methods. A simulation study and a real data example are presented.

Key words: FDR; Full Bayesian models; Microarray analysis; p-value; Statistical significance

1 Introduction

In the past decade, the advance of high-throughput technologies has presented statisticians with the challenge of testing thousands of features simultaneously. Without loss of generality, we consider multiple testing in the context of detecting differentially expressed (DE) genes in microarray experiments. Many statistical methods, including frequentist methods (Tusher *et al.*, 2001; Cui *et al.*, 2005), empirical Bayes methods (Efron *et al.*, 2001; Lonnstedt and Speed, 2002), and full Bayesian models (Newton *et al.*, 2004; Do *et al.*, 2005; Lewin *et al.*, 2006), have been proposed in microarray analysis. To control the error rate and compare performance across different methods, it is important to assess statistical significance such as the p-value and the false discovery rate (FDR). The FDR, which is the proportion of false rejections relative to the total number of rejections, has been shown to be more useful in balancing between the numbers of true/false positives in large-scale multiple testing (Benjamini and Hochberg, 1995; Storey, 2002; Genovese and Wasserman, 2002). It should be pointed out that the estimation of either significance measure requires the null distribution of the relevant test statistic.

For most frequentist and empirical Bayes (EB) methods, such as the SAM (Tusher *et al.*, 2001) and the posterior odds (Kendziorski *et al.*, 2003), the test statistic does not have a theoretical null distribution. Furthermore, for some methods whose test statistic does have a theoretical null distribution (eg., the moderated t-statistic, Smyth, 2004), the theoretical null might fail. Efron (2008) listed four reasons why it happens, including failed mathematical assumptions, unobserved covariates, correlation across arrays, and correlation across genes. When a test statistic does not have a theoretical null or it does not follow the theoretical null, a simple way is to construct an empirical null distribution using a simulated null data set.

For example, let $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_i, Y_i), i = 1, \cdots, n\}$ be the collection of expression mea-

surements from n genes, where X_i and Y_i are obtained under control and treatment, respectively. Most frequentist and EB testing procedures are based on a test statistic function $T(\cdot)$. Any gene with $T(X_i, Y_i)$ above a certain threshold is flagged as DE. Suppose we have a null data set, denoted as $(\boldsymbol{X}^0, \boldsymbol{Y}^0) = \{(X_k^0, Y_k^0), k = 1 \cdots, n^0\}$. Then the null distribution of the test statistic is approximated by $\{T(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$. We can estimate the p-value of gene i by

$$\widehat{p}_i = \frac{\sum_{k=1}^{n^0} I(T(X_k^0, Y_k^0) \geq T(X_i, Y_i))}{n^0}.$$

Storey *et al.* (2007) presented a procedure to estimate the FDR based on $\{T(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$.

Resampling-based procedures (i.e., boostrap or permutation) have being developed to generate the null data set (Tusher *et al.*, 2001; Storey and Tibshirani, 2003; Storey *et al.*, 2007). The generation of $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ is beyond the scope of this paper. We assume $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ to approximate the distribution of null gene expressions adequately well. Because the same null data set can be utilized by different testing methods to assess the p-value and the FDR, the resampling-based procedures provide a platform to objectively compare performance across different methods (Storey *et al.*, 2007).

In full Bayesian methods, the posterior probability of a gene being DE is often used as the test statistic and it does not have a theoretical null. To our knowledge, no approach has been proposed for full Bayesian methods to objectively evaluate statistical significance in the resampling-based framework. We will demonstrate that simply applying the Bayesian model on $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ does not produce a valid empirical null distribution for the posterior probabilities. Newton *et al.* (2004) proposed the Bayesian FDR (BFDR), which is the posterior proportion of false positives relative to the total number of rejections. It has been widely used in full Bayesian methods to assess statistical significance. However, the BFDR can be sensitive to prior specification, and its assessment of statistical significance

can be inaccurate. Because the BFDR is not based on the empirical null of the test statistic, it is incomparable to the resampling-based FDR (Storey *et al.*, 2007). In this paper, we present an approach to objectively assessing statistical significance for full Bayesian methods by constructing an empirical null for the test statistic using $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$. We show that this approach is robust to prior specification, and it allows a fair comparison between full Bayesian methods and other testing procedures.

The remainder of the paper is organized as follows. In Section 2 we introduce a generic full Bayesian model and review the BFDR. In Section 3 we present the approach to assessing statistical significance for full Bayesian methods in the resampling-based framework. In Section 4 and 5, we illustrate the proposed approach in a simulation study and a real microarray experiment, respectively. We conclude with a brief discussion in Section 6.

2 A Full Bayesian Model and the BFDR

We present a generic full Bayesian model for the detection of DE genes under two conditions. For $i = 1, \cdots, n$, it is assumed that $X_i \mid \theta_{0i}, \eta_i, \xi \sim [X_i \mid \theta_{0i}, \eta_i, \xi]$ and

$$Y_i \mid \theta_{0i}, \theta_{1i}, r_i, \eta_i, \xi \sim \begin{cases} [Y_i \mid \theta_{0i}, \eta_i, \xi], & \text{if } r_i = 0, \\ [Y_i \mid \theta_{1i}, \eta_i, \xi], & \text{if } r_i = 1. \end{cases} \tag{1}$$

We use $[U \mid V]$ to denote the conditional distribution (or density) of U given V. Thus X_i and Y_i share the same probability model when gene i is non-DE ($r_i = 0$), and they follow different models when gene i is DE ($r_i = 1$). We use θ_{0i}/θ_{1i} to denote the distinctive model parameters under $r_i = 0/1$, η_i to denote the gene-specific parameters shared under the two conditions, and ξ to denote the parameters shared by all genes. Depending on the specific model, θ_{0i}, θ_{1i}, η_i, and ξ might be vectors, scalers, or empty sets. Parameter r_i is usually modeled by a Bernoulli distribution, $r_i \mid p_r \sim \text{Bernoulli}(p_r)$, where p_r quantifies the prior belief about the proportion of DE genes. Integrating with respect to r_i, we have a mixture

model for Y_i,

$$Y_i \mid \theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r \sim (1-p_r)[Y_i \mid \theta_{0i}, \eta_i, \xi] + p_r[Y_i \mid \theta_{1i}, \eta_i, \xi].$$

Hierarchical priors are assumed for θ_{0i}, θ_{1i} and η_i to promote sharing of information among genes. A general prior form can be written as $\theta_{vi} \mid \theta_v \sim [\theta_{vi} \mid \theta_v]$ for $v = 0, 1$, and $\eta_i \mid \eta \sim [\eta_i \mid \eta]$ (Lonnstedt and Britton, 2005; Lewin *et al.*, 2006; Cao *et al.*, 2009). We use $[\theta_0, \theta_1, \eta, \xi, p_r]$ to denote the hyper-prior. Let $\boldsymbol{\Theta}$ be the collection of model parameters. The joint posterior distribution of $\boldsymbol{\Theta}$ is

$$[\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Y}] \propto \prod_{i=1}^{n} \{[X_i \mid \theta_{0i}, \eta_i, \xi][Y_i \mid \theta_{0i}, \theta_{1i}, r_i, \eta_i, \xi][r_i \mid p_r][\theta_{0i} \mid \theta_0][\theta_{1i} \mid \theta_1][\eta_i \mid \eta]\} \cdot [\theta_0, \theta_1, \eta, \xi, p_r]. \quad (2)$$

The posterior inference is usually based on $z_i = P(r_i = 1 \mid \boldsymbol{X}, \boldsymbol{Y})$, the posterior probability of gene i being DE. Muller *et al.* (2004, 2007) showed that under several loss functions that combine false positive/negative counts (rates), the optimal decision rule is based on z_i. Gene i is flagged as DE if $z_i > \lambda$, where λ is a threshold. It can be shown that

$$z_i = \int P(r_i = 1 \mid \theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r) \cdot [\theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}] d\theta_{0i} d\theta_{1i} d\eta_i d\xi dp_r,$$

with

$$P(r_i = 1 \mid \theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r) = \frac{p_r[Y_i \mid \theta_{1i}, \eta_i, \xi]}{(1-p_r)[Y_i \mid \theta_{0i}, \eta_i, \xi] + p_r[Y_i \mid \theta_{1i}, \eta_i, \xi]}, \quad (3)$$

and $[\theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}]$ is the marginal posterior distribution of $(\theta_{0i}, \theta_{1i}, \eta_i, \xi, p_r)$ derived from $[\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Y}]$.

The BFDR (Newton *et al.*, 2004) has been employed to control the error rate for full Bayesian methods. It is estimated by

$$\widehat{BFDR}(\lambda) = \frac{\sum_{i=1}^{n}(1-z_i)\delta_i}{D}, \quad (4)$$

where $\delta_i = I(z_i > \lambda)$ is the decision (1 for DE and 0 for non-DE) on gene i at cutoff λ, and $D = \sum_{i=1}^{n} \delta_i$ is the total number of rejections. Note that $1 - z_i$ is the posterior probability

of gene i being non-DE. The BFDR can be interpreted as the posterior proportion of false positives in the list of identified genes. The straightforward interpretation and easy computation based on z_i have brought popularity for the BFDR. The Bayesian model, however, in most cases only provides an approximation to the unknown expression distributions. As a result, z_i may not be an accurate estimate of the unknown probability, even though it can be a good test statistic quantifying the evidence for DE. Taking z_i at the face value of estimated probability, the BFDR may produce an inaccurate estimate of the error rate. In the simulation study we illustrate this point using a Bayesian model with two different priors. We show that the ordering of z_i's is robust to prior specification but the BFDR is not.

3 Assessing Statistical Significance Based on $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$

In the resampling-based framework, plugging $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ into the test statistic function $T(\cdot)$, we can use $\{T(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$ to approximate the null distribution of the test statistic. Following this rationale, we rewrite the test statistic z_i in full Bayesian methods as a function of (X_i, Y_i),

$$z_i = P(r_i = 1 \mid \boldsymbol{X}, \boldsymbol{Y}) = P(r_i = 1 \mid X_i, Y_i, \boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}) = h_i(X_i, Y_i),$$

where $(\boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}) = \{(X_j, Y_j) : j = 1, \cdots, n \text{ and } j \neq i\}$, and $h_i(\cdot)$, depending on $(\boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)})$, is a function uniquely defined for gene i. Based on the null data set $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$, the null distribution of z_i is approximated by $\{h_i(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$.

The above discussion suggests that simply applying the Bayesian model on $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ will not provide a valid empirical null for z_i. Let

$$z_k^0 = P(r_k^0 = 1 \mid \boldsymbol{X}^0, \boldsymbol{Y}^0) = P(r_k^0 = 1 \mid X_k^0, Y_k^0, \boldsymbol{X}_{(-k)}^0, \boldsymbol{Y}_{(-k)}^0) = h_k^0(X_k^0, Y_k^0),$$

where the superscript 0 in the expression indicates that it is for a gene in $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$. The

function $h_k^0(\cdot)$ is defined depending on $(\boldsymbol{X}_{(-k)}^0, \boldsymbol{Y}_{(-k)}^0)$, and it is incomparable to $h_i(\cdot)$, which indicates that the null distribution of z_i can not be approximated by $\{z_k^0, k = 1, \cdots, n^0\}$. We borrowed Table 1 from Do *et al.* (2005), which uses a simulation study to demonstrate how the inference on a gene with the same measurements (observed difference score) is affected by p_r, the true proportion of DE genes. We see that z_i increases/decreases when p_r increases/decreases. This observation suggests that using $\{z_k^0, k = 1, \cdots, n^0\}$ to approximate the null distribution of z_i will result in a gross inflation of significance, because the proportion of DE genes is often much lower in $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ than that in $(\boldsymbol{X}, \boldsymbol{Y})$.

Table 1: Comparison of z_i under different p_r

	Observed difference scores										
p_r	-5.0	-4.0	-3.0	-2.0	-1.0	-0.0	1.0	2.0	3.0	4.0	5.0
0.6	1.00	1.00	0.98	0.87	0.46	0.19	0.43	0.85	0.98	1.00	1.00
0.2	0.94	0.90	0.75	0.41	0.14	0.07	0.13	0.44	0.91	0.93	0.96
0.05	0.46	0.42	0.27	0.11	0.05	0.03	0.04	0.10	0.28	0.43	0.50

For $i = 1, \cdots, n$, the null distribution of z_i can be approximated by $\{h_i(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$. This approach to constructing the empirical null for z_i $(i = 1, \cdots, n)$ poses a great computational challenge. Specifically, we need to fit the Bayesian model on $n \times n^0$ data sets, i.e., $\{X_k^0, Y_k^0, \boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}\}$ for $k = 1, \cdots, n^0$ and $i = 1, \cdots, n$. Because n and n^0 are both on a large scale and most full Bayesian models require MCMC simulation, the above procedure is usually computationally infeasible.

To reduce the computational burden, we propose to approximate $h_i(X_k^0, Y_k^0)$ by

$$s(X_k^0, Y_k^0) = \int P(r_k^0 = 1 \mid \theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \xi, p_r)[\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0, \theta_1, \eta, \xi, p_r] \cdot [\theta_1, \theta_0, \eta, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}] d\theta_{0k}^0 d\theta_{1k}^0 d\eta_k^0 d\theta_0 d\theta_1 d\eta d\xi dp_r. \quad (5)$$

Here $(r_k^0, \theta_{0k}^0, \theta_{1k}^0, \eta_k^0)$ denotes the model parameters for (X_k^0, Y_k^0), and $[\theta_1, \theta_0, \eta, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}]$

is the marginal posterior distribution of $(\theta_1, \theta_0, \eta, \xi, p_r)$ obtained from $[\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Y}]$, which represents the statistical learning from $(\boldsymbol{X}, \boldsymbol{Y})$. Furthermore,

$$[\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0, \theta_1, \eta, \xi, p_r] \propto [X_k^0 \mid \theta_{0k}^0, \eta_k^0, \xi][Y_k^0 \mid \theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \xi, p_r] \cdot [\theta_{0k}^0 \mid \theta_0][\theta_{1k}^0 \mid \theta_1][\eta_k^0 \mid \eta]$$

is the conditional distribution of $(\theta_{0k}^0, \theta_{1k}^0, \eta_k^0)$ given $(\theta_0, \theta_1, \eta, \xi, p_r)$ and (X_k^0, Y_k^0), and $P(r_k^0 = 1 \mid \theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \xi, p_r)$ is similarly defined as in (3). We interpret $s(X_k^0, Y_k^0)$ as the predictive probability of a gene with measurements (X_k^0, Y_k^0) being DE given $(\boldsymbol{X}, \boldsymbol{Y})$.

Theorem 1 Under the assumed Bayesian model (1), $h_i(X_k^0, Y_k^0) - s(X_k^0, Y_k^0) \to 0$ for $i = 1, \cdots, n$, as $n \to +\infty$.

Proof. See Appendix.

We use $\{s(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$ to approximate the null distribution of of z_i for $i = 1, \cdots, n$. In the following we provide an algorithm to estimate $\{s(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$, which only requires fitting the Bayesian model once based on $(\boldsymbol{X}, \boldsymbol{Y})$.

Algorithm 1

1. For $l = 1, \cdots, L$,
   (a) Simulate $(\theta_0^{(l)}, \theta_1^{(l)}, \eta^{(l)}, \xi^{(l)}, p_r^{(l)})$ from $[\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Y}]$.
   (b) For $k = 1, \cdots, n^0$, simulate $(\theta_{0k}^{0(l)}, \theta_{1k}^{0(l)}, \eta_k^{0(l)})$ from the conditional distribution $[\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0^{(l)}, \theta_1^{(l)}, \eta^{(l)}, \xi^{(l)}, p_r^{(l)}]$.
2. For $k = 1, \cdots, n^0$, approximate $s(X_k^0, Y_k^0)$ by

$$\widehat{s}(X_k^0, Y_k^0) = \frac{1}{L} \sum_{l=1}^{L} P(r_k^0 = 1 \mid \theta_{0k}^{0(l)}, \theta_{1k}^{0(l)}, \eta_k^{0(l)}, \xi^{(l)}, p_r^{(l)}).$$

Step 1a is usually accomplished by the MCMC simulation for the Bayesian model based on $(\boldsymbol{X}, \boldsymbol{Y})$. In Step 1b, if $[\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0^{(l)}, \theta_1^{(l)}, \eta^{(l)}, \xi^{(l)}, p_r^{(l)}]$ does not have a closed form, a nested MCMC simulation given data (X_k^0, Y_k^0) can be employed to simulate $(\theta_{0k}^{0(l)}, \theta_{1k}^{0(l)}, \eta_k^{0(l)})$.

Different measures of statistical significance can then be computed based on $\{\widehat{s}(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$. For example, the p-value of gene i is approximated by

$$\widehat{P}_i = \frac{\sum_{k=1}^{n^0} I(\widehat{s}(X_k^0, Y_k^0) > z_i)}{n^0}.$$

Using the procedure in Storey *et al.* (2007), we can estimate the FDR by

$$\widehat{PBFDR}(\lambda) = \frac{\widehat{\pi}_0 n \sum_{k=1}^{n^0} I(\widehat{s}(X_k^0, Y_k^0) > \lambda)}{n^0 \sum_{i=1}^{n} I(z_i > \lambda)}, \tag{6}$$

where $\hat{\pi}_0$ is the estimated proportion of true nulls, computed based on $\widehat{P}_i$ (Storey and Tibshirani, 2003). The PBFDR stands for the predictive Bayesian FDR, indicating that it is computed based on $s(X_k^0, Y_k^0)$, the predictive probability of a gene with measurements (X_k^0, Y_k^0) being DE given $(\boldsymbol{X}, \boldsymbol{Y})$.

4 Simulation Study

In this section we compared the performance of the PBFDR and the BFDR. We used the full Bayesian model in Cao *et al.* (2009) as an example. Let $X_i = (x_{i1}, \cdots, x_{im})'$ and $Y_i = (y_{i1}, \cdots, y_{ig})'$ be the expression measurements from the ith $(i = 1, \cdots, n)$ gene. Here m and g denote the number of arrays under the control and treatment, respectively. Through a proper transformation, x_{ij} and y_{ij} are modeled by normal distributions: $x_{ij} \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2)$ and

$$y_{ij} \mid \mu_i, \Delta_i, \sigma_i^2, r_i \sim \begin{cases} N(\mu_i, \sigma_i^2), & \text{if } r_i = 0, \\ N(\mu_i + \Delta_i, \sigma_i^2), & \text{if } r_i = 1, \end{cases}$$

where $r_i = 0/1$ indicates that gene i is non-DE/DE. It is assumed that $\Delta_i \sim N(0, s_\Delta^2)$ and $r_i \mid p_r \sim Bernoulli(p_r)$. To encourage sharing of information, a mixture structure is introduced on the variances, $\sigma_i^2 \mid \sigma_0^2, p_v \sim (1-p_v)\delta(\sigma_0^2) + p_v IG(a_\sigma, b_\sigma)$. Here p_v is the mixing probability, $\delta(\sigma_0^2)$ denotes a point mass at σ_0^2, and $IG(a_\sigma, b_\sigma)$ denotes an inverse Gamma distribution parameterized such that the mean equals $b_\sigma/(a_\sigma - 1)$. The Bayesian model

includes hyper-priors, $\mu_i \sim N(0, s_\mu^2)$, $\sigma_0^2 \sim IG(a_0, b_0)$, $p_r \sim U(0,1)$, and $p_v \sim U(0,1)$. More details can be found in Cao *et al.* (2009). This model fits in the generic Bayesian model framework in (1).

To demonstrate how prior specification affects the PBFDR and the BFDR differently, we considered two specifications of $(a_\sigma, a_0, b_\sigma, b_0)$, denoted as Prior 1 and Prior 2. Prior 1 is data dependent, where we set $a_\sigma = a_0 = 2.0$ and both b_σ and b_0 (the prior mean) equal to the average of the pooled sample variances over all genes. Prior 1 is a diffuse prior with an infinite variance. For Prior 2, we set $a_\sigma = a_0 = b_\sigma = b_0 = 0.01$, which is also a commonly used diffuse prior.

The simulated dataset contains $n = 1000$ genes and 6 replicates per gene per condition. We generated x_{ij} and y_{ij} using $\mu_i = 0$, $p_r = 0.1$, $\Delta_i \sim N(0,1)$, and $\sigma_i^2 \sim IG(4,1)$. For each simulated data set, we generated the null data set using the permutation procedure described in Storey and Tibshirani (2003). We repeated the simulation 100 times. MCMC simulation was conducted to fit the Bayesian model.

Figure 1 plots the estimated z_i's under Prior 1 versus those under Prior 2 based on one simulation. It shows that z_i can take different values under different priors, but the ordering of z_i is well preserved (the correlation coefficient is 0.989). Thus z_i as a test statistic to quantify the DE evidence is robust to prior specification. However, z_i as the estimate of probability of a gene being DE is sensitive to prior specification.

Figure 2 plots the true FDR, the BFDR, and the PBFDR versus the the total number of rejections under Prior 1 and Prior 2, averaged over 100 simulations. The two curves of the true FDR are very close, suggesting that, as a testing procedure, the Bayesian model is robust to prior specification. The BFDR, which is calculated based on z_i, deviates from the true FDR and changes considerably between Prior 1 and Prior 2. By comparison, the PBFDR almost overlaps with the true FDR. We have computed the PBFDR under a number

of different priors and obtained similar results. It suggests that the PBFDR is robust to prior specification and it provides a reliable estimate of the true FDR.

5 Real Data Example

The real data comes from a microarray study comparing the gene expressions of breast cancer tumors with *BRCA1* mutations, *BRCA2* mutations, and sporadic tumors (Hedenfalk et al., 2001). The data set is available at http://research.nhgri.nih.gov/microarray/NEJM_Supplement. Here we only considered the *BRCA1* group and the *BRCA2* group. There are 3226 genes, with 7 arrays in the *BRCA1* group and 8 arrays in the *BRCA2* group. We analyzed the data on the log_2 scale. Following Storey and Tibshirani (2003), we eliminated genes with aberrantly large expression values (>20), which left us with measurements on $n = 3169$ genes.

We analyzed the data using the full Bayesian model in Cao *et al.* (2009). As in the simulation study, we estimated the PBFDR and the BFDR under Prior 1 and Prior 2. Figure 3 plots the FDR estimates versus the total number of rejections based on the breast cancer data. The PBFDR is relatively stable under the two priors. The BFDR deviates from the PBFDR and it changes substantially with different prior specifications.

Figure 3 also includes the permutation-based FDR for the SAM statistic (Tusher *et al.*, 2001), which follows the PBFDR closely. It suggests that the full Bayesian model and the SAM method have similar performance. This observation is supported by the large number of genes flagged by both methods. Among the top 100, 200, 300, 400, 500 selected genes, the number of genes selected by both the SAM and the Bayesian model (under Prior 1) are 78, 167, 267, 355, and 440, respectively.

6 Discussion

Full Bayesian methods are useful tools to handle complex data structure in high-throughput data analysis. In this paper we have proposed a generic approach to objectively evaluate statistical significance for full Bayesian methods in the resampling-based framework. Specifically, we constructed an empirical null distribution for the posterior probability of a gene being DE, based on which, commonly used significance measures, such as the p-value and the FDR, can be estimated following the same procedure employed by frequentist and EB methods. The resulting PBFDR is robust to prior specification and can produce accurate estimate of the true FDR. In addition, when computed based on the same null data set, the PBFDR is comparable to the resampling-based FDR estimate. It allows researchers to objectively compare the performance of full Bayesian methods with other frequentist and EB methods.

We have also proposed an algorithm to approximate the empirical null for full Bayesian methods. The algorithm only requires fitting the full Bayesian model once, which reduces the computational burden tremendously. However, the evaluation of the PBFDR is more computationally intensive than the BFDR. We plan to develop more efficient algorithms in future research.

REFERENCES

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.

Cao, J., Xie, X., Zhang, S., Whitehurst, A., and White, M. (2009). Bayesian optimal discovery procedure for simultaneous significance testing. *BMC Bioinformatics* **10**:5.

Cui X., Hwang, J.T.G., Qiu, J., Blades, N.J., and Churchill, G.A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6**, 59-75.

Do, K., Muller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Applied Statistics* **54**, 627-644.

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.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science* **23**, 1-47.

Ferguson, T.S.(1996). *A Course in Large Sample Theory.* Chapman & Hall, London.

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.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., and others (2002). Gene-expression profiles in hereditary breast cancer . *New England Journal of Medicine* **344**, 539-548.

Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**, 3899-3914.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics* **62**, 1-9.

Lonnstedt, I. and Britton, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics* **6**:279-291.

Lonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31-46.

Muller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* **99**, 990-1001.

Muller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In Bernardo, J. *et al.* (eds.) *Bayesian Statistics* Vol 8. Oxford University Press, Oxford.

Newton, M.A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **4**:155-176.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**(1):3.

Storey, J.D. (2002). A direct approach to false discovery rate. *Journal of the Royal Statistical Society, Series B* **64**:479-498.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445.

Storey, J.D., Dai, J.Y., and Leek, J.T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8**:414-432.

Tusher, V.G., Tibshirani, R., Chu. G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences* **98**:5116-5121.

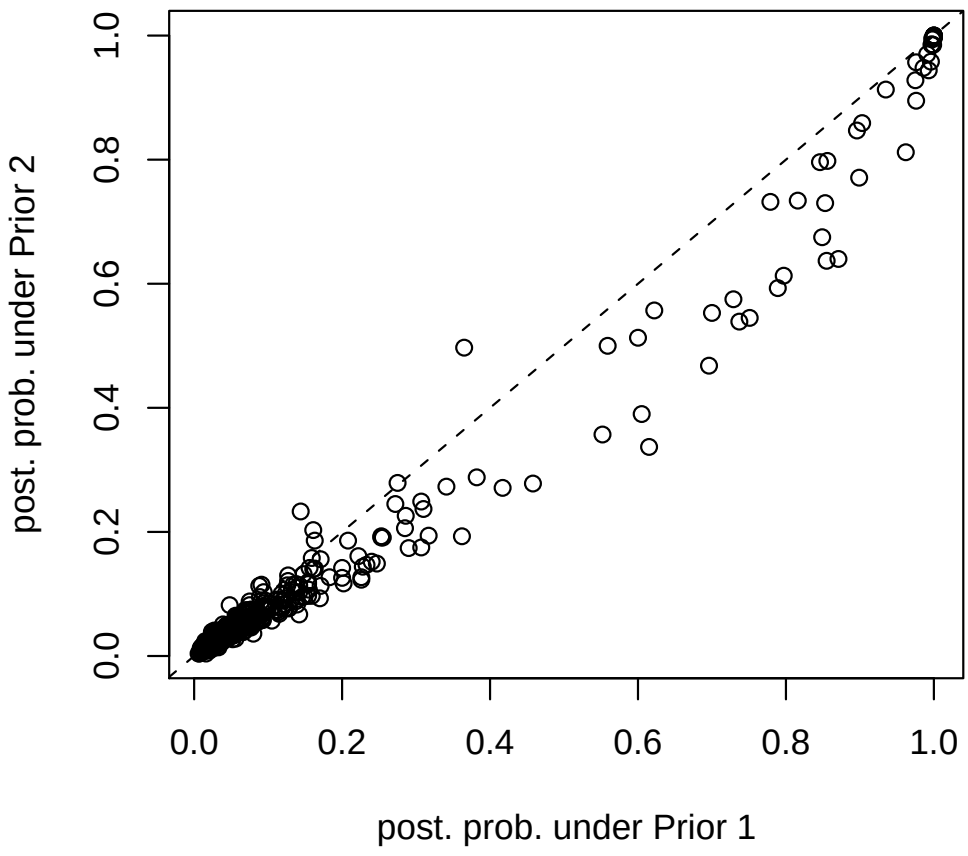

Figure 1: The scatter plot of the estimated z_i's under Prior 1 and Prior 2 in the simulation study.

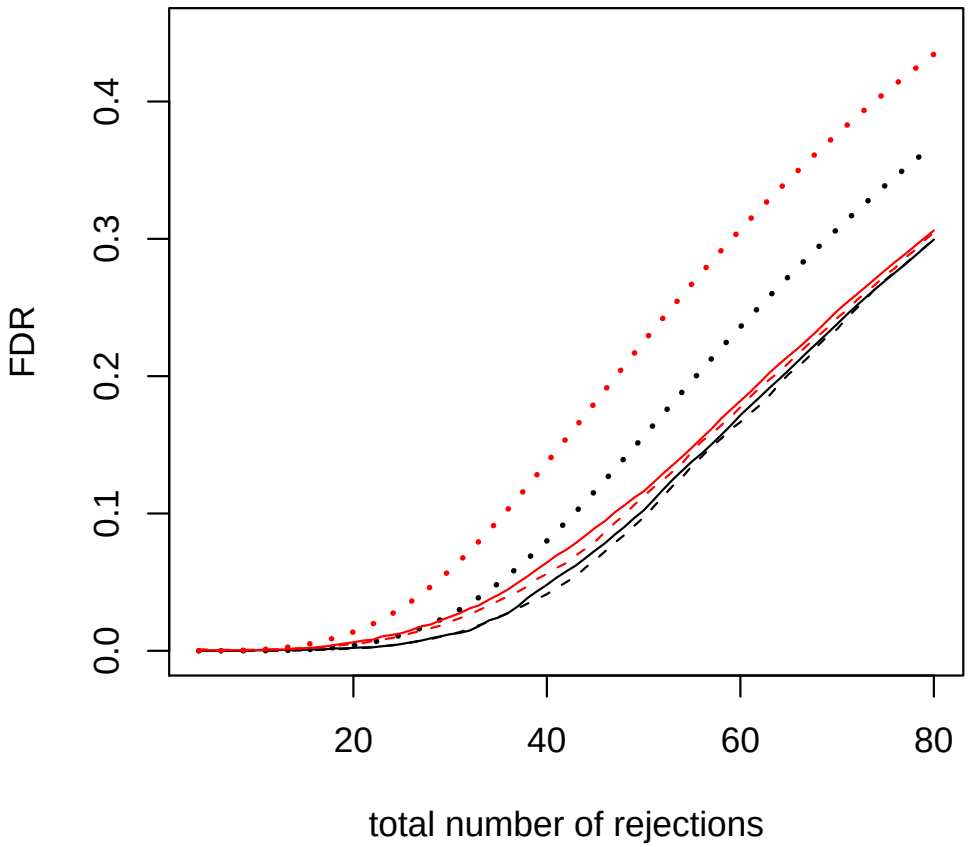

Figure 2: Comparison of the true FDR, the BFDR, and the PBFDR in the simulation study. The black curve is for Prior 1 and the red curve is for Prior 2. The solid curve denotes the true FDR, the dotted curve denotes the BFDR, and the dashed curve denotes the PBFDR.

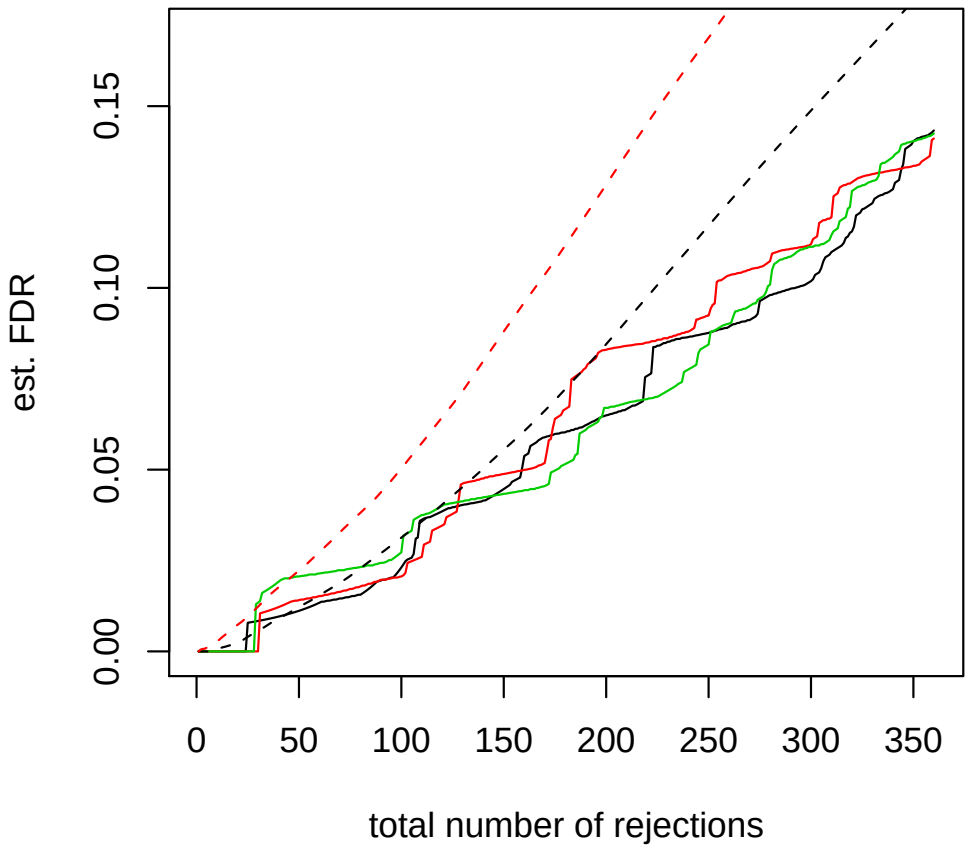

Figure 3: The plot depicts the estimated FDR versus the total number of rejections in the breast cancer data. The black solid curve is for the PBFDR under Prior 1, the black dashed curve is for the BFDR under Prior 1, the red solid curve is for the PBFDR under Prior 2, the red dashed curve is for the BFDR under Prior 2, and the green curve is for the permutation-based FDR of the SAM.

Appendix: Proof of Theorem 1

We rewrite $s(X_k^0, Y_k^0)$ in (5) as

$$s(X_k^0, Y_k^0) = \int w(\theta_0, \theta_1, \eta, \xi, p_r) \cdot [\theta_0, \theta_1, \eta, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}] d\theta_0 d\theta_1 d\eta d\xi dp_r, \tag{7}$$

where

$$w(\theta_0, \theta_1, \eta, \xi, p_r) = \int P(r_k^0 = 1 \mid \theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \xi, p_r)[\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0, \theta_1, \eta, \xi, p_r] d\theta_{0k}^0 \theta_{1k}^0 \eta_k^0.$$

Note that

$$\begin{aligned} h_i(X_k^0, Y_k^0) = \int & P(r_k^0 = 1 \mid \theta_{0k}^0, \theta_{0k}^0, \eta_k^0, \xi, p_r) \\ & \cdot [\theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \theta_0, \theta_1, \eta, \xi, p_r \mid X_k^0, Y_k^0, \boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}] d\theta_{0k}^0 d\theta_{1k}^0 d\eta_k^0 d\theta_0 d\theta_1 d\eta d\xi dp_r, \end{aligned}$$

where

$$\begin{aligned} & [\theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \theta_0, \theta_1, \eta, \xi, p_r \mid X_k^0, Y_k^0, \boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}] = \\ & \qquad [\theta_{0k}^0, \theta_{1k}^0, \eta_k^0 \mid X_k^0, Y_k^0, \theta_0, \theta_1, \eta, \xi, p_r] g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r), \end{aligned}$$

and

$$\begin{aligned} g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r) &= \frac{\{\prod_{j \neq i}^n [X_j, Y_j \mid \theta_0, \theta_1, \eta, \xi, p_r]\}[X_k^0, Y_k^0 \mid \theta_0, \theta_1, \eta, \xi, p_r][\theta_0, \theta_1, \eta, \xi, p_r]}{[X_k^0, Y_k^0, \boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)}]} \\ &\propto \{\prod_{j \neq i}^n [X_j, Y_j \mid \theta_0, \theta_1, \eta, \xi, p_r]\}\{[X_k^0, Y_k^0 \mid \theta_0, \theta_1, \eta, \xi, p_r][\theta_0, \theta_1, \eta, \xi, p_r]\}. \end{aligned}$$

Then we have

$$h_i(X_k^0, Y_k^0) = \int w(\theta_0, \theta_1, \eta, \xi, p_r) g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r) d\theta_0 d\theta_1 d\eta d\xi dp_r. \tag{8}$$

Note that $g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r)$ can be considered as the posterior distribution of $(\theta_0, \theta_1, \eta, \xi, p_r)$ given data $(\boldsymbol{X}_{(-i)}, \boldsymbol{Y}_{(-i)})$, with the likelihood being $\prod_{j \neq i}^n [X_j, Y_j \mid \theta_0, \theta_1, \eta, \xi, p_r]$ and the prior being $[X_k^0, Y_k^0 \mid \theta_0, \theta_1, \eta, \xi, p_r][\theta_0, \theta_1, \eta, \xi, p_r]$.

The Bernstein-von Mises theorem indicates that as $n \to +\infty$, both $[\theta_0, \theta_1, \eta, \xi, p_r \mid \boldsymbol{X}, \boldsymbol{Y}]$ and $g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r)$ converge almost surely to the same normal distribution centered at the MLE of $(\theta_0, \theta_1, \eta, \xi, p_r)$, where the normal density is denoted as $f(\theta_0, \theta_1, \eta, \xi, p_r)$.

Based on the convergence theorem in large sample theory (Ferguson, 1996), both $s(X_k^0, Y_k^0)$ in (7) and $h_i(X_k^0, Y_k^0)$ in (8) converge to the same quantity C,

$$C = \int w(\theta_0, \theta_1, \eta, \xi, p_r) f(\theta_0, \theta_1, \eta, \xi, p_r) d\theta_0 d\theta_1 d\eta d\xi dp_r,$$

if $w(\theta_0, \theta_1, \eta, \xi, p_r)$ is a bounded and continuous function. Then we have $h_i(X_k^0, Y_k^0) - s(X_k^0, Y_k^0) \to 0$ for $i = 1, \cdots, n$, as $n \to +\infty$.