A Sequential Bayesian Approach with Applications to Circadian Rhythm Microarray Gene Expression Data

Faming Liang,

*University of Florida*
In microarray data analysis, an important question is to identify genes that express differentially between two types of tissues or at different experimental conditions. Since a large number of genes is compared simultaneously, the use of significance testing methods, such as a Student’s t-test or Wilcoxon test, could lead to a large chance of false positive findings if the extremes of multiple samples are not properly accounted for.

Existing methods in the literature

(1) False discovery rate (FDR) method (Benjamini and Hochberg, 1995)

**FDR methods**

**Notations:** Let $H_1, \ldots, H_N$ denote the collection of $N$ null hypotheses, $Y_1, \ldots, Y_N$ denote the corresponding $N$ test statistics, and $P_1, \ldots, P_N$ denote the corresponding $N$ $p$-values of the tests.

<table>
<thead>
<tr>
<th></th>
<th>Accept $H_i$</th>
<th>Reject $H_i$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes for which $H_i$ is true:</td>
<td>$U$</td>
<td>$V$</td>
<td>$n$</td>
</tr>
<tr>
<td>Genes for which $H_i$ is false:</td>
<td>$T$</td>
<td>$S$</td>
<td>$n'$</td>
</tr>
<tr>
<td>Total</td>
<td>$W$</td>
<td>$R$</td>
<td>$N$</td>
</tr>
</tbody>
</table>
**Introduction**

**Literature review for FDR methods**

**Assumption:** The $N$ $p$-values are mutually independent and follow the mixture distribution

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p),$$

where $\pi_0$ denotes a *priori* probability that a null hypothesis is true and it 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. In the FDR method, it is generally assumed that $f_0$ is uniform $[0,1]$.

- Benjamini and Hochberg (1995) and Benjamini and Liu (1999) defines FDR as the expected proportion of false positive findings among all the rejected hypotheses, i.e.,

$$\text{FDR} = E\left(\frac{V}{R} | R > 0\right) Pr(R > 0).$$

Sequential $p$-value based testing procedures are proposed to control FDR to a desired level.
Storey (2002, 2003) and Storey, Taylor and Siegmund (2004) proposed a new class of testing procedures by incorporating the information of $\pi_0$. The tests have usually a higher power. Related quantities:

(i) Positive FDR:

$$pFDR = E\left(\frac{V}{R} \mid R > 0\right),$$

which is simpler than FDR conceptually.
(ii) If the rejection region is of the form $\Lambda = [0, \lambda]$, then $p$FDR is

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

where $P_{f_0}$ and $P_f$ are the probabilities of $\Lambda$ w.r.t. $f_0$ and $f$, respectively.

(iii) $q$-value:

$$q(p) = \inf_{\{\Lambda; p \in \Lambda\}} \{p\text{FDR}(\Lambda)\}.$$

Storey (2002) argued that the $q$-value is a natural $p$FDR 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.
**Literature review for empirical Bayes methods**

Unlike FDR methods, empirical Bayes methods work on the test statistics $Y_i$'s or the test scores $Z_i = \Phi^{-1}(P_i)$ or $Z_i = \Phi^{-1}(1 - P_i)$, and assume the score follow a mixture distribution,

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z),$$

where $\Phi$ is the CDF of the standard Gaussian, $f_0$ is a non-standard Gaussian and can be estimated from the data.

- Do, Muller and Tang (2004) estimated $f_0$ and $f_1$ using a nonparametric Bayesian approach.
Literature review for empirical Bayes methods

In empirical Bayes methods, the differentially expressed genes are usually identified using the local FDR

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

or the posterior expected FDR (Genovese and Wasserman, 2002, 2003).
Multiple Hypothesis Testing

Introduction

Difficulties of the two existing methods

- The distribution of the $p$-values of the non-differentially expressed genes may have a fair deviation from the uniform[0,1]. This renders the invalidity of the FDR methods.

- When the number of differentially expressed genes are small, the estimate of $f_1$ or $f$ may not be reliable.

Our approach estimates $f_0$ (and thus $F_0$) parametrically with a sequential procedure by taking advantage of missing data techniques, and estimate $F$ nonparametrically.

Note that estimating $F$ is much easier than estimating $f$ or $f_1$, as evidenced by the faster optimal convergence rate of the former.
Sequential Bayesian Procedure

Methodology

Assumptions:

- It works on the test scores $Z_i = \Phi^{-1}(1 - P_i)$.
- It assumes that the scores are mutually independent, the majority of the $Z_i$'s are from the null distribution $F_0$, and the others are from $F_1$. Here $F_1$ may be an arbitrarily complex distribution.
- There exists a number $C > \text{mode}(F_0)$ and any $Z < C$ follows the null distribution $F_0$. 
Let \( \{z_1, \ldots, z_N\} \) be a sample from the mixture distribution

\[
F(z) = \pi_0 F(z|\theta) + (1 - \pi_0) F_1(z),
\]

and suppose that \( \{z_1^*, \ldots, z_n^*\} \) are from \( F_0 \). Our goal is to determine the sample size \( n \) (\( n \leq N \)) and the parameters of \( F_0 \).

We assume that \( \{z_1^*, z_2^*, \ldots, z_m^*\} \) is an identical copy of the set of the \( m \) smallest \( z_i \)'s, and there exist a cut-off value \( c \) such that \( \max_{1 \leq i \leq m} z_i^* \leq c \) and \( \min_{m+1 \leq i \leq n} z_i^* > c \). Treating \( \{z_{m+1}^*, \ldots, z_n^*\} \) as missing, we have the posterior of \( n \) and \( \theta \) as follows,

\[
P(\theta, n|z_1^*, \ldots, z_m^*) \propto \binom{n}{m} \prod_{i=1}^{m} f_0(z_i^*|\theta)[1 - F_0(c|\theta)]^{n-m} P(n) P(\theta).
\]
Bayes Specification

- Specify $f_0$ as a generalized Gaussian distribution (Box and Tiao, 1973),

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

where $\theta = (\mu, \alpha, \beta)$ with $\alpha > 0$ and $\beta > 0$. The location parameter $\mu$ represents the center of the distribution, the scale parameter $\alpha$ represents the standard deviation, and the shape parameter $\beta$ represents the rate of exponential decay.

- The priors for $\theta = (\mu, \alpha, \beta)$:

  $$f(\mu) \propto 1, \quad f(\alpha) \propto \frac{1}{\alpha} \quad (\mu < c),$$

  and $\beta$ follows a $\chi^2(\nu)$ distribution with degree of freedom $\nu$; that is,

  $$f(\beta) \propto \beta^{\nu/2-1}e^{-\beta/2}.$$
Based on the symmetry of $f_0$,

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

where $\lambda$ is a hyperparameter, and $n_0 = n_0^*$ if $n_0^* < 0.95N$, and $0.95N$ otherwise; here, $n_0^*$ is defined to be twice of $\#\{z_i : z_i \leq \text{mode}(F_0), i = 1, \ldots, N\}$. Setting $\lambda = 0.001$, which corresponds to a vague prior on $n$. 
The posterior distribution of \( n \) and \( \theta \) is

\[
P(\theta, n | z_1^*, \ldots, z_m^*) \propto \left( \frac{n}{m} \right) \left( \frac{\beta}{2\alpha \Gamma(1/\beta)} \right)^m \exp\left\{ -\sum_{i=1}^{m} (|z_i^* - \mu|/\alpha)^\beta \right\}
\]

\[
\left[ \frac{1}{2} - \frac{1}{2\Gamma(1/\beta)} \int_0^{(c_\alpha)^\beta} t^{1/\beta-1} e^{-t} dt \right]^{n-m} e^{-\lambda |n-n_0|} \frac{\beta^{\nu/2-1}}{\alpha} e^{-\beta/2},
\]

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

The scheme for sampling from \( P(\theta, n | z_1^*, \ldots, z_m^*) \).

(a) Simulate \( n \) from the conditional posterior \( f(n | \theta, z_1^*, \ldots, z_m^*) \) using the Metropolis-Hastings algorithm.

(b) Simulate \( \theta \) from the conditional posterior \( f(\theta | n, z_1^*, \ldots, z_m^*) \) using the Metropolis-Hastings algorithm.
Sequential Bayesian Procedure

With the estimate of $\theta$ obtained in the above simulation, we can test if $z_{(m+1)}, \ldots, z_{(m+s)} \sim f_0(z|\hat{\theta})$, where $z_{(m+1)}, \ldots, z_{(m+s)} \in (c, c + \delta \hat{\alpha}]$. Here $\delta$ is usually set to a value between 0.01 and 0.1.

- If we decide to accept the hypothesis that $z_{(m+1)} \ldots z_{(m+s)} \sim f_0(z|\hat{\theta})$, then we let the new $m \leftarrow m + s$, repeat the simulation, and re-estimate $\theta$ with the data $\{z(1), \ldots, z_{(m+s)}\}$.

The above procedure will be iterated until no more samples can be added to $f_0$. 
The Sequential Bayesian Estimation Algorithm

(0) Set $c$ to the $Q^{th}$ 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 stages used for estimation of the parameters $n$ and $\theta$. 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.

(1) Simulate samples $(n_1, \theta_1), \ldots, (n_M, \theta_M)$ from the joint posterior $L(n, \theta|z_1^*, \ldots, z_m^*)$ 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 $\hat{n}_0 = 0$, and $M_0$ is the number of burn-in steps in each stage. Estimate $\mu$, $\alpha$, and $\beta$ similarly.
(2) Test the hypothesis $H_0 : z_{(m+1)} \ldots z_{(m+s)} \sim f_0(z|\hat{\theta}_I)$ versus $H_1 : z_{(m+1)} \ldots z_{(m+s)} \sim f_0(z|\hat{\theta}_I)$ at a pre-specified level $\gamma$. If $H_0$ is accepted, set $m \leftarrow m + s$, $c \leftarrow c + \delta \hat{\alpha}$, $n_0 = \hat{n}_I$, $\hat{\theta}_0 = \hat{\theta}_I$, and $I = 1$, go to step (1); otherwise, keep $m$ and $c$ unchanged and set $I \leftarrow I + 1$, go to step(3).

(3) If $I > S$, go to step (4); otherwise, go to step (1).

(4) Fix $m$ to the current value, re-simulate samples $(\hat{n}_1, \hat{\theta}_1), \ldots, (\hat{n}_M, \hat{\theta}_M)$ from $f(n, \theta|z_1^*, \ldots, z_m^*)$ by the Metropolis-within-Gibbs sampler, calculate $\hat{q}(z_i)$ for $i = 1, \ldots, N$, and identify the differentially expressed genes according to the $\hat{q}(z_i)$'s.

Default setting: $Q = 80$, $M_0 = 200$, $M = 1000$, $\delta = 0.025$, $\gamma = 0.05$, $S = 3$, and $M' = 1000$. 
Estimator of FDR in the Sequential Bayesian Procedure

If the rejection region takes the form $\Lambda = [\omega, \infty)$, we have

$$\widehat{p_{FDR}}(\Lambda) = \frac{1}{M'} \sum_{i=1}^{M'} \frac{\hat{n}_i 1 - F_0(\omega|\hat{\theta}_i)}{N \left( 1 - \hat{F}(\omega) \right)}$$

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

$$\hat{q}(z) = \inf_{\omega \leq z} \{\widehat{p_{FDR}}(\omega)\}.$$
Example 1 This example comprises 20 datasets. Each dataset consists of 2100 test scores, of which the first 2000 scores are generated from the standard normal distribution, and the remaining 100 scores are generated from a left-truncated student-\(t\) distribution with the degree of freedom \(df = 5\) and the truncation threshold \(T = 3\); that is,

\[ f(z) = \pi_0\phi(z) + (1 - \pi_0)\tilde{t}(z|df = 5, T = 3). \]

The histogram of one of the 20 datasets is shown in Figure 1(a).

Five runs of the sequential procedure for this dataset produce the following estimate: \(\hat{\pi}_0 = 0.9526\) with standard deviation 0.0009, which is almost identical to the true value 0.9524; and \(\hat{\theta} = (\hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.0158, 1.4179, 1.9914)\) with standard deviation (.0026, .0039, .0070), which is also very close to the true value (0.0, 1.414, 2.0).
Figure 1: (a) Histogram of the 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 scores versus the \( q \)-value cut-off values. (d) The true false discovery rates (tFDR) versus the \( q \)-values.
Figure 2: 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 $\pi_0$ and $\theta$ in different stages.
### Table 1: True false discovery rates for different rejection regions.

The true false discovery rates and their standard deviations, the numbers in the below parentheses, are calculated by averaging over 20 runs. Storey’s procedure: [http://faculty.washington.edu/jstorey/](http://faculty.washington.edu/jstorey/).

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\hat{\pi}_0$</th>
<th>$\hat{q}(z) \leq 0.2$</th>
<th>$\hat{q}(z) \leq 0.15$</th>
<th>$\hat{q}(z) \leq 0.1$</th>
<th>$\hat{q}(z) \leq 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>0.948</td>
<td>(0.002)</td>
<td>0.208</td>
<td>0.161</td>
<td>0.107</td>
</tr>
<tr>
<td>Storey</td>
<td>0.934</td>
<td>(0.012)</td>
<td>0.206</td>
<td>0.156</td>
<td>0.109</td>
</tr>
</tbody>
</table>

**Example 2** It comprises 20 datasets. Each of the dataset consists of 2100 test score, of which the first 2000 scores are generated from $N(0, 1.5^2)$, and the remaining 100 scores are generated from $N(4, 1)$; that is, $f(z) = \pi_0 \phi(z/1.5) + (1 - \pi_0)\phi(z - 4)$.

Figure 3 shows the histogram of one dataset. For this dataset, five runs of the sequential procedure produce the following estimates, $(\hat{\pi}_0, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.949, 0.05, 2.148, 1.98)$ with the standard deviations (0.001, 0.002, 0.006, 0.007). The true value is $(\pi_0, \mu, \alpha, \beta) = (0.952, 0.0, 2.121, 2.0)$. 
Figure 3: Histogram of one dataset generated by the mixture model of Example 2.
## True false discovery rate

$$\hat{q}(z) \leq 0.2 \quad \hat{q}(z) \leq 0.15 \quad \hat{q}(z) \leq 0.1 \quad \hat{q}(z) \leq 0.05$$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\hat{\pi}_0$</th>
<th>$\hat{q}(z) \leq 0.2$</th>
<th>$\hat{q}(z) \leq 0.15$</th>
<th>$\hat{q}(z) \leq 0.1$</th>
<th>$\hat{q}(z) \leq 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>0.952</td>
<td>(0.003)</td>
<td>0.199</td>
<td>(0.020)</td>
<td>0.146</td>
</tr>
<tr>
<td>Storey</td>
<td>1.0</td>
<td>(0.000)</td>
<td>0.687</td>
<td>(0.006)</td>
<td>0.638</td>
</tr>
</tbody>
</table>

Table 2: True false discovery rates for different rejection regions. The true false discovery rates and their standard deviations, the numbers in the below parentheses, are calculated by averaging over 20 runs.
**Data Description:** 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, 22 hour Zeitgeber time (ZT) to obtain mRNA to produce adequate cDNA libraries. Under both LD and DD conditions, $p$-values, $P_i$’s, for testing the existence of different time effects were produced and transformed to test scores using $\Phi(1 - P_i)$.
Figure 4: 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. Estimate: $(\hat{\pi}_0, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.863, 1.339, 2.203, 2.540)$ with the standard deviation $(0.0002, 0.0019, 0.0021, 0.0147)$. Among the identified 1400 differentially expressed genes, there are about 400 genes ($\approx 1400 \times 28\%$) which are false positive.
Figure 5: 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.
Figure 6: 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. Estimate: $(\hat{\pi}_0, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = (0.972, 0.361, 0.942, 1.284)$ with the standard deviation $(0.0004, 0.0020, 0.0021, 0.0146)$. 
Figure 7: Computational results of the sequential procedure for the DD data with the significant level $\gamma = 0.001$ of the data addition tests. (a)&(d): the histogram of the scores and the fitted $f_0$. (b)&(e): the $q$-values versus the test scores. (c)&(f): the numbers of Significant genes versus the $q$-value cut-off values.
Figure 8: Computational results of Storey’s procedure for the DD data. (a) The $q$-values versus the test scores. (c) The numbers of Significant genes versus the $q$-value cut-off values.
• Suspiciously differentially expressed genes can be identified by comparing the density curve of the null distribution and the histogram of the test scores of genes.

• The null distribution can be modeled by different parametric distributions, say, a mixture of the generalized normal distribution, 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 $m$. 