A ROBUST SEQUENTIAL BAYESIAN METHOD FOR IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

Faming Liang[1], Chuanhai Liu[2] and Naisyin Wang[1]

[1]*Texas A&M University* and [2]*Purdue University*

Abstract: A DNA microarray experiment simultaneously measures the expression levels of thousands of genes. An important question is to identify genes that express differentially between two types of tissues or at different experimental conditions. Since large numbers of genes are compared simultaneously, simple use of significance tests can easily lead to false positive findings. We propose a sequential procedure for estimating the empirical null distribution of multiple hypothesis testing and apply the procedure to identify differentially expressed genes in microarray experiments. Our procedure can be viewed as a new method to estimate the q-value proposed by Storey (2002). The key intuition is to obtain an estimate of the null distribution that is robust to the observations from the alternative distribution. Technically, we borrow strength from the missing data literature so that we can avoid estimating the density function corresponding to differentially expressed genes nonparametrically, but can focus on estimating the null density. Numerical comparisons between our method and Storey's original method were conducted in simulated and real data examples. The numerical results show that our procedure outperforms the originally estimated q-values in almost all scenarios.

Key words and phrases: False discovery rate, Markov chain Monte Carlo, microarray data analysis, missing data, multiple hypothesis testing.

1. Introduction

A DNA microarray experiment measures the expression levels of thousands of genes simultaneously. An important question is to identify genes that express differentially between two types of tissues or at different experimental conditions. Since large numbers of genes are compared simultaneously, the use of significance testing methods, such as a Student's t-test or Wilcoxon test, can easily lead to false positive findings if the extremes of multiple samples are not properly accounted for. See Ge, Dudoit and Speed (2003) for more discussion on this issue. Two recent methods that effectively account for the extremes in multiple testing are the false discovery rate (FDR) method and the empirical Bayes method.

To set notation, let $H_1, \ldots, H_N$ denote the collection of N null hypotheses, and $P_1, \ldots, P_N$ denote the corresponding p-values of the N tests. The outcome of testing N genes simultaneously can be summarized as follows.

	Accept H_i	Reject H_i	Total
Genes for which H_i is true:	U	V	n
Genes for which H_i is false:	T	S	n'
Total	W	R	N

The FDR method is due to Benjamini and Hochberg (1995), where

$$\mathrm{FDR} = E\Big(\frac{V}{R}|R > 0\Big)Pr(R > 0),$$

i.e., the false discovery rate is the expected proportion of false positive findings among all the rejected hypotheses. Under the null hypothesis that the p-values resulted from testing the non-differentially expressed genes are independent and uniformly distributed on $[0, 1]$, Benjamini and Hochberg (1995) and Benjamini and Liu (1999) proposed sequential p-value procedures that can control FDR to a desired level. Under the same null hypothesis, Storey (2002, 2003) and Storey, Taylot and Siegmund (2004) proposed a new class of testing procedures which incorporate the information of n/N in the test, and thus have a higher power. Storey (2002) also defined two new quantities, the positive FDR and the q-value. The positive FDR is

$$p\mathrm{FDR} = E\Big(\frac{V}{R}|R > 0\Big)$$

and has certain conceptual advantages over FDR (Storey (2002)). In addition, Storey (2002) showed that, for $\Lambda = [0, \lambda]$, for a test that rejects when the p-value $\leq \lambda$, pFDR can be written

$$p\mathrm{FDR}(\Lambda) = \frac{\pi_0 P_{f_0}(\Lambda)}{P_f(\Lambda)}$$

if the N p-values are mutually independent and follow the mixture distribution

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0)f_1(p). \tag{1}$$

Here π_0 denotes the a *priori* probability that a null hypothesis is true and is typically near 1, say $\pi_0 \geq 0.9$; f_0 and f_1 denote the distributions of the p-values corresponding to the null and alternative hypotheses, respectively; $P_{f_0}(\Lambda)$ and $P_f(\Lambda)$ are the probabilities of Λ with respect to the densities f_0 and f, respectively. Note that in the FDR method, it is generally assumed that f_0 is uniform[0, 1]. For an observed p-value p, the q-value is

$$q(p) = \inf_{\{\Lambda:p\in\Lambda\}} \{p\mathrm{FDR}(\Lambda)\}. \tag{2}$$

Storey (2002) argued that the q-value is a natural pFDR analogue of the p-value used in the conventional single hypothesis test, and suggested that the q-value could be used as a reference quantity for decision of multiple tests. More discussion on the FDR method can be found in Benjamini and Yekutieli (2001), Finner and Roters (2002), Genovese and Wasserman (2002, 2003), Reiner, Yekutieli and Benjamini (2003) and Black (2004).

The empirical Bayes method was proposed by Efron, Tibshirani, Storey and Tusher (2001) and Efron (2004). Unlike the FDR method, the empirical Bayes method works on the test statistics Y_i's, or the test scores $z_i = \Phi^{-1}(P_i)$ or $z_i = \Phi^{-1}(1 - P_i)$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution. Efron (2004) assumed that the test scores follow a mixture distribution

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \tag{3}$$

where f_0 is assumed to be a non-standard normal distribution that can be estimated from the data. In this sense, the method is empirical, and the estimate of f_0 is called the empirical null distribution. Efron (2004) estimated f_0 and f using a spline method. This idea was further extended by Do, Müller and Tang (2005). They assumed no parametric structure on f_1 nor on f, and estimated the densities using a nonparametric Bayesian approach. We note that there are other papers which also focused on the mixture (1); a non-exhaustive list includes Allison, Gadbury, Heo, Fernandez, Lee, Prolla and Weindruch (2002), Pounds and Morris (2003) and Liao, Lin, Selvanayagam and Shih (2004). They assumed that f_0 is uniform[0, 1] and formulated f_1 as a beta or mixture of beta distributions. In the empirical methods, the differentially expressed genes are usually identified using the local FDR (Efron et al. (2001), Efron (2004) and Liao et al. (2004))

$$fdr(z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)},$$

or the posterior expected FDR (Genovese and Wasserman (2002, 2003)). Although these methods work well for many problems, they often need to model or estimate both densities f_0 and f_1. Estimation of f_1 or f can be difficult for certain values of z, particularly when the number of differentially expressed genes is small. The problem can be worse when the distribution of differentially expressed genes has a complex structure. In most of the unsuccessful cases, the estimate of FDR deteriorates simply because one can not accurately estimate f_1 or f.

To address this issue, we propose a sequential Bayesian procedure for identifying differentially expressed genes. Our approach is motivated by methods from the missing data literature and takes advantage of the fact that f_0 usually

has a known parametric structure, at least approximately, as indicated in the literature. Let F, F_0 and F_1 denote the CDFs of f, f_0 and f_1, respectively. Note that it is much easier to estimate a CDF than to estimate a density function nonparametrically. This is evidenced by the faster optimal convergence rate of the former. We thus focus on building an estimated pFDR on a parametrically estimated $\hat{F}_0$ and a nonparametrically estimated $\hat{F}$. This new procedure naturally avoids the difficulties embedded in estimating f_1 or f nonparametrically, with f_0 to be estimated simultaneously. As a result, our estimation of the null distribution is robust to the observations from the alternative distribution f_1.

The rest of the article is organized as follows. In Section 2, the sequential Bayesian procedure is described. In Section 3, the procedure is demonstrated and compared with Storey's procedure through three simulated examples. In Section 4, the new procedure is applied to a data set. In Section 5, we conclude with a brief discussion.

2. The sequential Bayesian Method

In Section 2.1, we give a detailed description for the sequential Bayesian procedure, where the generalized normal distribution is employed to model the null test scores. In this article, the null test scores refer to the test scores for which the null hypotheses are true. In Section 2.2, we justify the validity of the generalized normal distribution for modeling the null test scores through a simulation.

2.1. The sequential Bayesian procedure

We consider the test scores, $z_1, \ldots, z_N$, where $z_i = \Phi^{-1}(1 - P_i)$ with P_i being the p-value corresponding to the ith test. For simplicity, we assume that the z_i's are mutually independent. We assume the following.

(1) The z_i's come from two different populations. The majority of the z_i's are from the null distribution $F_0(z|\theta)$, parameterized by a vector θ; the other z_i's are from an arbitrary unknown distribution $F_1(z)$.

(2) There exists an $m < N$ such that the smallest m z_i's are from the null distribution $F_0(z|\theta)$ with mode(F_0); the mode of the null distribution $F_0(z|\theta)$ is inside the range of these m smallest z_i's. In other words, there exists a number $c > \text{mode}(F_0)$ such that $F_1(c)$ is practically 0 and $1 - F_0(c)$ remains positive, so any test score $z < c$ comes from the null distribution F_0.

Our goal is to find the sample size n and the n z_i's that correspond to non-differentially expressed genes. We first explain the intuition behind our method, the description of the detailed algorithm follows. We consider the following procedure.

Let $\{z_1^*, \ldots, z_n^*\}$ denote the samples in the set $\{z_1, \ldots, z_N\}$ which are from $F_0(z|\theta)$; $n \leq N$. We assume that $\{z_1^*, \ldots, z_m^*\}$ is a copy of the set of the m smallest z_i's, and treat $\{z_{m+1}^*, \ldots, z_n^*\}$ as missing. Furthermore, we assume that $\{z_1^*, z_2^*, \ldots, z_m^*\}$ satisfies the condition $\max_{1 \leq i \leq m} z_i^* \leq c$ and $\min_{m+1 \leq i \leq n} z_i^* > c$. Note that c is here treated as a working parameter, fixed at each cycle of our Bayesian procedure, and is effectively increased from cycle to cycle to improve efficiency by including more data that are potentially from f_0. The joint posterior distribution of n and θ can be written as

$$f(\theta, n|c, z_1, \ldots z_N) \propto \binom{n}{m} \prod_{i=1}^{m} f_0(z_i^*|\theta)[1 - F_0(c|\theta)]^{n-m} P(n)P(\theta), \qquad (4)$$

where the binomial coefficient is the number of all possible arrangements of the n samples with m samples less than or equal to c and others greater than c; $f_0(z|\theta)$ is the probability density function of the null distribution; and $P(n)$ and $P(\theta)$ denote the prior distributions of n and θ, respectively.

To accommodate the possible deviation of the null distribution from the normal distribution, we assume that $f_0(z|\theta)$ belongs to the family of generalized normal distributions (Box and Tiao (1973)),

$$f_0(z|\theta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \exp\left\{ -\left(\frac{|z - \mu|}{\alpha}\right)^\beta \right\}, \qquad (5)$$

where $\theta = (\mu, \alpha, \beta)$ with $\alpha > 0$ and $\beta > 0$. The location parameter μ represents the center of the distribution, the scale parameter α represents the standard deviation, and the shape parameter β represents the rate of exponential decay. For $\beta = 2$, the distribution is $N(\mu, \alpha^2)$; for $0 < \beta < 2$, the distribution has longer tails than the normal distribution; and for $\beta > 2$, the distribution has shorter tails than the normal distribution.

For a Bayes analysis, we specify the following prior distributions for μ, α, β and n. We assume

$$f(\mu) \propto 1, \quad f(\alpha) \propto \frac{1}{\alpha} \quad (\mu < c), \quad f(\beta) \propto \beta^{\frac{\nu}{2}-1} e^{-\frac{\beta}{2}}.$$

For our examples, we set $\nu = 2$ so β has a prior mean value of 2. Based on the symmetry property of the generalized normal distribution, we set

$$P(n) \propto \exp\{-\lambda|n - n_0|\}, \quad n = m, m+1, \ldots, N,$$

where λ is a hyperparameter, and $n_0 = n_0^*$ if $n_0^* < 0.95N$, $n_0 = 0.95N$ otherwise; here, n_0^* is defined to be twice $\#\{z_i : z_i \leq \text{mode}(F_0), i = 1, \ldots, N\}$. Note that $\text{mode}(F_0)$ can be easily identified as the biggest mode of F, the distribution of

all test scores. We use a simple histogram-based method to estimate mode(F_0), though better methods can be used. Since our method is a sequential approach, the current estimate of μ can be used as an estimate of mode(F_0) for the next cycle. The hyperparameter λ represents our belief on how much n is different from n_0. For our examples the $\lambda = 0.001$, which corresponds to a vague prior on n.

With above specifications, we have the following posterior distribution of n and θ,

$$
f(\theta, n | c, z_1, \ldots, Z_N)
$$
$$
\propto \binom{n}{m} \left(\frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \right)^m \exp\left\{ -\sum_{i=1}^{m} \left(\frac{|z_i^* - \mu|}{\alpha} \right)^\beta \right\}
$$
$$
\times \left[\frac{1}{2} - \frac{1}{2\Gamma(\frac{1}{\beta})} \int_0^{(\frac{c-\mu}{\alpha})^\beta} t^{\frac{1}{\beta}-1} e^{-t} dt \right]^{n-m} e^{-\lambda|n-n_0|} \frac{\beta^{\frac{\nu}{2}-1}}{\alpha} e^{-\frac{\beta}{2}}, \qquad (6)
$$

where $\mu < c$, $\alpha > 0$, $\beta > 0$, and $n \in \{m, m+1, ..., N\}$.

Given the observations $z_1, \ldots, z_N$, we can simulate the sample size n and the parameter vector θ from (4) using the Metropolis-within-Gibbs sampler (Robert and Casella (1999)) as follows.

(a) Simulate n from the conditional posterior $f(n \mid \theta, c, z_1, \ldots, z_N)$ using the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and Hastings (1970)).

(b) Simulate θ from the conditional posterior $f(\theta \mid n, c, z_1, \ldots, z_N)$ using the Metropolis-Hastings algorithm.

With the estimate of θ obtained in the above simulation (how to estimate θ is described below), we can test whether $z_{(m+1)}, \ldots, z_{(m+s)} \sim f_0(z|\hat{\theta})$, where $z_{(l)}$ denotes the lth order statistic of z_i's, and $z_{(m+1)}, \ldots, z_{(m+s)}$ are all the samples which are in the interval $(c, c+\delta\hat{\alpha}]$. Here δ is the step size of the increment of c. If δ is too large, the estimate of c, hence the estimate of n, can not be very accurate, -they will be too small or too large relative to their true values. Otherwise, the procedure will need to iterate too many steps for m to reach n. Our experience suggests that a number between 0.01 and 0.1 is often a good choice for δ. If we decide to accept the hypothesis that $z_{(m+1)} \ldots z_{(m+s)} \sim f_0(z|\hat{\theta})$, with the testing procedure described below, then we let the new $m \leftarrow m + s$, is repeat the above procedures, and re-estimate θ with more observations. The above procedure is iterated until no more samples can be added to f_0. The detailed algorithm is summarized at the end of this section.

To test if $\{z_{(m+1)} \ldots z_{(m+s)}\} \sim f_0(z|\hat{\theta})$, we design the following test which is referred to as a null-score-addition test hereafter. We note that testing $\{z_{(m+1)} \ldots$

$z_{(m+s)}\} \sim f_0(z|\hat{\theta})$ is approximately equivalent to testing that $s \sim \text{Binomial}(\hat{n} - m, \hat{\eta})$, with $\hat{\eta} = [F_0(c + \delta\hat{\alpha}|\hat{\theta}) - F_0(c|\hat{\theta})]/[1 - F_0(c|\hat{\theta})]$. Considering the additivity of the components of the mixture model (3), we conduct the following single-sided test. For the null hypothesized value $\eta_0 = \hat{\eta}$ the hypothesis is stated as

$$H_0 : \eta \leq \eta_0, \qquad H_1 : \eta > \eta_0, \tag{7}$$

which rejects the hypothesis $\{z_{(m+1)} \ldots z_{(m+s)}\} \sim f_0(z|\hat{\theta})$ when s is too large. The p-value of the test can be calculated as

$$p = \sum_{i=s}^{\hat{n}-m} \binom{\hat{n} - m}{i} \hat{\eta}^i (1 - \hat{\eta})^{\hat{n}-m-i}. \tag{8}$$

When $\hat{n} - m$ is large and $\hat{\eta}$ is small, p can be calculated by Poisson approximation.

Once a stable value of m has been established, re-simulate $(\hat{n}_1, \hat{\theta}_1), \ldots, (\hat{n}_{M'}, \hat{\theta}_{M'})$ from $f(n, \theta|z_1^*, \ldots, z_m^*)$ by the Metropolis-within-Gibbs sampler. For a particular rejection region Λ specified for the scores under study, we estimate the pFDR by

$$\widehat{p\text{FDR}}(\Lambda) = \frac{1}{M'} \sum_{i=1}^{M'} \frac{\hat{n}_i}{N} \frac{P_{f_0(\cdot|\hat{\theta})}(\Lambda|\hat{\theta}_i)}{\widehat{P}_f(\Lambda)}, \tag{9}$$

where $\widehat{P}_f(\Lambda)$ is estimated by $\#\{z_i : z_i \in \Lambda, i = 1, \ldots, N\}/N$, as in Storey (2002). If the rejection region takes the form $[\omega, \infty)$, we denote $\widehat{p\text{FDR}}([\omega, \infty))$ by $\widehat{p\text{FDR}}(\omega)$ and re-write (9) as

$$\widehat{p\text{FDR}}(\omega) = \frac{1}{M'} \sum_{i=1}^{M'} \frac{\hat{n}_i}{N} \frac{1 - F_0(\omega|\hat{\theta}_i)}{1 - \hat{F}(\omega)}, \tag{10}$$

where $\hat{F}(\omega) = \#\{z_i : z_i \leq \omega\}/N$. Furthermore, for the nested rejection regions of the form $[\omega, \infty)$, we can estimate the q-value of the observed score z by

$$\hat{q}(z) = \inf_{\omega \leq z} \{\widehat{p\text{FDR}}(\omega)\}. \tag{11}$$

The genes in the set $\{z_i : \hat{q}(z_i) < \zeta\}$ are then identified as the differentially expressed genes at the significance level ζ. We now summarize the algorithm.

The Sequential Bayesian Estimation Algorithm

(a) Set c to the Qth percentile of z, and determine the value of m such that $z_{(m)} < c$ and $z_{(m+1)} > c$. Let I denote the number of consecutive sampling stages used for estimation of the parameters n and θ. Let $I = 1$, $\hat{n}_0 = m$, $\hat{\beta}_0 = 2$, and $\hat{\mu}_0$ and $\hat{\alpha}_0$ be the mean and standard deviation of $z_{(1)}, \ldots, z_{(m)}$, respectively.

(b) Simulate samples $(n_1, \theta_1), \ldots, (n_M, \theta_M)$ from the joint posterior $f(n, \theta|c, z_1, \ldots, z_N)$ for M steps starting with the current estimate $(\hat{n}_{I-1}, \hat{\theta}_{I-1})$. Estimate n by

$$\hat{n}_I = (1 - \frac{1}{I})\hat{n}_{I-1} + \frac{1}{(M - M_0)I} \sum_{i=M_0+1}^{M} n_i,$$

where M_0 is the number of burn-in steps. Estimate μ, α and β similarly.

(c) Calculate the p-value at (8). If $p \leq 0.1$ and $I < S$, set $I \leftarrow I + 1$ and go to step(b). Otherwise, go to step (d).

(d) Test the hypothesis $H_0 : z_{(m+1)} \ldots z_{(m+s)} \sim f_0(z|\hat{\theta}_I)$ versus $H_1 : z_{(m+1)} \ldots z_{(m+s)} \nsim f_0(z|\hat{\theta}_I)$ at a pre-specified level γ. If H_0 is accepted, set $m \leftarrow m+s$, $c \leftarrow c + \delta\hat{\alpha}$, $\hat{n}_0 = \hat{n}_I$, $\hat{\theta}_0 = \hat{\theta}_I$ and $I = 1$, go to step (b). Otherwise, go to step (e).

(e) Fix c to its current value, re-simulate samples $(\hat{n}_1, \hat{\theta}_1), \ldots, (\hat{n}_{M'}, \hat{\theta}_{M'})$ from $f(n, \theta|c, z_1, \ldots, z_N)$ by the Metropolis-within-Gibbs sampler, calculate $\hat{q}(z_i)$ for $i = 1, \ldots, N$, and identify the differentially expressed genes according to the $\hat{q}(z_i)$'s.

In the sequential procedure, the parameters that need to be specified by the user include Q, M_0, M, δ, γ, S and M'. The default setting used in our examples is $Q = 80$, $M_0 = 200$, $M = 1,000$, $\delta = 0.025$, $\gamma = 0.05$, $S = 3$, and $M' = 1,000$. Otherwise, the setting used will be specified in context.

The default setting of Q is 80. This is quite reasonable as π_0 often takes a value greater than 0.9. The problem of identifying differentially expressed genes is truly a clustering problem. The idea underlying the sequential procedure is similar to that of semi-supervised learning (Basu, Banerjee and Mooney (2002)), Q can be as large as possible if we are sure that the genes included in the learning dataset (i.e., $\{z_{(1)}, \ldots, z_{(m)}\}$) express non-differentially. In our experience, we obtain similar outcomes provided we choose Q from a reasonable range, say, $Q > 50$. This "robust" property actually made finding an optimal Q a practically difficult and unrewarding procedure in the sense that the results are not very differential with respect to Q, and the "optimal" Q provides very limited improvement. The values of M_0, M and M' are problem dependent. The more complex the posterior distribution is, the larger the M_0, M and M' used. The setting $\gamma = 0.05$ is appropriate for most examples, although this may be too strict. The parameter δ controls the step size of null test score addition. If δ is too large, the resulting estimate of c, and thus the estimate of n may be far from its true value. Otherwise, the procedure will need to iterate too many steps to have m reach n. Our experience suggests that a number between 0.01 and 0.1 is often a good choice for δ. Fortunately, the outcome of the sequential procedure

will not be affected much by the setting of γ and δ as shown by one of our examples. The parameter S together with the parameter M controls the accuracy of the estimate of c. The larger the product MS is, the smaller variance the estimates of c have. Introducing the parameter S improves the efficiency of the sequential procedure, as it can make a moderate M acceptable.

2.2. Numerical evaluation of the generalized normal modeling

Let $z_1, \ldots, z_n$ denote the null test scores. The null test scores have been modeled using different distributions by different authors. Storey (2002) modeled the scores using the standard normal distribution $N(0, 1)$ by assuming that the corresponding p-values follow the uniform distribution Unif(0, 1). Efron (2004) relaxed Storey's assumption and modeled the scores using a non-standard normal distribution $N(\mu, \sigma^2)$, where μ and σ are estimated from the data using a spline method. In the sequential procedure, we relaxed Efron's assumption and modeled the null test scores using a generalized normal distribution. In the following simulation, we illustrate the flexibility of the use of the generalized normal distribution.

Suppose that the expression levels of n genes are measured under two experimental conditions on ten microarray chips. Let $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,10})$ be the expression levels of gene i, where $x_{i,1}, \ldots, x_{i,5}$ were measured under condition 1, and $x_{i,6}, \ldots, x_{i,10}$ were measured under condition 2. We assume that the distribution of the expression levels can deviate from normal and that there could exist dependence among genes. Furthermore, we assume that not all gene expression measurements are usable and we keep all the genes which have at least two usable expression levels per condition.

First, we allowed error distribution to be non-normal. Let x_{ij} be the gene expression level described above, so

$$\frac{x_{ij} - \mu_i^{(1)}}{\sigma_i} \sim t(v), \quad j = 1, \ldots, 5; \quad \frac{x_{ij} - \mu_i^{(2)}}{\sigma_i} \sim t(v), \quad j = 6, \ldots, 10; \quad (12)$$

where $t(v)$ denotes the student t-distribution with degree of freedom v, $\mu_i^{(1)}$ and $\mu_i^{(2)}$ are the respective mean expression levels of gene i under condition 1 and condition 2, and σ_i^2 is a random variable distributed according to the inverse gamma distribution $IG(2.5, 0.5)$. The parameters of the inverse gamma distribution are computed from a gene expression dataset, the avian pineal gland gene expression data analyzed in Section 4. In analyzing the dataset, we also found that the distribution of the gene expression levels was significantly different from the normal distribution and closer to a student t-distribution with the degrees of freedom ranging from 3 to 5. Hence, we tried $v = 3, 4$ and 5 in (12) and (13).

Next, we created some level of dependence among genes. To correlate gene i with gene l in expression, conditional on the expression levels of gene l, we let

$$\frac{x_{ij} - \mu_{ij|l}^{(1)}}{\sigma_{i|l}} \sim t(v), \quad j = 1, \dots, 5; \quad \frac{x_{ij} - \mu_{ij|l}^{(2)}}{\sigma_{i|l}} \sim t(v), \quad j = 6, \dots, 10, \quad (13)$$

where $\mu_{ij|l}^{(a)} = \mu_i^{(a)} + \rho_{il}\sigma_i/\sigma_l(x_{lj} - \mu_l^{(a)})$ $(a = 1, 2)$, $\sigma_{i|l} = \sigma_i\sqrt{1 - \rho_{il}^2}$, and ρ_{il} is a random variable drawn from the uniform distribution Unif$[-1, 1]$. Note that if $t(v)$ is replaced by $N(0, 1)$ in (13), then x_{ij} and x_{lj} are normal random variables and their correlation coefficient is equal to ρ_{il} exactly. Here, with x_{ij} and x_{lj} as student-t random variables, our simulation results show a correlation coefficient slightly larger than ρ_{il}.

At last, for each gene we randomly discard some expression levels as disqualified and thus missing values, but per condition we retain at least two expression levels as usable values.

Based on the principles described above, we have the following algorithm for simulating gene expression x_s, $s = 1, \dots, n$.

(a) Set $v = 3$, 4, or 5 and $s = 1$; generate $\boldsymbol{x}_1$ according to (12).

(b) Repeat steps (b.1)-(b.3) below for $s = 2$ to n.

    (b.1) Randomly choose an l uniformly from the set $\{0, 1, \dots, s - 1\}$.

    (b.2) If $l = 0$, generate $\boldsymbol{x}_s$ according to (12).

    (b.3) If $l \neq 0$, generate $\boldsymbol{x}_s$ according to (13), conditional on the expression levels of gene l.

(c) Repeat steps (c.1)-(c.2) below for $s = 1$ to n.

    (c.1) Randomly choose k_{s1} and k_{s2} uniformly from the set $\{2, 3, 4, 5\}$.

    (c.2) Retain $x_{s,1}, \dots, x_{s,k_{s1}}$ and $x_{s,6}, \dots, x_{s,5+k_{s2}}$ as usable expression levels of gene s under condition 1 and condition 2, respectively.

Figure 1 (1)-(4) show the quantile-quantile (Q-Q) normal plots of the expression levels of 5,200 genes simulated with $v = 4$, $\mu_i^{(1)} = 0$ for all i, $\mu_i^{(2)} = 0$ for $i = 1, \dots, 5,000$ and $\mu_i^{(2)}$ drawn from $N(5, 1)$ for $i = 5,001, \dots, 5,200$. The data shown in Figures 1 (1)-(4) correspond to $x_{.1}$, $x_{.2}$, $x_{.6}$ and $x_{.7}$, respectively. Hence, Figures 1, 3 and 4 mimic datasets that include differentially expressed genes. For comparison, we also show the Q-Q normal plots of the expression levels of the avian genes (Figure 1, 5-12). Refer to Section 4 for more information on the dataset. Visually, the simulated datasets seem to share similar features with gene expression data when the data are not normally distributed.
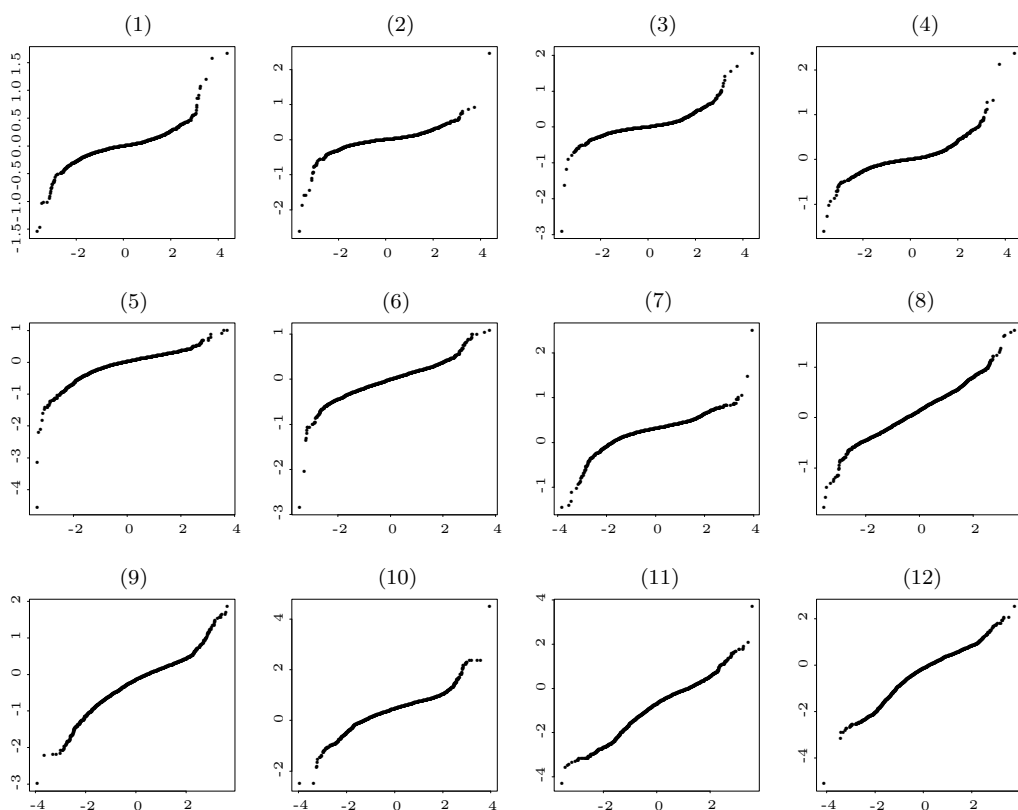
Figure 1. Comparison of the distributions of the simulated and gene expression levels (scaled). Plots 1-4: Q-Q normal plots of the simulated gene expression data. Plots 5-8: Q-Q normal plots of the avian LD data measured in four chips at a time point. Plots 9-12: Q-Q normal plots of the avian DD data measured in four chips at a time point. The horizontal axis is the quantile of the standard normal distribution, and the vertical axis is the quantile of the gene expression levels.

With the simulated data, we test the hypotheses $H_{i0} : \mu_i^{(1)} = \mu_i^{(2)}$ versus $H_{i1} : \mu_i^{(1)} \neq \mu_i^{(2)}$ using the two sample t-test statistic

$$Y_i = \frac{\bar{x}_i^{(1)} - \bar{x}_i^{(2)}}{\sqrt{S_i^2 \left(\frac{1}{k_{i1}} + \frac{1}{k_{i2}} \right)}},$$

where $\bar{x}_i^{(1)}$ and $\bar{x}_i^{(2)}$ denote the averages of the observed expression levels of gene i under conditions 1 and 2, respectively; and S_i^2 is the pooled variance estimate

for σ_i^2. We assume that $Y_i \sim t(k_{i1} + k_{i2} - 2)$ by treating x_{ij}'s as samples drawn from a normal distribution, calculate the p-value, and then convert the p-value to the score $z_i = \Phi^{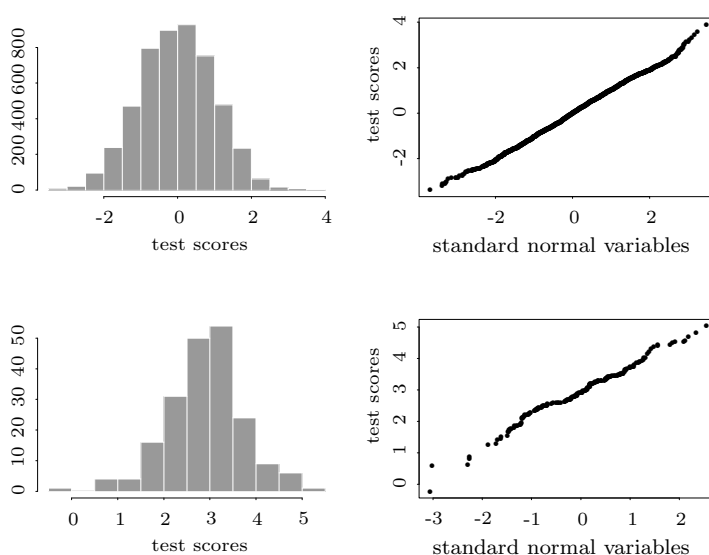-1}(1 - P_i)$. Figure 2 summarizes the characteristics of the test scores for a simulated data set. Figure 2 (a) and (b) are the histogram and Q-Q normal plot of the test scores for the non-differentially expressed genes, while Figure 2c and 2d are for the differentially expressed genes. Figure 2b indicates that a normal distribution may not be the best distribution for modeling null test scores of the gene expression data. This is further evidenced by the avian DD data shown in Figure 8a.

Figure 2. Test scores of a simulated data set. Plots a and b are the histogram and Q-Q normal plot of the test scores of the non-differentially expressed genes. Plots c and d are the histogram and Q-Q normal plot of the test scores of the differentially expressed genes.

For a simulation study to compare the performance of the normal and generalized normal distributions, the above procedure was replicated 30 times and 30 independent datasets were created. The degree of freedom of the t-distribution was set to $v = 3, 4$ and 5 for the first, second and third 10 datasets, respectively. The null test scores were then fitted by both the normal and the generalized normal distributions. The fitness was measured using the BIC statistic

$$BIC_d = -\text{log-likelihood} + \frac{p_d}{2}\log(n), \qquad (14)$$

where $d \in \{1, 2\}$ denotes the choice of the fitting distribution (1 for the normal distribution and 2 for the generalized normal distribution), and p_d is the number

of parameters of the fitting distribution. According to the BIC criterion, the fitting distribution which results in a smaller BIC value is preferred.

Table 1. Comparison of the normal and generalized normal distributions for modeling the simulated null test scores: v is the degrees of freedom of the t-distribution used in simulating the data; $\#\{BIC_1 \leq BIC_2\}$ is the number of datasets favoring the normal distribution; $\#\{BIC_2 < BIC_1\}$ is the number of datasets favoring the generalized normal distribution; $BIC_1 - BIC_2$ is the average (over 10 datasets) of the difference of the BIC values for the two distributions; the number in parentheses is the standard deviation of the corresponding average value.

v	$\#\{BIC_1 \leq BIC_2\}$	$\#\{BIC_2 < BIC_1\}$	$BIC_1 - BIC_2$
3	0	10	5.61 (1.41)
4	1	9	2.51 (0.77)
5	5	5	-0.03 (1.09)

Table 1 summarizes the BIC values calculated for the 30 datasets. It indicates that the generalized normal distribution is a good choice for modeling simulated null test scores when v is less than 5, and the normal distribution is preferred by BIC otherwise. Nonetheless, even when normal assumption is preferable, the generalized normal distribution still allows normality as a special case but at the cost of a higher number of parameters. From this result and Figure 1, we can see that the generalized normal distribution is not a bad choice for avian gene expression data. In fact, it is the necessary choice for the DD data shown in Figures 8 and 9.

3. Simulated Examples

In this section, the sequential Bayesian procedure is compared with Storey's procedure (2002) through three simulated examples, one of which satisfies our assumptions while the others do not.

3.1. Example 1

This example comprises 20 datasets. Each dataset consists of 2,100 test scores, of which the first 2,000 scores are generated from the standard normal distribution, and the remaining 100 scores are generated from a left-truncated student-t distribution with $df = 5$ and the truncation threshold $T = 3$. In this section, each score is viewed as a different gene, and the terms score and gene are used exchangeably. In terms of mixture models, the density function of the data can be written as

$$f(z) = \pi_0 \phi(z) + (1 - \pi_0)\widetilde{t}(z|df = 5, T = 3), \tag{15}$$

where $\phi(z)$ denotes the density function of the standard normal, and $\widetilde{t}(z|df = 5, T = 3)$ denotes the density function of the left-truncated student-t distribution. Here $\phi(\cdot)$ and $\widetilde{t}(\cdot)$ correspond to the f_0 and f_1 in (3), respectively. The histogram of one of the 20 datasets is shown in Figure 3a. The long right tail corresponds to the left-truncated student-t distribution. Since $f_1(z)$ is left-truncated, this example satisfies our assumptions that the majority of the scores are from $f_0(z)$ and that there exists a number c such that all z_i's less than c are from f_0.
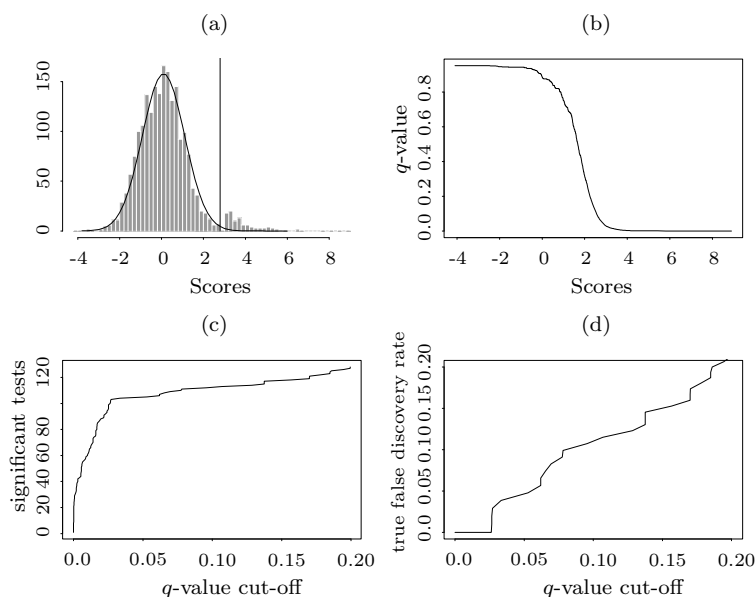
Figure 3. a: Histogram of the test scores. The vertical line shows the cut-off point, the value of c, obtained in a run of the sequential procedure at the final stage. b: the q-values versus the test scores; c: the numbers of significant genes versus the q-value cut-off values; d: the true false discovery rates (tFDR) versus the q-values.

We first applied the sequential procedure to the dataset shown in Figure 3a. The computational results are summarized in Figures 3 and 4. The vertical line in Figure 3a shows the cut-off point, the value of c, obtained in a run of the sequential procedure at the final stage. It splits the dataset into two parts, the part used (left) and the part not used (right) for estimation of f_0 and π_0. Five runs of the sequential procedure produced $\hat{\pi}_0 = 0.9519$ with standard deviation $2e - 5$, which is almost identical to the true value 0.9524; and $\hat{\theta} = (\hat{\mu}, \hat{\alpha}, \hat{\beta}) = (-0.0124, 1.4298, 2.0169)$ with standard deviation (0.0007, 0.0016, 0.0037), also very close to the true value (0.0, 1.414, 2.0). These results show the validity of the sequential procedure for estimation of the empirical null distribution. Figure

3b shows the q-values, which decrease monotonically with test scores. Figure 3c shows the number of significant scores versus different q-value cut-off values. Figure 3d compares the q-value and the true false discovery rate (tFDR), where the tFDR is defined for a given rejection region as the ratio V/R if $R > 0$ and 0 otherwise. The tFDR is calculable only when V and R are both available. The approximate equality of the q-value and tFDR in Figure 3d shows that the q-value defined in (11) is a valid measure for the false discovery rate. Note that the results shown in Figure 3 are all from a single run of the sequential procedure. They are almost identical to those obtained by averaging over multiple runs. This can be seen from the small standard deviations of the estimates of π_0 and θ.
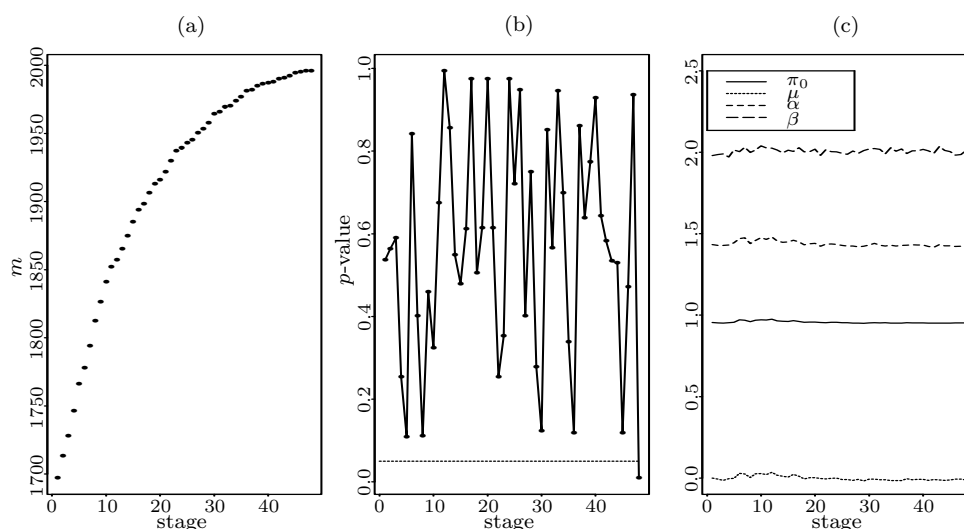
Figure 4. Intermediate results produced in a run of the sequential procedure. a: the increasing process of m; b: the p-values of the null-score-addition tests; c: the estimates of π_0 and θ in different stages.

Figure 4 shows some of the intermediate results produced during a run of the sequential procedure. Figure 4a shows the increasing progress of the value of m, the number of scores used for estimation of f_0. It starts with 1,680, the 80th percentile of the dataset, and then increases sequentially until the null-score-addition test is rejected. The number of observations added in each stage tends to decrease as the procedure evolves. Figure 4b shows the p-values of the null-score-addition tests. Figure 4c shows the estimates of π_0 and θ at different stages. It is easy to see that these estimates are quite stable over the whole run of the sequential procedure, although they tend to have large variations in the early stages. This implies that the value of m is not crucial to the sequential procedure. At the beginning of the sequential procedure, we have almost enough data for estimation of f_0, the sequential procedure makes the estimate more accurate.

The sequential procedure was then run once for each of the 20 datasets. The results of the 20 runs are summarized in Table 2, which shows the tFDRs for different rejection regions. For example, the tFDR of the rejection region $\Lambda(0.25) = \{z : \hat{q}(z) \leq 0.25\}$ is 0.252 with standard deviation 0.017, where the tFDR and its standard deviation are calculated by averaging over the 20 runs. Here we suppose that the rejection regions are determined according to the q-values. The consistency of the tFDR and the claimed q-value for each of the rejection regions considered in Table 2 suggests that the q-value defined in (11) provides a good approximation to the tFDR, and can work as a reasonable test statistic for multiple tests.

Table 2. True false discovery rates for Example 1. The true false discovery rates and their standard deviations (the numbers in parentheses) are calculated by averaging over 20 runs.

Methods	$\hat{\pi}_0$	True false discovery rate					
		$\Lambda(0.3)$	$\Lambda(0.25)$	$\Lambda(0.2)$	$\Lambda(0.15)$	$\Lambda(0.1)$	$\Lambda(0.05)$
Sequential	0.945	0.303	0.252	0.198	0.146	0.096	0.049
	(0.004)	(0.018)	(0.017)	(0.016)	(0.016)	(0.012)	(0.006)
Storey	0.943	0.296	0.250	0.201	0.136	0.090	0.046
	(0.010)	(0.007)	(0.008)	(0.008)	(0.009)	(0.009)	(0.005)

For comparison, we also applied Storey's FDR procedure (Storey (2002)) to this example. The software was downloaded from http://faculty.washington .edu/~jstorey/. It was run at the default setting for all examples. The p-values used in Storey's procedure are transformed from the Z-scores as $P = 1 - \Phi(z)$. In Storey's procedure, f_0 is assumed to be uniform[0, 1]. This is equivalent to using the true null distribution $N(0, 1)$ in (10) and (11) while we consider a more general family in (5). Hence, we expect that Storey's procedure will outperform the sequential procedure for this example with $f_0(z)$ being the standard normal. It is encouraging to note that two procedures perform similarly, and the sequential procedure performs even better than Storey's procedure for small q rejection regions. The true FDRs produced by the sequential procedures are closer to the theoretical q-values than those produced by Storey's procedure for these rejection regions. In practice, the p-values may not be exactly uniform[0, 1] and our procedure should have more flexibility.

Table 3 displays the agreement/disagreement between the two procedures on the identification of differentially expressed genes. The table can be read as follows. For example, if the rejection region is chosen as $\Lambda(0.10)$, on average (over the 20 runs) there are 109.1 genes identified by both procedures as differentially

expressed genes. The sequential procedure identifies 111 (=109.1+1.9) genes, and Storey's procedure identifies 110.05 (=109.1+0.95) genes. The respective FDRs of the two procedures are displayed in Table 2. For completeness, we also give in the 'Disagree/Disagree' column the average numbers of differentially expressed genes that are not identified by either of two procedures. This number should not be related with the comparison of the two procedures. Table 3 indicates that these two procedures are almost identical for this ideal example on identification for differentially expressed genes. A minor difference is that the sequential procedure tends to identify slightly more genes as differentially expressed genes for this example, and this makes the resulting true FDRs closer to nominal levels.

Table 3. Agreement/disagreement between the sequential and Storey's procedures on the identification of differentially expressed genes for Example 1. Agree/Agree: the average number of differentially expressed genes identified by both procedures. Agree/Disagree: the average number of differentially expressed genes identified by the sequential procedure but not by Storey's procedure. Agree/Disagree: the average number of differentially expressed genes identified by the sequential procedure but not by Storey's procedure. Disagree/Disagree: the average number of truly differentially expressed genes not identified by either of two procedures. The averages are calculated based on the 20 runs, and the numbers in parentheses are the standard deviations of the averages.

Sequential/Storey	Agree/Agree	Agree/Disagree	Disagree/Agree	Disagree/Disagree
$\Lambda(0.30)$	137.50(1.84)	8.15 (3.14)	4.80(1.35)	0.0 (0.0)
$\Lambda(0.25)$	129.75(2.03)	5.40 (1.87)	3.85(1.32)	0.0 (0.0)
$\Lambda(0.20)$	121.60(1.68)	4.10 (1.46)	3.80(1.29)	0.0 (0.0)
$\Lambda(0.15)$	113.85(1.38)	4.05 (1.38)	2.05(0.60)	0.0 (0.0)
$\Lambda(0.10)$	109.10(1.14)	1.90 (1.02)	0.95(0.34)	0.0 (0.0)
$\Lambda(0.05)$	104.20(0.65)	0.90 (0.33)	0.70(0.25)	0.0 (0.0)

3.2. Example 2

This example also comprises 20 datasets. Each dataset consists of 2,100 test scores, of which the first 2,000 are generated from $N(0, 1.5^2)$, and the remaining 100 are generated from $N(4, 1)$. Similar to (15), we write the mixture model as

$$f(z) = \pi_0\phi(\frac{z}{1.5}) + (1 - \pi_0)\phi(z - 4), \qquad (16)$$

where $\phi(z/1.5)$ and $\phi(z - 4)$ correspond to $f_0(z)$ and $f_1(z)$, respectively. Figure 5a shows the histogram of one dataset. The shape of the histogram shows the difficulty and practicality of the dataset: the right tail is slightly long, and there

is no a clear split between f_0 and f_1. In addition, our second assumption may be violated: the data used for estimation of f_0 may not be all from f_0.
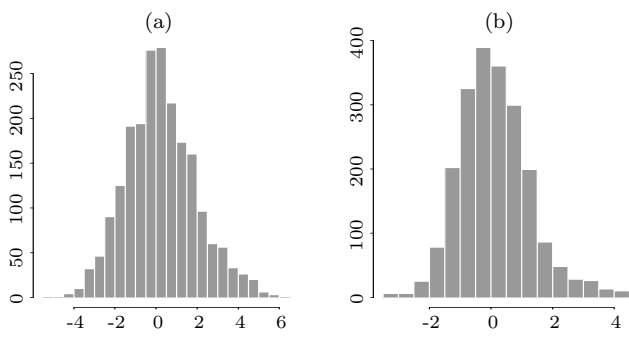

Figure 5. a: histogram of test scores from Example 2; b: histogram of test scores from Example 3.

Table 4. True false discovery rates for Example 2. The numbers in parentheses are standard deviations of the true FDRs. The upper panel shows the sensitivity of the sequential procedure to the choice of λ. The middle panel shows the sensitivity of the sequential procedure to the choice of ν.

| Methods | (λ, ν) | $\hat{\pi}_0$ | True false discovery rate | | | | | |
			$\Lambda(0.3)$	$\Lambda(0.25)$	$\Lambda(0.2)$	$\Lambda(0.15)$	$\Lambda(0.1)$	$\Lambda(0.05)$
Sequential	(0.0005,2)	0.952 (0.002)	0.306 (0.019)	0.252 (0.020)	0.200 (0.019)	0.150 (0.016)	0.087 (0.018)	0.046 (0.013)
	(0.001,2)	0.953 (0.003)	0.306 (0.019)	0.250 (0.020)	0.196 (0.019)	0.148 (0.015)	0.082 (0.015)	0.046 (0.013)
	(0.002,2)	0.950 (0.002)	0.318 (0.018)	0.258 (0.020)	0.206 (0.018)	0.154 (0.015)	0.089 (0.018)	0.045 (0.013)
Sequential	(0.001,1)	0.952 (0.003)	0.306 (0.019)	0.249 (0.019)	0.195 (0.019)	0.142 (0.017)	0.070 (0.016)	0.044 (0.012)
	(0.001,2)	0.953 (0.003)	0.306 (0.019)	0.250 (0.020)	0.196 (0.019)	0.148 (0.015)	0.082 (0.015)	0.046 (0.013)
	(0.001,3)	0.952 (0.002)	0.311 (0.017)	0.251 (0.018)	0.202 (0.018)	0.148 (0.016)	0.081 (0.016)	0.054 (0.016)
Storey		1.0 (0.000)	0.750 (0.004)	0.721 (0.005)	0.689 (0.006)	0.638 (0.006)	0.577 (0.007)	0.479 (0.009)

Both the sequential procedure and Storey's procedure were applied to this example. The results are summarized in Table 4 (the row with $\lambda = 0.001$ and $\nu = 2$). For preparing the p-values used in Storey's procedure, the transformation $P = 1 - \Phi(z)$ was applied to the Z-scores as in Example 1. Table 4 shows that this example is much more difficult than Example 1, and our assumptions are

possibly violated, but the sequential procedure still works well. The true FDRs are consistent with their nominal levels, and π_0 is estimated rather accurately. Note that the true value of π_0 is 0.952. However, Storey's procedure completely fails here, the true FDRs are much higher than their nominal levels and π_0 is over-estimated in all runs.

To assess the dependence of the performance of the sequential procedure on the prior distributions, sensitivity analysis was done for the hyperparameters λ and ν that control the strength of the prior information on n and β, respectively. The sequential procedure was re-run for the choices of (λ, β) given in Table 4. The numerical results show that the sequential procedure is not sensitive to the choice of the hyperparameters. Note that the change of ν from 2 to 1 or 3 represents a significant change in our belief about the tail behavior of f_0. The choices $\nu = 1, 2, 3$ drive the tail of the fitted distribution $\widehat{f_0}$ toward exponential (with rate 1), normal and Weibull (with the shape parameter 3), respectively. The change of λ from 0.001 to 0.0005 or 0.002 also represents a big change in our thinking about n_0, the initial guess at n.

Table 5. Agreement/disagreement between the sequential and Storey's procedures on the identification of differentially expressed genes for Example 2. The notation is that of Table 3.

Sequential/Storey	Agree/Agree	Agree/Disagree	Disagree/Agree	Disagree/Disagree
$\Lambda(0.30)$	112.80(6.11)	0.0(0.0)	285.35(10.19)	0.85 (0.22)
$\Lambda(0.25)$	96.10(5.80)	0.0(0.0)	259.20(9.46)	1.35 (0.33)
$\Lambda(0.20)$	78.15(5.75)	0.0(0.0)	236.05(8.60)	2.15 (0.38)
$\Lambda(0.15)$	57.55(5.42)	0.0(0.0)	211.00(7.67)	3.15 (0.44)
$\Lambda(0.10)$	33.80(4.60)	0.0(0.0)	192.50(6.40)	4.85 (0.58)
$\Lambda(0.05)$	13.20(3.06)	0.0(0.0)	163.00(4.80)	8.75 (0.92)

Table 5 displays the agreement/disagreement between the two procedures on the identification of differentially expressed genes. It shows that Storey's procedure tends to identify too many genes as differentially expressed. The significant genes identified by the sequential procedure is a subset of those identified by Storey's procedure. Note that a plot like Figure 3c would be helpful in determining an appropriate cut-off value of q.

3.3. Example 3

This example also comprises 20 datasets. Each dataset consists of 2,100 test scores as described in Section 2.2. The first 2,000 genes were generated as non-differentially expressed genes, and the last 100 genes were generated as differentially expressed genes.

Both the sequential procedure and Storey's procedure were applied. The computational results are summarized in Table 6 and Table 7. They show that even though this example violates our assumptions on test score independence and normality of gene expression levels, the sequential procedure still works well. The tFDR is still close to the nominal level, as shown in Table 6, and the estimate of π_0 is rather accurate. However, Storey's procedure fails here since the tFDR is much lower than its nominal level. Table 7 just confirms the results presented in Table 6. The sequential procedure tends to identify more differentially expressed genes than Storey's procedure and this brings the resulting true FDRs close to their nominal levels.

Table 6. True false discovery rates for Example 3. The notation is that of Table 2.

| Methods | $\hat{\pi}_0$ | True false discovery rate | | | | | |
		$\Lambda(0.3)$	$\Lambda(0.25)$	$\Lambda(0.2)$	$\Lambda(0.15)$	$\Lambda(0.1)$	$\Lambda(0.05)$
Sequential	0.956	0.296	0.242	0.193	0.142	0.097	0.055
	(0.003)	(0.017)	(0.016)	(0.016)	(0.013)	(0.012)	(0.009)
Storey	0.956	0.245	0.201	0.151	0.103	0.067	0.027
	(0.008)	(0.014)	(0.011)	(0.010)	(0.009)	(0.007)	(0.005)

Table 7. Agreement/disagreement between the sequential and Storey's procedures on the identification of differentially expressed genes for Example 3. The notations are as in Table 3.

Sequential/Storey	Agree/Agree	Agree/Disagree	Disagree/Agree	Disagree/Disagree
$\Lambda(0.30)$	108.65(2.37)	13.25(3.74)	1.30(1.30)	15.15 (0.77)
$\Lambda(0.25)$	98.15(1.81)	10.75(2.51)	0.95(0.77)	18.05 (0.91)
$\Lambda(0.20)$	86.55(1.49)	11.35(2.45)	1.10(1.10)	21.50 (0.87)
$\Lambda(0.15)$	74.80(1.61)	10.70(2.01)	0.85(0.85)	26.50 (1.11)
$\Lambda(0.10)$	63.65(1.62)	9.60(1.94)	0.75(0.65)	33.70 (1.47)
$\Lambda(0.05)$	45.80(1.22)	11.90(2.19)	0.55(0.50)	45.20 (1.95)

4. Avian Pineal Gland Gene Expressions Data

The avian pineal gland contains both circadian oscillators and photoreceptors to produce rhythms in biosynthesis of the hormone melatonin in vivo and in vitro. It is of great interest to understand the genetic mechanisms driving the rhythms. For this purpose, a sequence of cDNA microarrays of birds' pineal gland transcripts under light-dark (LD) and constant darkness (DD) conditions were

generated. Under LD, birds were euthanized at 2, 6, 10, 14, 18 and 22 hour Zeit-geber time (ZT) to obtain mRNA for adequate cDNA libraries. Four microarray chips per time point were produced, and there are two replicates for each gene in each chip. The experiment was then repeated under DD. Throughout, samples from LD ZT18 were used as controls. Relative gene expression levels to the controls were recorded and processed. The initial goal is to identify genes that are differentially expressed at different time points. Mixed effect analysis, with the fixed effect being the different time points and the random effects corresponding to chips and biological batches, was applied to the relative gene expression levels in log-scale. Normalization procedures were adopted but will not be listed here since they are not the focus of the paper. Under both LD and DD conditions, the p-values P_i for testing the existence of different time effects were produced and transformed to test scores using $\Phi^{-1}(1 - P_i)$.
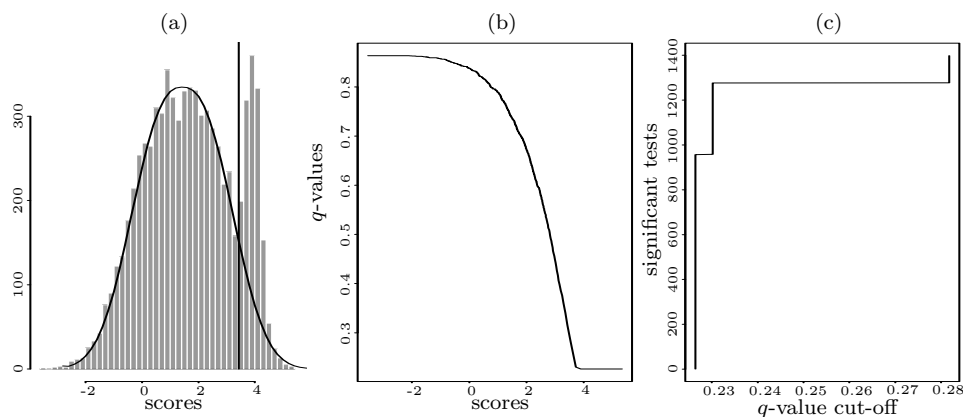

Figure 6. Computational results of the sequential procedure for the LD data. a: the histogram of the scores and the fitted f_0 in one run of the sequential procedure; b: the q-values versus the test scores; c: the numbers of significant genes versus the q-value cut-off values.

4.1. LD data

We first applied the sequential procedure to the scores of the LD data. The computational results are summarized in Figure 6. Figure 6a shows the histogram of the scores and the estimated density curve f_0 by a run of the sequential procedure. From this plot, we can see that f_0 can be well-estimated, the estimate is actually quite stable. We repeated the procedure five times. And found the estimate $(\hat{\pi}_0, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.863, 1.341, 2.207, 2.521)$ with standard deviation (0.0003, 0.0018, 0.0041, 0.0092). The vertical line shows the split point between the scores that are used and those not used for estimation of f_0 in the run. The interaction

of the fitted f_0 density curve and the histogram bars suggests that there are as many as 1,400 genes which might be differentially expressed. The test scores of these genes are all greater than 3.5. Figure 6b shows the q-values versus the test scores. Figure 6c shows the numbers of significant genes versus the q-value cut-off values. Our analysis suggests that among the identified 1,400 differentially expressed genes, there are about 400 genes ($\approx 1,400 \times 28\%$) which are false positive and about 1,000 genes which are really differentially expressed.

For comparison, we also applied Storey's procedure to the p-values constructed above for the LD data. The computational results are summarized in Figure 7. Figure 7a shows the q-values, and Figure 7b shows the numbers of significant genes versus the q-value cut-off values. Figure 7 implies that even we include as many as 4,000 genes in the set of differently expressed genes, the false discovery rate is as low as 5%. In addition, Storey's procedure produces the estimate $\hat{\pi}_0 = 0.235$. These results are quite different from what we have obtained. Figure 4a seems to suggest that $\hat{\pi}_0 = 0.235$ severely underestimates the true π_0. Note that $\hat{\beta} = 2.54$ in our analysis, and this implies that the original null p-values may not be uniformly distributed, as assumed by Storey. We believe our analysis provides a more reasonable outcome than does Storey's.
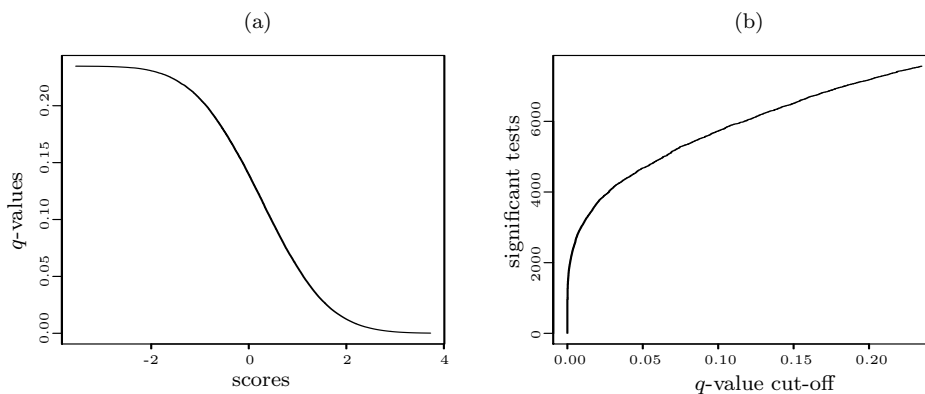

Figure 7. Computational results of Storey's procedure for the LD data. a: the q-values versus the test scores; b: the numbers of Significant genes versus the q-value cut-off values.

4.2. DD data

We also applied the sequential procedure to the test scores of the DD data. The computational results are summarized in Figure 8. This plot tells us again that f_0 can be well-estimated by the sequential procedure. We repeated the procedure five times as before, and obtained the estimate $(\hat{\pi}_0, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.970, 0.357,$

$0.939, 1.285$) with standard deviation ($0.0005, 0.0011, 0.0033, 0.0043$). The vertical line shows the split in one run between the scores which are used and those which are not used for estimation of f_0.
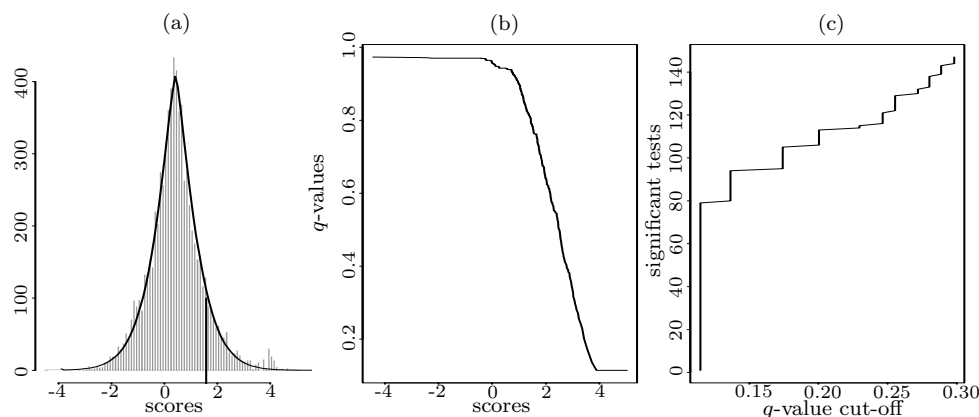

Figure 8. Computational results of the sequential procedure for the DD data. a: the histogram of the scores and the fitted f_0 in one run of the sequential procedure; b: the q-values versus the test scores; c: the numbers of Significant genes versus the q-value cut-off values.

As pointed out before, the number of scores, m, used in the estimation of f_0, or the size γ of the null-score-addition test designed to increase m, is not crucial for the proposed sequential procedure. To illustrate this, we reduce the significant level of the null-score-addition test from the default of $\gamma = 0.05$ to $\gamma = 0.001$, so a larger choice of final m could result. We then purposely chose two different split points from two different runs. One is at the 93.68^{th} percentile, and the other one at the 98.47^{th} percentile, shown by the vertical lines in Figure 9a and Figure 9d, respectively. Taking a closer look at the fitted density curves in the three runs, we can see that the curve shown in 9d is slightly shifted to the right compared to the other two. However, the effect on the final outcome is minor. The plots in Figure 8a, Figure 9a and Figure 9d suggest that there are about 100 suspiciously differentially expressed genes with test scores greater than 3.6, and the false discovery rate of the identified significant genes is about 20%.

For comparison, we also applied Storey's procedure to the p-values of the DD data. The computational results are summarized in Figure 10. As with the LD data, Storey's procedure again produces unreasonable outcomes. The estimate of π_0 by Storey's procedure is 0.396, which seems low. Compared to our findings, we also suspect that Storey's procedure severely underestimates the false discovery rate.
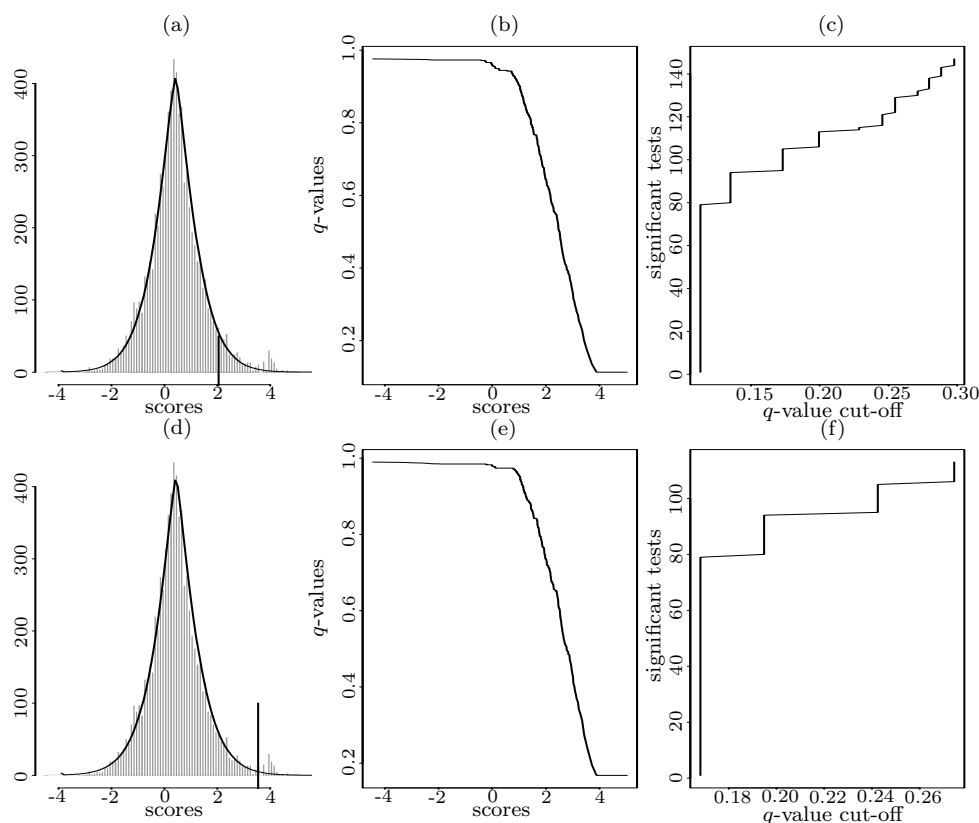
Figure 9. Computational results of the sequential procedure for the DD data with the significant level $\gamma = 0.001$ of the data addition tests. a and d: the histogram of the scores and the fitted f_0; b and e: q-values versus the test scores; c and f: the numbers of significant genes versus q-value cut-off values.
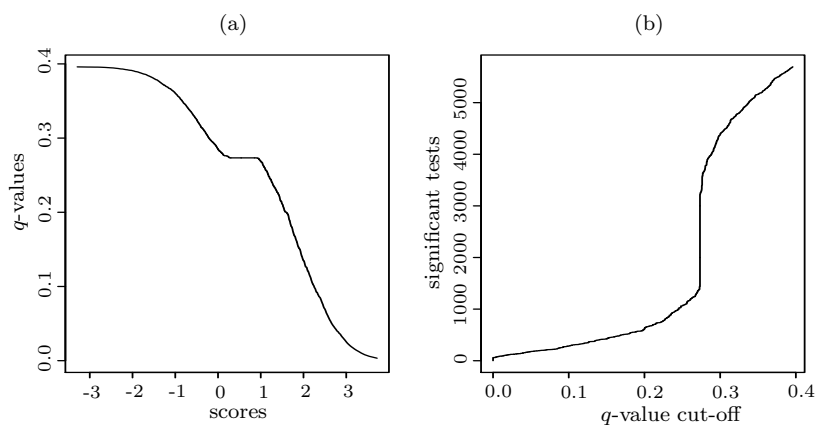

Figure 10. Computational results of Storey's procedure for the DD data. a: q-values versus test scores; b: the numbers of Significant genes versus the q-value cut-off values.

5. Discussion

Although our sequential procedure works well for many problems, it can be modified, extended or improved further in several aspects as listed below.

The null distribution f_0 can be modeled by different parametric distributions, a mixture of the generalized normal distributions for example, to reflect the biological background of the "null" condition for the dataset under study. The null score addition process can be slightly improved by adding a backward deletion step even though the final outcomes seem to be fairly indifferent toward the choice of m.

The sequential procedure is developed based on the assumption that the test scores are independent. In reality, this is not true. Test scores are often correlated due to the correlations among functionally related genes. An extension of the sequential procedure to correlated test scores would be of interest, although the current procedure works well for a simulated (Example 3) and two real datasets where the genes are correlated in expression. We note that the SAM procedure presented in Tusher, Tibshirani and Chu (2001) retains the correlation structure in gene expression. However, SAM and our procedure estimate different targets. As pointed out by Dudoit, Shaffer and Boldrick (2003), the definition of FDR in SAM is different from the standard one given in Benjamini and Hochberg (1995): the SAM FDR is estimating $E(V|H_0^c)/R$ and not E(V/R) as in Benjamini and Hochberg (1995), where H_0^c denotes the complete null hypotheses. Consequently the SAM FDR can be greater than 1 (e.g., Table 3 in Chu et al. (2000, p.16), Other FDR procedures which have accounted for the dependence structure of gene expression include Benjamini and Yekutieli (2001), Storey and Tibshirani (2001, 2003), Storey, Taylor and Siegmund (2004), among others. A major problem with these procedures is an unrealistic assumption: the p-values are uniformly distributed under the null hypotheses. As discussed in Efron (2005), the violation of the uniformity assumption of p-values for the FDR procedures may be more harmful to FDR estimation than the violation of the independence assumption of test statistics for the empirical Bayes procedures.

As in many other papers on FDR estimation (e.g., Efron (2004)), we start our analysis with z-scores (or, equivalently, p-values) by treating them as observations. Due to the randomness and limited sample size of the microarray experiments, these analyses may not be able to identify all genes which express differentially, but they should be able to identify the genes which express at a level significantly different from the normal one. This can be seen in our Example 3, where the majority of the differentially expressed genes are identified by the sequential procedure. Of course, further research on how to take into account the observation errors in FDR estimation is of great interest. However, with the

small sample sizes commonly seen in microarray experiments, it is hard to say whether the practice of taking into account this further uncertainty could be a practical one.

Acknowledgements

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. *Comput. Statist. Data Anal.*, **39**, 1-20.

Basu, S., Banerjee, A. and Mooney, R. J. (2002). Semi-supervised Clustering by Seeding. *Proceedings of the Nineteenth International Conference on Machine Learning*, 19-26, Sydney, Australia, July 2002.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypothesis procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82**, 163-170.

Benjamini, Y. and Yekutieli, D. (2001). On the control of false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165-1188.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *J. Roy. Statist. Soc. Ser. B* **66**, 297-304.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Wiley, New York.

Chu, G., Goss, V., Narasimhan, B. and Tibshirani, R. (2000). SAM (Significance Analysis of Microarrays)–Users guide and technical document. Technical Report, Stanford University.

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

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71-103.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96-104.

Efron, B. (2005). Correlation and large-scale simultaneous significance testing. *Technical Report.* Department of Statistics, Stanford University, Stanford.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Finner, H. and Roters, M. (2002). Multiple hypothesis testing and expected number of type I errors. *Ann. Statist.* **30**, 220-238.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *Test* **12**, 1-77.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extension of the FDR procedure. *J. Roy. Statist. Soc. Ser. B* **64**, 499-517.

Genovese, C. and Wasserman, L. (2003). Bayesian and frequentist multiple testing. In *Bayesian Statistics 7*, Oxford University Press, Oxford.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Liao, J. G., Lin, Y., Selvanayagam, Z. E. and Shih, W.J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics* **20**, 2694-2701.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1091.

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 the p-values. *Bioinformatics*, **19**, 1236-1242.

Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer, New York.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64**, 479-498.

Storey, J. D. (2003). The positive false discovery rate: an Bayesian interpretation and the Q-value. *Ann. Statist.* **31**, 2013-2035.

Storey, J. D., Taylot, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. Ser. B* **66**, 187-205.

Storey, J. D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001-18. Department of Statistics, Stanford University, Stanford.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440-9445.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116-5121.

Department of Statistics, Texas A&M University, College Station, TX 77843-3143.

E-mail: fliang@stat.tamu.edu.

Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067.

E-mail: chuanhai@stat.purdue.edu.

Department of Statistics, Texas A&M University, College Station, TX 77843-3143.

E-mail: nwang@stat.tamu.edu.