Bayesian Analysis (2010) **5**, Number 2, pp. 413–428

Measuring Statistical Significance for Full Bayesian Methods in Microarray Analyses

Jing Cao* and Song Zhang†

Abstract. Full Bayesian methods are useful tools to account for complex data structures in high-throughput data analyses. The Bayesian FDR, which is the posterior proportion of false positives relative to the total number of rejections, has been widely used to measure statistical significance for full Bayesian methods in microarray analyses. 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 a computationally efficient algorithm to evaluate the statistical significance for full Bayesian methods in the resampling-based framework. The resulting predictive Bayesian FDR is robust to prior specifications and it can produce a more accurate estimate of error rate. In addition, the proposed approach provides a general framework for the objective comparison of performance between full Bayesian methods and the other frequentist and empirical Bayes methods in microarray analyses, which has been an unaddressed issue. A simulation study and a real data example are presented.

Keywords: FDR, Bayesian models, microarray analysis, resampling method, statistical significance

1 Introduction

The advance of high-throughput technologies has presented statisticians with the challenge of testing thousands of genes simultaneously. In this paper, 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 2005) , empirical Bayes (EB) methods (Efron et al. 2001; Lönnstedt 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 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 p-value or FDR requires the null distribution of the relevant test statistic.

*Department of Statistical Science, Southern Methodist University, Dallas, TX, http://smu.edu/statistics/faculty/cao.htm

†Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, mailto:song.zhang@utsouthwestern.edu

 DOI:10.1214/10-BA608

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

Let $(\boldsymbol{X}, \boldsymbol{Y}) = \{(X_i, Y_i), i = 1, \cdots, n\}$ be the collection of expression measurements 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 $(\boldsymbol{X}^0, \boldsymbol{Y}^0) = \{(X_k^0, Y_k^0), k = 1 \cdots, n^0\}$. Note that (X_k^0, Y_k^0) represents expression measurements for any non-DE gene in the experiment, while (X_i, Y_i) denotes the observed expression values for gene i. 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 (eg., bootstrap and permutation) have been developed to generate the null data set (Rubin 1981; 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 that through a certain procedure, we have a null data set $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ which approximates 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 FDR, the resampling-based procedures provide a general framework to objectively compare performance across different methods (Storey et al. 2007).

For full Bayesian methods, the posterior probability of a gene being DE is often used as the test statistic and its null distribution does not have a closed form. 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. Furthermore, estimated based on the posterior probabilities, the BFDR is incomparable to the resampling-based FDR adopted by most frequentist and EB methods (Storey et al. 2007). Objectively comparing the performance between full Bayesian methods and the other methods has

remained a challenge in microarray analyses. In this paper, we present an approach to assessing statistical significance for full Bayesian methods in the resampling-based framework. 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 Generic 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 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 is the proportion of DE genes. Such a specification implies an equivalent model for Y_i in the mixture form,

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

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 $\eta_i \mid \eta \sim [\eta_i \mid \eta]$ (Lönnstedt 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, which includes $\theta_{0i}, \theta_{1i}, \eta_i, r_i, \theta_0, \theta_1, \eta, \xi, p_r$. The joint posterior distribution of $\boldsymbol{\Theta}$ is

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

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. Müller 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

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

with

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

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 widely employed to control the error rate for full Bayesian methods. It is estimated by

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

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. A Bayesian model, however, in most cases only provides an approximation to the unknown true expression distribution. As a result, z_i may not accurately estimate the unknown probability, even though it may serve as a good test statistic in screening DE genes. Taking z_i at the face value of estimated probability, the BFDR may produce an inaccurate assessment 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.

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

In most frequentist and EB microarray screening procedures, the inference is based on a certain test statistic function $T(\cdot)$. Any gene with $T(X_i, Y_i)$ above a certain threshold is flagged as DE. Under the resampling-based framework, by plugging $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ into $T(\cdot)$, researchers have used $\{T(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$ to approximate the null distribution of the test statistic. It forms the basis for the assessment of statistical significance, such as p-value and FDR. 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 each approximated by $\{h_i(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$, for $i = 1, \cdots, n$.

The definition of $h_i(X_k^0, Y_k^0)$ suggests that simply applying the Bayesian model on $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ will not produce a valid empirical null distribution for z_i. To further explain it, 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 indicates that the data and parameters are for null genes in $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$. Different from $h_i(\cdot)$, the definition of $h_k^0(\cdot)$ depends on $(\boldsymbol{X}_{(-k)}^0, \boldsymbol{Y}_{(-k)}^0)$, which means 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 estimation of z_i is affected by p_r, the true proportion of DE genes, given the same difference score (defined as $\bar{X}_i - \bar{Y}_i$). Each column of Table 1 compares the estimated z_i under different p_r. We notice 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 in significance, because the proportion of DE genes is usually much lower in $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ than 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 each be approximated by $\{h_i(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$. Such an approach to constructing the empirical null distribution of z_i 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 usually large (in thousands) and most full Bayesian models require MCMC simulation, the above procedure is computationally infeasible.

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

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

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 $[\boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Y}]$, representing 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 $P(r_k^0 = 1 \mid \theta_{0k}^0, \theta_{1k}^0, \eta_k^0, \xi, p_r, Y_k^0)$ 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), $h_i(X_k^0, Y_k^0) - s(X_k^0, Y_k^0) \xrightarrow{a.s.} 0$ uniformly for each i, $i = 1, \cdots, n$, as $n \to +\infty$.

Proof. See Appendix.

Because $s(X_k^0, Y_k^0)$ provides good approximation to $h_i(X_k^0, Y_k^0)$, we propose to use $\{s(X_k^0, Y_k^0), k = 1, \cdots, n^0\}$, instead of $\{h_i(X_k^0, Y_k^0),\ k = 1, \cdots, n^0\}$, to approximate the null distribution of z_i $(i = 1, \cdots, n)$. In the following we provide an algorithm to efficiently 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 iteration $l = 1, \cdots, L$,
   (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 $k = 1, \cdots, n^0$, estimate $s(X_k^0, Y_k^0)$ by

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

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 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 null genes, 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 $(\boldsymbol{X}, \boldsymbol{Y})$.

3 Simulation Study

In this section we compare 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 $r_i \mid p_r \sim \text{Bernoulli}(p_r)$. 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 means) 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 data set 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 screen DE genes 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 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 estimation of

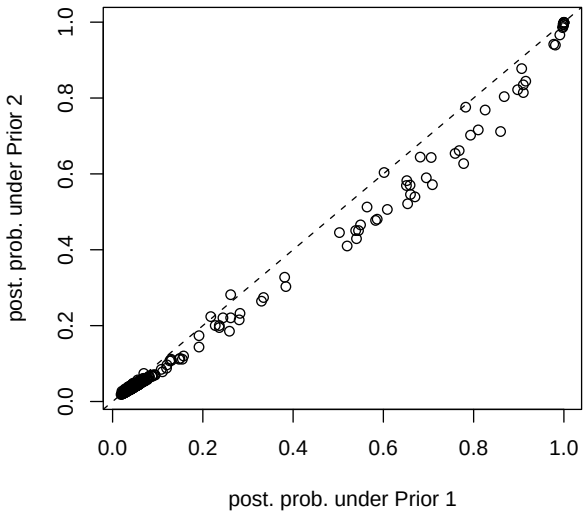

Figure 1: The scatter plot of the estimated z_i's under Prior 1 and Prior 2 in the simulation study.

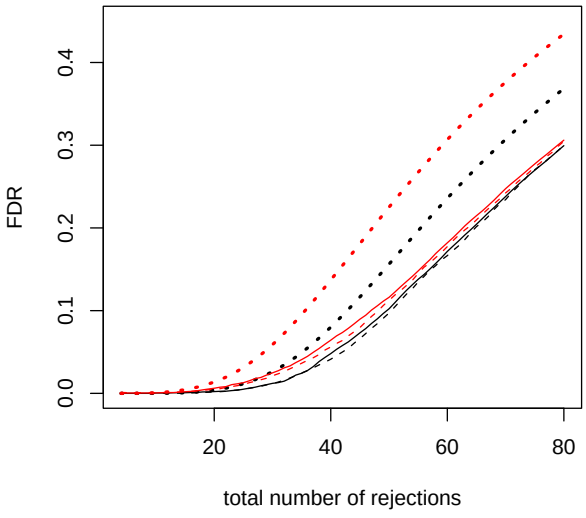

Figure 2: Comparison of the true FDR, the BFDR, and the PBFDR in the first simulation study. The black curve is for Prior 1 and the red curve is for Prior 2. The solid curve denotes the true FDR, the dotted curve denotes the BFDR, and the dashed curve denotes the PBFDR.

the true FDR.

In the first simulation study, the gene expression measurements are generated based on the assumed model. We conducted another simulation study with a more realistic gene expression distribution. Specifically, let X_i and Y_i be the observed expression levels for gene i from a real microarray study. Define the residual vector $\mathbf{e_i} = (e_{i1}, \cdots, e_{i,m+g})'$ by

$$e_{il} = \begin{cases} x_{il} - \bar{x}_i, & \text{for } l = 1, \cdots, m, \\ y_{i,(l-m)} - \bar{y}_i, & \text{for } l = m+1, \cdots, m+g, \end{cases}$$

where $\bar{x}_i = \sum_{j=1}^{m} x_{ij}/m$ and $\bar{y}_i = \sum_{j=1}^{g} y_{ij}/m$. Then $\mathbf{e_i}$ can be considered as a set of random errors sampled based on the true gene expression distribution (Storey et al. 2007). We simulated 100 data sets according to the following steps. For iteration t $(t = 1, \cdots, 100)$ and gene i $(i = 1, \cdots, n)$,

1. obtain a random permutation of $(e_{i1}, \cdots, e_{i,m+g})$, denoted by $\mathbf{e_i}^{(t)}$;
2. generate $\Delta_i^{(t)}$ as described in the previous simulation study;
3. for $j = 1, \cdots, m$, compute $x_{ij}^{(t)} = e_{ij}^{(t)}$, and for $k = 1, \cdots, g$, compute $y_{ik}^{(t)} = \Delta_i^{(t)} + e_{i,(m+k)}^{(t)}$, where $e_{ij}^{(t)}$ is the jth element of $\mathbf{e_i}^{(t)}$.

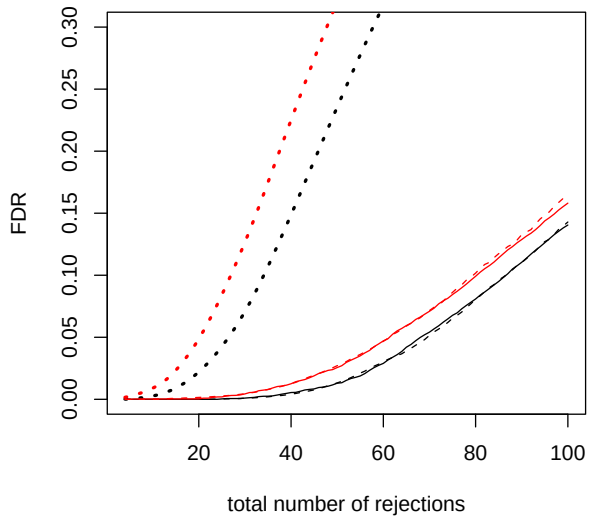

Figure 3: Comparison of the true FDR, the BFDR, and the PBFDR in the second simulation study. The black curve is for Prior 1 and the red curve is for Prior 2. The solid curve denotes the true FDR, the dotted curve denotes the BFDR, and the dashed curve denotes the PBFDR.

The real data comes from a study comparing the gene expressions of breast cancer tumors with *BRCA1* mutations, *BRCA2* mutations, and sporadic tumors (Hedenfalk

et al. 2001), 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. Figure 3 compares the true FDR, the BFDR, and the PBFDR under Prior 1 and Prior 2, where the residual vector $\mathbf{e_i}$ was constructed based on the breast cancer data. We kept the same replicate number in the experiment, with 7 replicates per gene in one group and 8 replicates in the other group.

The BFDR is sensitive to prior specification and its assessment of statistical significance can be inaccurate. In the second simulation, the BFDR deviates substantially from the true FDR because the assumed model was not an adequate fit to the data (Figure 3). By comparison, the PBFDR under either prior closely follows the true FDR. This is because the PBFDR utilizes an (artificially) augmented data set that includes $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$ to construct the empirical null distribution for z_i. The null data can be generated independent of the working model. With a well constructed null data set, the empirical null distribution of z_i can characterize the behavior of test statistics for non-DE genes, whether the working model is true or not, which leads to an accurate and robust assessment of FDR.

4 Real Data Example

We analyzed the breast cancer microarray data (Hedenfalk et al. 2001) 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. The null data set was generated using the permutation procedure (Storey and Tibshirani 2003). Figure 4 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.

We also used Figure 4 to demonstrate that the PBFDR allows objectively comparing the performance between full Bayesian methods and frequentist and EB methods. Specifically, we plot the permutation-based FDR for the SAM statistic (Tusher et al. 2001). It follows the PBFDR closely, suggesting that the full Bayesian model and the SAM method have similar performance. This conclusion is supported by the large number of overlapping 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.

5 Discussion

Full Bayesian methods are useful tools to handle complex data structures 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

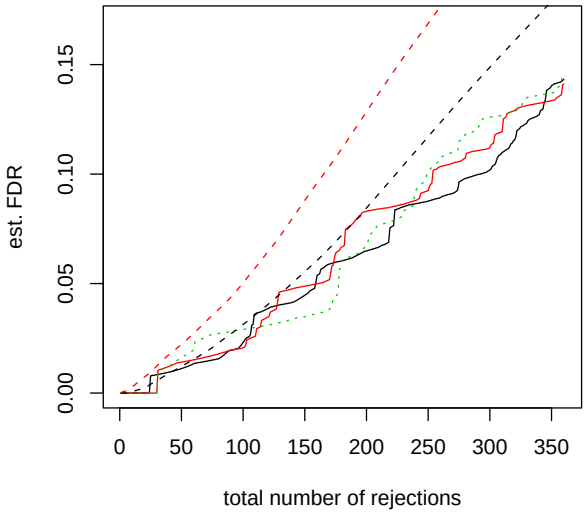

Figure 4: The estimated FDR versus the total number of rejections in the breast cancer data. The black solid curve is for the PBFDR under Prior 1, the black dashed curve is for the BFDR under Prior 1, the red solid curve is for the PBFDR under Prior 2, the red dashed curve is for the BFDR under Prior 2, and the green dotted curve is for the permutation-based FDR of the SAM.

framework. The key idea is to construct an empirical null distribution for the posterior probabilities (z_i) using a computationally efficient algorithm. Based on this empirical null distribution, 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 estimated for other testing procedures. It allows researchers to objectively compare the performance of full Bayesian methods with other frequentist and EB methods.

The proposed algorithm only requires fitting the full Bayesian model once, which reduces the computational burden tremendously. We acknowledge that 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, Series B: Methodological*, 57: 289–300. 413

Cao, J., Xie, X. ., Zhang, S., Whitehurst, A., and White, M. A. (2009). "Bayesian opti-

mal discovery procedure for simultaneous significance testing." *BMC Bioinformatics*, 10. 415, 419, 422

Cui, X. (2005). "Improved Statistical Tests for Differential Gene Expression by Shrinking Variance Components Estimates." *Biostatistics (Oxford)*, 6(1): 59–75. 413

Do, K.-A., Müller, P., and Tang, F. (2005). "A Bayesian Mixture Model for Differential Gene Expression." *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 54(3): 627–644. 413, 417

Efron, B. (2008). "Microarrays, empirical bayes and the two-groups model." *Statistical Science*, 23(1): 1–22. 414

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(456): 1151–1160. 413

Ferguson, T. (1996). *A Course in Large Sample Theory*. Chapman & Hall. 426

Genovese, C. and Wasserman, L. (2002). "Operating Characteristics and Extensions of the False Discovery Rate Procedure." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(3): 499–517. 413

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O. ., Borg, A., and Trent, J. (2001). "Gene-expression profiles in hereditary breast cancer." *New England Journal of Medicine*, 344(8): 539–548. 421, 422

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(22): 3899–3914. 414

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). "Bayesian Modeling of Differential Gene Expression." *Biometrics*, 62(1): 10–18. 413, 415

Lönnstedt, I. and Britton, T. (2005). "Hierarchical Bayes Models for CDNA Microarray Gene Expression." *Biostatistics (Oxford)*, 6(2): 279–291. 415

Lönnstedt, I. and Speed, T. (2002). "Replicated microarray data." *Statistica Sinica*, 12(1): 31–46. 413

Müller, P., Parmigiani, G., and Rice, K. (2007). "FDR and Bayesian multiple comparisons rules." In *Bayesian Statistics 8*. Oxford University Press. 415

Müller, 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(468): 990–1001. 415

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5(2): 155–176. 413, 414, 416

Rubin, D. B. (1981). "The Bayesian Bootstrap." *The Annals of Statistics*, 9: 130–134. 414

Smyth, G. K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology*, 3(1): 1–27. 414

Storey, J. and Tibshirani, R. (2003). "Statistical significance for genome-wide studies." In *Proceedings of the National Academy of Sciences*, volume 100, 9440–9445. 414, 418, 419, 422

Storey, J. D. (2002). "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64(3): 479–498. 413

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(2): 414–432. 414, 418, 421

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of the National Academy of Sciences of the United States of America*, 98(9): 5116–5121. 413, 414, 422

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

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

Effectively, we have $w(\theta_0, \theta_1, \eta, \xi, p_r) = P(r_k^0 = 1 \mid \theta_0, \theta_1, \eta, \xi, p_r, X_k^0, Y_k^0)$, which is the predictive probability of a gene with measurements (X_k^0, Y_k^0) being DE, based on parameters $(\theta_0, \theta_1, \eta, \xi, p_r)$.

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

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

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

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

We can consider $g_{ik}(\theta_0, \theta_1, \eta, \xi, p_r)$ 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 (Ferguson 1996) 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 at a rate of $1/\sqrt{n}$ 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 almost surely 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 $h_i(X_k^0, Y_k^0) - s(X_k^0, Y_k^0) \to 0$ almost surely for any i as $n \to +\infty$. Note that how fast $s(X_k^0, Y_k^0)$ and $h_i(X_k^0, Y_k^0)$ converge depends on the specific function of $w(\theta_0, \theta_1, \eta, \xi, p_r)$.

Acknowledgments

This study is supported in part by NIH grants UL1 RR024982 and P50 CA70907. The authors thank the two reviewers and associate editor for their constructive comments and suggestions.