Biometrika (2008), **95**, 4, *pp*. 961–977
doi: 10.1093/biomet/asn036
Advance Access publication 26 September 2008

Estimating the false discovery rate using the stochastic approximation algorithm

BY FAMING LIANG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
fliang@stat.tamu.edu

AND JIAN ZHANG

Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, U.K.
j.zhang@kent.ac.uk

SUMMARY

Testing of multiple hypotheses involves statistics that are strongly dependent in some applications, but most work on this subject is based on the assumption of independence. We propose a new method for estimating the false discovery rate of multiple hypothesis tests, in which the density of test scores is estimated parametrically by minimizing the Kullback–Leibler distance between the unknown density and its estimator using the stochastic approximation algorithm, and the false discovery rate is estimated using the ensemble averaging method. Our method is applicable under general dependence between test statistics. Numerical comparisons between our method and several competitors, conducted on simulated and real data examples, show that our method achieves more accurate control of the false discovery rate in almost all scenarios.

Some key words: Ensemble averaging; False discovery rate; Microarray data analysis; Multiple hypothesis testing; Stochastic approximation.

1. INTRODUCTION

In microarray data analysis and elsewhere, one often needs to test a large number of hypotheses simultaneously, and frequentist methods and empirical Bayes methods have been developed to control the probability of erroneously rejecting true null hypotheses.

Let $H_1, \ldots, H_N$ denote the collection of N null hypotheses, $P_1 \ldots, P_N$ denote the corresponding p-values of the N tests and $Z_i = \Phi^{-1}(1 - P_i)$ denote the corresponding test scores, where Φ is the standard normal cumulative distribution function. For the outcome of the multiple tests, let V and U denote, respectively, the numbers of true null hypotheses that are erroneously rejected and correctly accepted, and let T and S denote, respectively, the numbers of false null hypotheses that are erroneously accepted and correctly rejected.

Among the frequentist methods, the false discovery rate control method (Benjamini & Hochberg, 1995) has received much attention, where the false discovery rate,

$$\text{FDR} \equiv E\left(\frac{V}{R} \mid R > 0\right) \text{pr}(R > 0),$$

in which $R = V + S$ denotes the total number of rejected hypotheses, is the expected proportion of false positive findings among all the rejected hypotheses. Under the assumption that the null p-values, i.e. the p-values for which the null hypotheses are true, are independent and uniformly distributed on [0, 1], Benjamini & Hochberg (1995) and Benjamini & Liu (1999) proposed several testing procedures which can control FDR at a given level. These procedures were further improved by Storey (2002), Storey et al. (2004) and Benjamini et al. (2006) by incorporating information about n/N into the test, where n denotes the total number of true hypotheses. Benjamini & Yekutieli (2001, 2005) and Genovese & Wasserman (2002) provide more discussion of these methods. Dependence between test statistics has been considered by Benjamini & Yekutieli (2001), Storey et al. (2004) and Benjamini et al. (2006). The main difficulty with the frequentist methods comes from the unrealistic assumption that the null p-values are uniformly distributed. When this assumption is violated, the resulting estimator of FDR may be unreliable, being either too liberal or too conservative (Pounds & Cheng 2006).

The empirical Bayes methods (Efron et al., 2001; Efron, 2004, 2007) overcome this difficulty by using an empirical null distribution estimated from the data. The methods assume that the test scores follow a mixture density

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

where π_0 is the prior probability that a null hypothesis is true, f_0 is the empirical null distribution and f_1 is the alternative distribution. In contrast to the empirical null distribution, the standard normal distribution is called the theoretical null distribution in this context. Bordes et al. (2006) show that the model (1) is nonidentifiable in general even if one component is completely specified. To achieve identifiability, some constraints, such as symmetry, have to be imposed on the unknown component. In this paper, $f(z)$ is modelled by a multiple-component parametric mixture distribution which is identifiable, so that f_0 and f_1 are identifiable under some clustering criterion on the components of the mixture distribution. Section 2 provides the details of the estimation of the model.

In the empirical Bayes methods, the differentially expressed genes are usually identified on the basis of the local false discovery rate (Efron et al., 2001), which is defined as

$$\text{FDR}(z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}.$$

The procedure developed by Efron (2004) allows for dependence between test scores. Many articles focus on the mixture model, including Allison et al. (2002), Pan et al. (2003), Do et al. (2005) and Liang et al. (2007).

In this paper, we extend Efron's work to model $f(z)$ by a mixture of exponential power distributions, and estimate the parameters of the mixture distribution by minimizing the Kullback–Leibler distance

$$\text{KL}(f_\theta, f) \equiv -\int \log\left\{\frac{f(z \mid \theta)}{f(z)}\right\} f(z)dz, \tag{2}$$

where $f(z \mid \theta)$ denotes the density of the mixture distribution and θ is the vector of the parameters. Jensen's inequality implies that $\text{KL}(f_\theta, f) \geqslant 0$. We show that the Kullback–Leibler distance can be minimized using the stochastic approximation method (Robbins & Monro, 1951), which allows general dependence between test scores. Pan et al. (2003) and Do et al. (2005) use a mixture of normal distributions, but in both articles, the models are estimated based on the assumption of mutual independence between test scores. Pan et al. (2003) employed the EM algorithm, and Do et al. (2005) used Markov chain Monte Carlo simulation.

2. Estimation of the false discovery rate

2·1. *The false discovery rate*

As discussed by Efron et al. (2001), the methods for controlling the false discovery rate and the empirical Bayes methods are closely related. Let $F(z)$ and $F_0(z)$ be the cumulative distribution functions corresponding to the densities $f(z)$ and $f_0(z)$, respectively, let the test scores be transformed from p-values via the function $Z_i = \Phi^{-1}(1 - P_i)$, and consider a rejection rule $\Lambda = \{Z_i \geqslant z_0\}$. The conditional expectation of $\mathrm{FDR}(z)$ given Λ is then

$$\mathrm{FDR}(\Lambda) = \frac{\int_{z_0}^{\infty} \mathrm{FDR}(z) f(z) dz}{\int_{z_0}^{\infty} f(z) dz} = \frac{\pi_0\{1 - F_0(z_0)\}}{1 - F(z_0)},$$

which corresponds to Benjamini and Hochberg's tail-area false discovery rate and is also called the 'Bayesian false discovery rate' in Efron (2004). When the test scores are independently and identically distributed, $\mathrm{FDR}(\Lambda)$ reduces to the positive false discovery rate introduced by Storey (2002). One natural estimator of $\mathrm{FDR}(\Lambda)$ is

$$\widehat{\mathrm{FDR}}(\Lambda) = \frac{N\hat{\pi}_0\{1 - \hat{F}_0(z_0)\}}{\#\{z_i : z_i \geqslant z_0\}},$$

where $\#\{z_i : z_i \geqslant z_0\}$ denotes the number of tests with scores greater than z_0, and $\hat{\pi}_0$ and $\hat{F}_0$ denote the estimators of π_0 and F_0, respectively. The quantity $\widehat{\mathrm{FDR}}(\Lambda)$ can be interpreted as the expected proportion of null cases among those having $z_i \geqslant z_0$. This estimator has been used in Liang et al. (2007) and Efron (2007). Similarly to Storey (2002), we define the q-value as

$$q(z) \equiv \inf_{\{\Lambda : z \in \Lambda\}} \mathrm{FDR}(\Lambda), \tag{3}$$

which we call the Bayesian q-value and which can be used as a reference quantity for the decision in multiple hypothesis tests. In the remainder of this section, we describe how to estimate $q(z)$ when $f(z)$ is fitted by a mixture of exponential power distributions.

2·2. *Distribution modelling for test scores*

Efron (2004, 2007) argued that correlations between test statistics can considerably widen or narrow the theoretical null distribution, and thus suggested the use of a general normal distribution as the null distribution. In our study of microarray data, we found that even this suggestion is inappropriate for some datasets. For example, for the dark-dark dataset studied in § 4·3, an exponential power distribution is apparently more appropriate than is a normal distribution because the decay rate of the tail area of the histogram, see Fig. 2(d), is much lower than that of any normal distribution.

We model $f(z)$ by a mixture of exponential power distributions,

$$f(z \mid \tilde{\theta}) = \sum_{i=1}^{m} \omega_i \varphi_i(z \mid \mu_i, \alpha_i, \beta_i), \tag{4}$$

where $\tilde{\theta} = (\mu_1, \alpha_1, \beta_1; \ldots; \mu_m, \alpha_m, \beta_m; \omega_1, \ldots, \omega_{m-1})$ contains all the parameters of the model, m is the total number of components, ω_i is the weight of the ith component, with $0 < \omega_i \leqslant 1$ and $\sum_{i=1}^{m} \omega_i = 1$, and

$$\varphi_i(z \mid \mu_i, \alpha_i, \beta_i) = \frac{\beta_i}{2\alpha_i \Gamma(1/\beta_i)} \exp\{-(|\, z - \mu_i \,| / \alpha_i)^{\beta_i}\}, \quad -\infty < \mu_i < \infty,\ \alpha_i > 0,\ \beta_i > 1, \tag{5}$$

where the location parameter μ_i represents the centre of the distribution, the scale parameter α_i represents the dispersion and the shape parameter β_i represents the rate of exponential decay. For $\beta_i = 2$, the distribution (6) is $N(\mu_i, \alpha_i^2/2)$; for $1 < \beta_i < 2$, the distributed is heavy-tailed; and for $\beta_i > 2$, the distribution is light-tailed. To avoid singularity of the Kullback–Leibler distance, we restrict β_i to be bounded above. To make the components of the mixture distribution identifiable, we impose the constraint $\mu_1 < \cdots < \mu_m$ on the means of the components in (5). The identifiability of a finite mixture of exponential power distributions has been established by Holzmann et al. (2006). For mathematical simplicity, we use the reparameterization $\alpha_i^* = \log(\alpha_i)$, $\beta_i^* = \log(\beta_i - 1)$ and $\omega_i = \exp(\omega_i^*)/\sum_{j=1}^m \exp(\omega_j^*)$, for $i = 1, \ldots, m-1$, with $\omega_m^* = 0$. For uniformity of notation, we define $\mu_i^* \equiv \mu_i$. In what follows, we denote $f(z \mid \tilde{\theta})$ by $f(z \mid \theta)$, where $\theta = (\mu_1^*, \alpha_1^*, \beta_1^*; \cdots; \mu_m^*, \alpha_m^*, \beta_m^*; \omega_1^*, \ldots, \omega_{m-1}^*)$ lies in the space $\mathbb{R}^{4m-1}$.

Given the number of components, the mixture distribution can be estimated by minimizing the Kullback–Leibler distance defined in (2). The Kullback–Leibler distance is a suitable measure for density estimation and unsupervised machine learning (White, 1989). Since the true density $f(z)$ is unknown, the stochastic approximation method can be used to solve this indirectly solvable problem. After differentiating $\mathrm{KL}(f_\theta, f)$ with respect to θ, we have

$$\frac{\partial \mathrm{KL}(f_\theta, f)}{\partial \mu_i^*} = -\int P(i \mid z) \left\{ (-1)^{I(z<\mu_i)} \frac{\beta_i}{\alpha_i} \left| \frac{z-\mu_i}{\alpha_i} \right|^{\beta_i - 1} \right\} dF(z), \tag{6}$$

$$\frac{\partial \mathrm{KL}(f_\theta, f)}{\partial \alpha_i^*} = -\int P(i \mid z) \left\{ -\frac{1}{\alpha_i} + \frac{\beta_i}{\alpha_i} \left(\frac{|z-\mu_i|}{\alpha_i} \right)^{\beta_i} \right\} \alpha_i dF(z), \tag{7}$$

$$\frac{\partial \mathrm{KL}(f_\theta, f)}{\partial \beta_i^*} = -\int P(i \mid z) \left\{ \frac{1}{\beta_i} + \frac{1}{(\beta_i)^2} \frac{\Gamma'(1/\beta_i)}{\Gamma(1/\beta_i)} - \left| \frac{z-\mu_i}{\alpha_i} \right|^{\beta_i} \log \left| \frac{z-\mu_i}{\alpha_i} \right| \right\} (\beta_i - 1) dF(z), \tag{8}$$

$$\frac{\partial \mathrm{KL}(f_\theta, f)}{\partial \omega_i^*} = -\int \{P(i \mid z) - \omega_i\} \, dF(z), \tag{9}$$

where $I(\cdot)$ is the indicator function, the index i ranges from 1 to m in equations (6)–(8) and from 1 to $m-1$ in equation (9); $P(i \mid z) = \omega_i \varphi_i(z \mid \mu_i, \alpha_i, \beta_i)/f(z \mid \theta)$; and $\Gamma'(x)/\Gamma(x) = -1/x - \upsilon + \sum_{k=1}^{\infty}\{1/k - 1/(k+x)\}$ with $\upsilon \doteq 0{\cdot}577216$ being Euler's constant. Define

$$H_{\mu_i^*}(\theta, z) = P(i \mid z) \left\{ (-1)^{I(z<\mu_i)} \frac{\beta_i}{\alpha_i} \left| \frac{z-\mu_i}{\alpha_i} \right|^{\beta_i - 1} \right\}, \tag{10}$$

$$H_{\alpha_i^*}(\theta, z) = P(i \mid z) \left\{ -\frac{1}{\alpha_i} + \frac{\beta_i}{\alpha_i} \left(\frac{|z-\mu_i|}{\alpha_i} \right)^{\beta_i} \right\} \alpha_i, \tag{11}$$

$$H_{\beta_i^*}(\theta, z) = P(i \mid z) \left\{ \frac{1}{\beta_i} + \frac{1}{(\beta_i)^2} \frac{\Gamma'(1/\beta_i)}{\Gamma(1/\beta_i)} - \left| \frac{z-\mu_i}{\alpha_i} \right|^{\beta_i} \log \left| \frac{z-\mu_i}{\alpha_i} \right| \right\} (\beta_i - 1), \tag{12}$$

$$H_{\omega_i^*}(\theta, z) = P(i \mid z) - \omega_i, \tag{13}$$

where the index i ranges from 1 to m in (10)–(12) and from 1 to $m-1$ in (13).

Let $\{z_1, \ldots, z_N\}$ denote a finite set of samples drawn from $F(z)$. Conditioned on the samples $\{z_1, \ldots, z_N\}$, the following stochastic approximation algorithm is used to solve the system of

equations

$$\begin{cases} \int H_{\mu_i^*}(\theta, z)dF(z) = 0 \quad (i = 1, \ldots, m), \\ \int H_{\alpha_i^*}(\theta, z)dF(z) = 0 \quad (i = 1, \ldots, m), \\ \int H_{\beta_i^*}(\theta, z)dF(z) = 0 \quad (i = 1, \ldots, m), \\ \int H_{\omega_i^*}(\theta, z)dF(z) = 0 \quad (i = 1, \ldots, m-1). \end{cases} \tag{14}$$

Let $\theta_t = (\mu_1^{*(t)}, \alpha_1^{*(t)}, \beta_1^{*(t)}; \cdots; \mu_m^{*(t)}, \alpha_m^{*(t)}, \beta_m^{*(t)}; \omega_1^{*(t)}, \ldots, \omega_{m-1}^{*(t)})$ denote the working estimate of θ obtained at iteration t. One iteration of the algorithm is as follows.

Stage 1. Draw Z_t from the set $\{z_1, \ldots, z_N\}$ at random and with replacement.

Stage 2. Update θ_t in the following equations:

$$\begin{aligned} \mu_i^{*(t+1)} &= \mu_i^{*(t)} + \gamma_t H_{\mu_i^*}(\theta_t, Z_t) \quad (i = 1, \ldots, m), \\ \alpha_i^{*(t+1)} &= \alpha_i^{*(t)} + \gamma_t H_{\alpha_i^*}(\theta_t, Z_t) \quad (i = 1, \ldots, m), \\ \beta_i^{*(t+1)} &= \beta_i^{*(t)} + \gamma_t H_{\beta_i^*}(\theta_t, Z_t) \quad (i = 1, \ldots, m), \\ \omega_i^{*(t+1)} &= \omega_i^{*(t)} + \gamma_t H_{\omega_i^*}(\theta_t, Z_t) \quad (i = 1, \ldots, m-1). \end{aligned} \tag{15}$$

In a supporting document, available from http://www.stat.tamu.edu/∼fliang, we show that, as $t \to \infty$, θ_t converges to a solution of (14) almost surely conditioned on the set $\{z_1, \ldots, z_N\}$, where the test scores can be generally dependent. In practice, the convergence can be diagnosed by checking the positive definiteness of the Hessian matrix of (2), under the assumption that the Hessian matrices exist and are positive definite at the solution points. It follows from the main theorem of Furrer (2002) that the resulting M-estimator is consistent under regularity conditions. In equation (15), γ_t is called the gain factor and is subject to the conditions

$$\gamma_t > 0, \quad \lim_{t\to} \gamma_t = 0, \quad \sum_{t=1}^{\infty} \gamma_t = \infty, \quad \sum_{t=1}^{\infty} \gamma_t^{1+\epsilon} < \infty,$$

where $\epsilon > 0$ can be any positive number. In this paper, we set

$$\gamma_t = \frac{\gamma_0 t_0}{\max(t_0, t)} \quad (t = 1, 2, \ldots),$$

for some values of $t_0 > 1$ and $\gamma_0 > 0$. The default setting is $t_0 = 10\,000$ and $\gamma_0 = 0{\cdot}02$. In addition, we set a default value of $500 \times N$ for the total number of iterations, where $N = \#\{z_1, \ldots, z_N\}$ is the number of genes in the dataset.

The subject of stochastic approximation was founded by Robbins & Monro (1951), and convergence of the algorithm has been studied by Kushner (1981), Yin & Zhu (1989), Tadić (1997) and Chen (1998), among others. A common condition required by them is that the functions $H_{\mu_i^*}(\theta, z)$, $H_{\alpha_i^*}(\theta, z)$, $H_{\beta_i^*}(\theta, z)$ and $H_{\omega_i^*}(\theta, z)$ be locally Lipschitz-continuous with respect to θ in $\mathbb{R}^d$, where d is the dimension of θ. It is easy to see that this condition is not satisfied by the above algorithm, where $H_{\mu_i^*}(\theta, z)$ is not locally Lipschitz-continuous with respect to μ. In the supporting document, we establish convergence of the algorithm under the condition of Hölder continuity, which extends the range of applications.

2·3. *Estimation of* π_0

Since the test scores are transformed from p-values via the function $Z_i = \Phi^{-1}(1 - P_i)$, a larger test score gives more evidence that the alternative hypothesis is true. Thus, the components

included in $f_0(z)$ should have smaller means than those included in $f_1(z)$. Given $f(z \mid \hat{\theta})$, where $\hat{\theta}$ denotes an estimator of θ, the problem of estimating FDR is reduced to a clustering problem: we evaluate the distances between the neighbouring components of $f(z \mid \hat{\theta})$ and propose a clustering criterion for the components. Suppose that the components of $f(z \mid \hat{\theta})$ have been ordered such that $\hat{\mu}_1 \leqslant \cdots \leqslant \hat{\mu}_m$. The distance between two neighbouring components is defined as

$$d_{\mathrm{KL}}(\hat{\varphi}_i, \hat{\varphi}_{i+1}) = \left\{ \mathrm{KL}(\hat{\varphi}_i, \hat{\varphi}_{i+1}) + \frac{\mathrm{KL}(\hat{\varphi}_{i+1}, \hat{\varphi}_i)}{2} \right\} \quad (i = 1, \ldots, m-1), \tag{16}$$

where $\hat{\varphi}_j$ denotes component j of $f(z \mid \hat{\theta})$. Expression (16) has been used by Cook & Weisberg (1992, p. 163) as a distance measure between two distributions. Since $d_{\mathrm{KL}}(\hat{\varphi}_i, \hat{\varphi}_{i+1})$ is not analytically available for exponential power distributions, it is estimated using Metropolis–Hastings samples simulated from $\hat{\varphi}_i$ and $\hat{\varphi}_{i+1}$. Given the distance sequence, we cluster the first m_0 components into the group f_0 with

$$\begin{aligned} m_0 = \min \Bigg\{ & i : d_{\mathrm{KL}}(\hat{\varphi}_i, \hat{\varphi}_{i+1}) > \max\{d_{\mathrm{KL}}(\hat{\varphi}_{i+1}, \hat{\varphi}_{i+2}), d_{\mathrm{KL}}(\hat{\varphi}_{i-1}, \hat{\varphi}_i)\}, \\ & \mu_i - \mu_b > \left\{ \frac{\Gamma(3/\hat{\beta}_b)}{\Gamma(1/\hat{\beta}_b)} \right\}^{1/2} \hat{\alpha}_b, i = b+1, \ldots, m-1 \Bigg\}, \end{aligned} \tag{17}$$

where $d_{\mathrm{KL}}(\hat{\varphi}_0, \hat{\varphi}_1) = d_{\mathrm{KL}}(\hat{\varphi}_m, \hat{\varphi}_{m+1}) = 0$, $b = \arg\max_i \hat{\omega}_i$ corresponds to the component with the largest probability, $(\hat{\mu}_b, \hat{\alpha}_b, \hat{\beta}_b)$ are the parameters of component b and $\{\Gamma(3/\hat{\beta}_b)/\,\Gamma(1/\hat{\beta}_b)\}\hat{\alpha}_b^2$ is the variance of component b (Johnson et al., 1980). Under the criterion (18), $d_{\mathrm{KL}}(\hat{\varphi}_{m_0}, \hat{\varphi}_{m_0+1})$ corresponds to a local maximum of the distance sequence. Given m_0, π_0 can then be simply estimated by $\hat{\pi}_0 = \sum_{i=1}^{m_0} \omega_i$. The inequality constraint on the distance between μ_i and μ_b reflects our belief that π_0 is around or greater than 0·9. This constraint is only enforced when m is large enough such that the major component has only a small proportion value. Hence, it robustifies the estimation of π_0 to the choice of m. We also tried the criterion

$$m_0 = \arg\max_i d_{\mathrm{KL}}(\hat{\varphi}_i, \hat{\varphi}_{i+1}), \tag{18}$$

but found that the estimator of π_0 based on (18) is less robust to the choice of m than is the estimator based on (17). When m is very large, π_0 tends to be overestimated under the criterion (18) because of the unreliable estimators for the components with small proportion values. Other distance measures, such as the Hellinger distance, could be used in place of the Kullback–Leibler distance.

2·4. *Estimation*

It is necessary to determine m, the number of components of the mixture distribution. Rather than choosing a single value of m, we choose a range of m and then use the ensemble averaging method, discussed by Wolpert (1992) and Hashem (1997), to estimate FDR.

Let x denote the new input data of a model, let y denote the corresponding response, and let $S_1(x), \ldots, S_L(x)$ denote L input-output functions realized by the model based on the same training dataset. These functions can vary in various respects, such as model structure and training conditions. The ensemble averaging method uses a combined predictor,

$$S_{\mathrm{EA}}(x) = \sum_{i=1}^{L} \lambda_i S_i(x),$$

to predict $E(Y \mid X = x)$, where λ_i is called the ensemble weight of $S_i(x)$, $0 \leqslant \lambda_i \leqslant 1$, for all i, and $\sum_{i=1}^{L} \lambda_i = 1$. The choice of λ can depend on the predictors $S_1(x), \ldots, S_L(x)$. The simplest

approach is to set $\lambda_i = 1/L$ for $i = 1, \ldots, L$; the resulting predictor $S_{\mathrm{EA}}(x)$ will have a smaller variance than any $S_i(x)$. Another use of ensemble averaging is to reduce the prediction bias caused by underfitting or overfitting by averaging over different models. We use ensemble averaging to estimate FDR as follows.

Step 1. Determine the model range $M = M_l, \ldots, M_u$.

Step 2. For each value of m, run the stochastic approximation algorithm L_m times independently. Let $\widehat{\mathrm{FDR}}_{m,i}(\Lambda)$ denote the Bayesian estimate of FDR resulting from the ith run, for $i = 1, \ldots, L_m$.

Step 3. Evaluate the final estimate of $\mathrm{FDR}(\Lambda)$ as

$$\widehat{\mathrm{FDR}}_{\mathrm{EA}}(\Lambda) = \sum_{m=M_l}^{M_u} \frac{\lambda_m}{L_m} \sum_{i=1}^{L_m} \widehat{\mathrm{FDR}}_{m,i}(\Lambda),$$

where $0 \leqslant \lambda_m \leqslant 1$ and $\sum_{m=M_l}^{M_u} \lambda_m = 1$.

In Step 1, the model range can be determined by some pilot runs. First, run the stochastic approximation algorithm with different choices of m and calculate the statistic

$$\widetilde{\mathrm{BIC}}_m = -\frac{1}{N} \sum_{i=1}^{N} \log f^{(m)}(z_i \mid \hat{\theta}) + \frac{p_m \log(N)}{2N},$$

where $f^{(m)}(\cdot)$ denotes the mixture density of exponential power distributions with m components, and $p_m = 4m - 1$ is the number of parameters in $f^{(m)}(\cdot)$. Since the test scores can be generally dependent, we call $\widetilde{\mathrm{BIC}}_m$ the pseudo-BIC statistic. Next, identify the value of m that minimizes the pseudo-BIC values. Set $m_c = \arg\min_m \widetilde{\mathrm{BIC}}_m$, $M_l = \max\{2, m_c - h\}$ and $M_u = m_c + h$, where h is a positive integer and is usually set to be 1 or 2. Although BIC is defined for independent data, we expect it to give a rough estimate of the order of models for dependent data.

For simplicity, we set $L_m = 1$ in Step 2 and $\lambda_m = 1/(M_u - M_l + 1)$ in Step 3, although this setting may not be optimal. It may be better to link the setting of λ_m with $\widetilde{\mathrm{BIC}}_m$, perhaps setting $\lambda_m \propto \exp\{-\widetilde{\mathrm{BIC}}_m / t_s\}$, where $t_s > 0$ is a scale parameter. When t_s is large, each model in the ensemble is approximately equally weighted; when t_s is small, the models with smaller values of $\widetilde{\mathrm{BIC}}_m$ tend to be heavily weighted.

Given $\widehat{\mathrm{FDR}}_{\mathrm{EA}}(\Lambda)$, the Bayesian q-value can be estimated using the plug-in method by

$$\hat{q}_{\mathrm{EA}}(z) = \inf_{\{\Lambda : z \in \Lambda\}} \widehat{\mathrm{FDR}}_{\mathrm{EA}}(\Lambda), \tag{19}$$

for any z. In this paper, $\hat{q}_{\mathrm{EA}}(z)$ is used as a reference quantity for the decision of multiple hypothesis tests.

3. Simulation examples

3·1. *Example I*

This example was modified from a microarray example in Qiu et al. (2005). The datasets were generated as follows. First, we generated a 1255×40 matrix $X = (x_{ij})$, $i = 1, \ldots, 1255$, $j = 1, \ldots, 40$. The x_{ij} were all stochastically independent, but the elements with $i = 1, \ldots, 125$ and $j = 1, \ldots, 21$ were drawn from $N(2, 1)$, and the others from $N(0, 1)$. The first 125 rows model the differentially expressed genes. Next, we generated a 40-dimensional random vector, $a = (a_1, \ldots, a_{40})$, with elements drawn independently from $N(0, 1)$. Define

$$y_{ij} = \rho^{1/2} a_j + (1 - \rho)^{1/2} x_{ij} \quad (i = 1, \ldots, 1255, j = 1, \ldots, 40).$$

Thus, we have $\text{corr}(y_{i_1,j}, y_{i_2,j}) = \rho$ for any $i_1 \neq i_2$ and any j. For each gene, a two-sample t-test was employed to test whether or not its mean expression levels were different under the two experimental conditions, that is, the first 21 columns were compared to the other 19 columns. The corresponding p-values were calculated and test scores were transformed via the function $Z = \Phi^{-1}(1 - P)$. For each of the correlation coefficients $\rho = 0, 0{\cdot}3$ and $0{\cdot}6$, 50 independent datasets were generated by repeating the above process with different random seeds. To examine the effect of correlations on the performance of our method, the independent case, $\rho = 0$, is included.

For all of the datasets, the minimum pseudo-BIC values from our stochastic approximation method are attained at $m = 2$ or 3. Hence, we set $M_l = 2$ and $M_u = 4$ for this example. Figure 1(a)–(c) summarizes the computational results for a dataset generated with $\rho = 0{\cdot}6$. The sharp increase in the number of significant genes at $q = 0{\cdot}0$, as shown in Fig. 1(b), implies the effectiveness of our method. The 125 differentially expressed genes can be identified even with a very small Bayesian q-value. The approximate equality of the true false discovery rate and the Bayesian q-value, as shown in Fig. 1(c), implies the validity of the Bayesian q-value as a reference quantity for the decision of multiple tests.

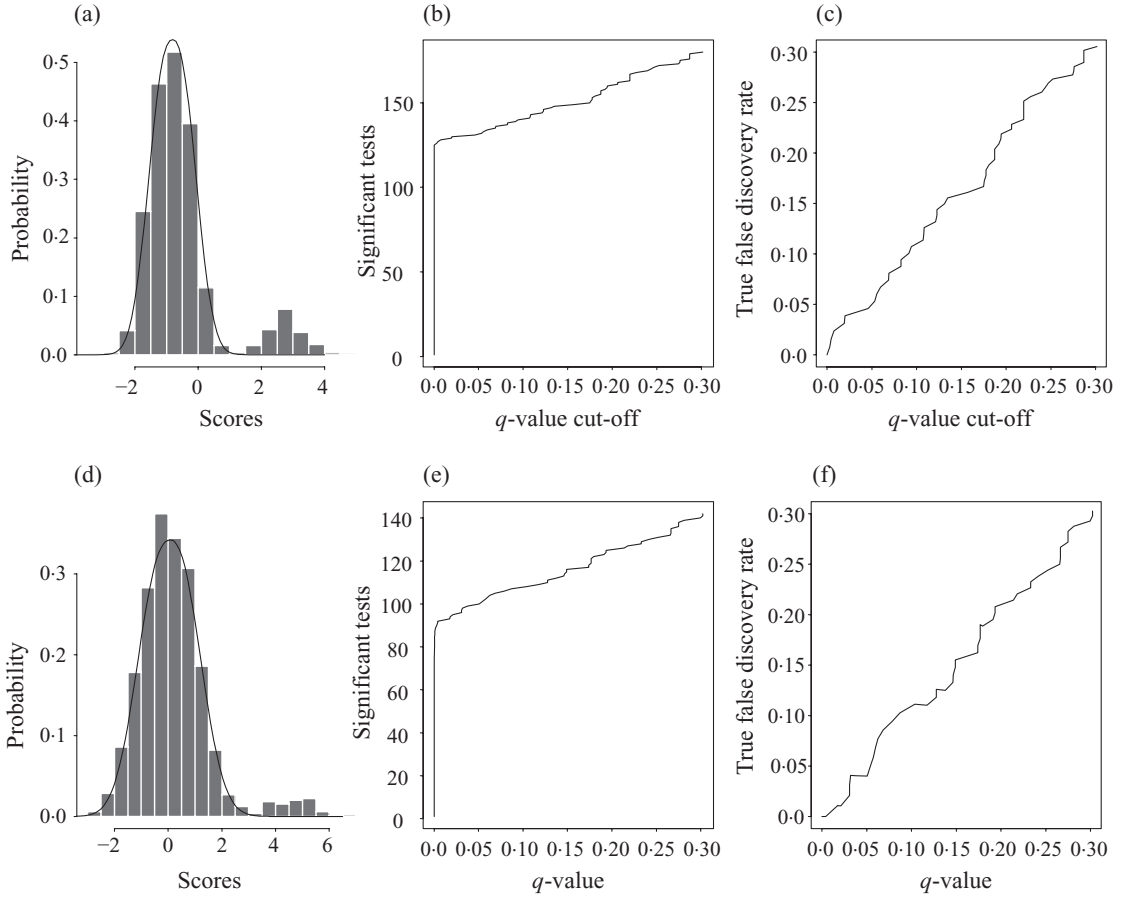

Fig. 1. Computational results for the stochastic approximation method for two simulated datasets, panels (a)–(c) for a dataset generated in Example I with $\rho = 0{\cdot}6$, and panels (d)–(f) for a dataset generated in Example II. Panels (a) and (d) show histograms of test scores and the fitted density curve of $f_0(z)$. Panels (b) and (e) show numbers of significant genes versus Bayesian q-values. Panels (c) and (f) show true false discovery rates versus Bayesian q-values.

Table 1. *Simulation study. Computational results for the datasets generated with* $\rho = 0$ *and* $\rho = 0{\cdot}6$. *The values are averages and estimated standard deviations, in parentheses, based on* 50 *datasets. Here* $\Lambda(q)$ *denotes a rejection region with the nominal value of* FDR *equal to* q. *For different methods,* q *has different meanings: for the* SA *method,* q *refers to* q_{EA} *given in* (19)*; for E-spline,* q *refers to the Bayesian* q*-value given in* (3)*; for S-*FDR, q *refers to the* q*-value defined in Storey* (2002)*; and for the two-stage method,* q *refers to* FDR *as defined in Benjamini & Hochberg* (1995)

Method	$\hat{\pi}_0$	Measure	Λ (0·3)	Λ (0·2)	Λ (0·1)	Λ (0·05)
			$\rho = 0$			
SA	0·9003 (0·0002)	tFDR	0·306 (0·005)	0·202 (0·005)	0·101 (0·003)	0·052 (0·003)
		Specificity	0·951 (0·001)	0·972 (0·001)	0·988 (0·001)	0·994 (0·0)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
E-Spline	0·8883 (0·0019)	tFDR	0·336 (0·014)	0·229 (0·011)	0·124 (0·007)	0·065 (0·004)
		Specificity	0·940 (0·004)	0·966 (0·002)	0·984 (0·001)	0·992 (0·001)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
S-FDR	0·9010 (0·0084)	tFDR	0·302 (0·007)	0·201 (0·006)	0·101 (0·004)	0·053 (0·003)
		Specificity	0·951 (0·002)	0·972 (0·001)	0·987 (0·001)	0·994 (0·0)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
Two-stage	—	tFDR	0·208 (0·005)	0·151 (0·005)	0·083 (0·003)	0·047 (0·002)
		Specificity	0·971 (0·001)	0·980 (0·001)	0·990 (0·0)	0·995 (0·0)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
			$\rho = 0{\cdot}6$			
SA	0·9003 (0·0002)	tFDR	0·304 (0·004)	0·202 (0·004)	0·099 (0·003)	0·047 (0·002)
		Specificity	0·951 (0·001)	0·972 (0·001)	0·988 (0·0)	0·994 (0·0)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
E-Spline	0·8896 (0·0024)	tFDR	0·331 (0·012)	0·226 (0·011)	0·124 (0·008)	0·065 (0·005)
		Specificity	0·942 (0·003)	0·966 (0·002)	0·984 (0·001)	0·992 (0·001)
		Sensitivity	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	0·999 (0·0)
S-FDR	0·6379 (0·0651)	tFDR	0·184 (0·051)	0·163 (0·049)	0·097 (0·038)	0·059 (0·030)
		Specificity	0·800 (0·057)	0·822 (0·054)	0·904 (0·041)	0·939 (0·034)
		Sensitivity	0·621 (0·067)	0·605 (0·066)	0·570 (0·066)	0·524 (0·065)
Two-stage	—	tFDR	0·153 (0·035)	0·103 (0·029)	0·059 (0·024)	0·039 (0·021)
		Specificity	0·933 (0·028)	0·950 (0·028)	0·969 (0·021)	0·977 (0·019)
		Sensitivity	0·948 (0·023)	0·930 (0·027)	0·899 (0·032)	0·847 (0·039)

tFDR, true false discovery rate; SA, the new stochastic approximation method; E-Spline, Efron's spline method (Efron, 2004); S-FDR, Storey et al.'s positive FDR method (Storey et al., 2004); Two-stage, the two-stage method of Benjamini et al. (2006).

For a thorough assessment, we also calculated the specificity and sensitivity of the tests. The specificity $U/(U + V)$ is the proportion of correctly identified genes that were not differentially expressed, and the sensitivity $S/(T + S)$ is the proportion of correctly identified differentially expressed genes. The average of the sensitivity values over multiple datasets, reported in Table 1, provides a natural estimate of the average power (Dudoit et al., 2003), a commonly adopted measure of the quality of multiple testing. To save space, numerical results for $\rho = 0{\cdot}3$ are not given but are available in a longer version of the paper available from the authors.

Table 2. *Computational results for the datasets generated with* $\rho = 0{\cdot}95$. *See caption to Table 1 for details*

Method	$\hat{\pi}_0$	Measure	Λ (0·3)	Λ (0·2)	Λ (0·1)	Λ (0·05)
SA	0·9003 (0·0002)	tFDR	0·294 (0·003)	0·192 (0·003)	0·095 (0·002)	0·048 (0·002)
		Spec.	0·954 (0·001)	0·974 (0·001)	0·988 (0·0)	0·994 (0·0)
		Sens.	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
E-Spline	0·8928 (0·0018)	tFDR	0·313 (0·010)	0·211 (0·008)	0·108 (0·005)	0·059 (0·003)
		Spec.	0·948 (0·003)	0·970 (0·002)	0·986 (0·001)	0·993 (0·0)
		Sens.	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)	1·0 (0·0)
S-FDR	0·8421 (0·0455)	tFDR	0·036 (0·025)	0·018 (0·018)	0·018 (0·018)	0·018 (0·018)
		Spec.	0·960 (0·028)	0·980 (0·020)	0·98 (0·020)	0·98 (0·020)
		Sens.	0·040 (0·028)	0·020 (0·020)	0·020 (0·020)	0·020 (0·020)
Two-stage	—	tFDR	0·109 (0·042)	0·090 (0·039)	0·072 (0·035)	0·072 (0·035)
		Spec.	0·880 (0·046)	0·900 (0·043)	0·920 (0·039)	0·92 (0·039)
		Sens.	0·233 (0·059)	0·199 (0·054)	0·132 (0·047)	0·098 (0·042)

tFDR, true false discovery rate; Spec., specificity; Sens., sensitivity; SA, the new stochastic approximation method.

For comparison, the positive false discovery rate method of Storey et al. (2004), the two-stage method of Benjamini et al. (2006) and the spline method of Efron (2004) were also applied to this example. These methods all allow certain forms of dependence between test statistics. The software for the first was downloaded from http://faculty.washington.edu/∼jstorey/ and was run with the default setting. The other two methods were implemented by F. Liang in an unpublished Texas A&M University technical report. In the third, $f(z)$ was estimated using the log-spline density estimation method (Kooperberg & Stone, 1992; Stone et al., 1997), but $f_0(z)$ and π_0 were estimated by the method proposed by Efron (2004). The log-spline method avoids the issue of sample interval selection involved in the Poisson regression method.

The numerical results reported in Table 1 indicate that our algorithm outperforms the other three methods for this example. It produces fairly accurate estimates for π_0 and the false discovery rate, and fairly high specificity and sensitivity values. For the datasets with $\rho = 0$, Storey's method performs as well as ours. This is not surprising as those datasets satisfy all the conditions Storey's method requires: the test scores are weakly dependent, and the null p-values are uniformly distributed on [0, 1]. However, when the correlation coefficient deviates from 0, Storey's method performs less well; the true false discovery rate tends to deviate widely from its nominal level, and both specificity and sensitivity decrease as ρ increases. The two-stage method works well for the cases $\rho = 0$ and 0·3: the true false discovery rate is below the nominal level, and the specificity and sensitivity are fairly high. However, as ρ increases, its performance deteriorates: both specificity and sensitivity decrease. This can be seen more clearly in Table 2. Since, in the two-stage method, estimation of π_0 depends on the nominal false discovery rate specified by the user, the estimate of π_0 by this method is not reported in the paper. Efron's method is robust to the correlations between test scores: neither specificity nor sensitivity is much changed as ρ increases from 0 to 0·6.

Now we consider an extreme case of this example. Fifty datasets were generated independently with $\rho = 0{\cdot}95$. As a result of the strong correlations between the simulated gene expression data, the test scores in these datasets tend to separate into two distinct groups, which correspond to $f_0(z)$ and $f_1(z)$, respectively, so we set $M_l = M_u = 2$ in the runs of our method. Different values of m were tried for our stochastic approximation method, the pseudo-BIC criterion suggesting the

same setting for the range of models. Table 2 shows that our and Efron's methods still work well, whereas the other have very low sensitivity values. Our method worked well even for $\rho = 0{\cdot}99$.

Ours and Efron's methods base inference about the false discovery rate on the empirical null distribution, whereas the others base inference about the FDR on the theoretical null distribution. The use of empirical null distributions is important for multiple hypothesis testing, especially when the correlation between test scores is high. As discussed by Efron (2007), correlations between test statistics can considerably widen or narrow the theoretical null distribution, and the use of the empirical null distribution helps to accommodate this.

3·2. *Example II*

This example was modified from an example studied in Liang et al. (2007). Let $x_i = (x_{i,1}, \ldots, x_{i,10})$ be the expression values simulated for gene i. Suppose that the distribution of the expression levels can deviate from normal and that there may be dependence among genes. First, we allow the error distribution to be nonnormal. Let

$$\frac{x_{ij} - \mu_i^{(1)}}{\sigma_i} \sim t(\nu)\ (j = 1, \ldots, 5), \quad \frac{x_{ij} - \mu_i^{(2)}}{\sigma_i} \sim t(\nu) \quad (j = 6, \ldots, 10), \tag{20}$$

where $\mu_i^{(1)}$ and $\mu_i^{(2)}$ are the respective mean expression levels of gene i under two different conditions, and σ_i^2 is a random variable drawn from the inverse gamma distribution IG(2·5, 0·5). The parameters of IG$(\cdot, \cdot)$ are computed from a real gene expression dataset, the light-dark data studied in §4. In analyzing the real dataset, we found that the distribution of the gene expression levels is consistent with a t-distribution, $t(\nu)$, with ν ranging from 3 to 5.

Secondly, we create some level of dependence among genes. Conditional on the expression levels of gene l, we set

$$\frac{x_{ij} - \mu_{ij|l}^{(1)}}{\sigma_{i|l}} \sim t(\nu),\ (j = 1, \ldots, 5), \quad \frac{x_{ij} - \mu_{ij|l}^{(2)}}{\sigma_{i|l}} \sim t(\nu),\ (j = 6, \ldots, 10), \tag{21}$$

where $\mu_{ij|l}^{(a)} = \mu_i^{(a)} + \rho_{il}\sigma_i/\sigma_l(x_{lj} - \mu_l^{(a)})$ for $a = 1, 2$, $\sigma_{i|l} = \sigma_i(1 - \rho_{il}^2)^{1/2}$, and $\rho_{il} \sim$ Un$[-1, 1]$. If $t(\nu)$ is replaced by $N(0, 1)$ in equation (21), then x_{ij} and x_{lj} are normal random variables with correlation coefficient ρ_{il}. Here x_{ij} and x_{lj} are Student-t random variables. Our simulation results show that they tend to have stronger correlations than ρ_{il}.

We generate the gene expression data x_s $(s = 1, \ldots, N)$, by first drawing l from the set $\{1, \ldots, s\}$ at random. Then, if $l = s$, we generate x_s according to equation (20); otherwise, we generate x_s according to equation (21) conditional on the expression profile of gene l.

A total of 50 datasets were generated independently. Each dataset consists of 2100 genes, which are generated with $\upsilon = 4$, $\mu_i^{(1)} = 0$, for all i, $\mu_i^{(2)} = 0$, for $i = 1, \ldots, 2000$, and $\mu_i^{(2)} \sim N(5, 1)$, for $i = 2001, \ldots, 2100$. Thus, the last 100 genes are differentially expressed in this example.

The four previous methods were applied to this example. For our stochastic approximation method, we set $M_l = 2$ and $M_u = 3$. Figure 1(d)–(f) indicates that, even if the error distribution of the gene expression values deviates from normal and the gene expression values are dependent, our method can still work well, and the Bayesian q-value can still work as a valid reference quantity for multiple hypothesis testing.

The numerical results reported in Table 3 indicate that the our new method outperforms the other three methods for this example, producing very accurate estimates for π_0 and the false discovery rate, and fairly high specificity, sensitivity and power.

Table 3. *Computational results for Example II. The values are averages and estimated standard deviations, in parentheses, based on* 50 *datasets. Here* $\Lambda(q)$ *denotes a rejection region with the nominal value of* FDR *being* q. *For different methods,* q *has different meanings: for the* SA *method,* q *refers to* q_{EA} *given in* (19)*; for E-spline,* q *refers to the Bayesian* q*-value given in* (3)*; for S-*FDR*,* q *refers to the* q*-value defined in Storey* (2002)*; and for the two-stage method,* q *refers to* FDR *as defined in Benjamini & Hochberg* (1995)

Method	$\hat{\pi}_0$	Measure	Λ (0·3)	Λ (0·2)	Λ (0·1)	Λ (0·05)
SA	0·9531 (0·0002)	tFDR	0·292 (0·004)	0·200 (0·004)	0·100 (0·004)	0·051 (0·003)
		Spec.	0·979 (0·0)	0·987 (0·0)	0·994 (0·0)	0·997 (0·0)
		Sens.	0·993 (0·001)	0·991 (0·001)	0·984 (0·002)	0·974 (0·003)
E-Spline	0·9519 (0·0066)	tFDR	0·316 (0·023)	0·223 (0·019)	0·127 (0·013)	0·072 (0·008)
		Spec.	0·969 (0·006)	0·983 (0·003)	0·992 (0·001)	0·996 (0·001)
		Sens.	0·993 (0·001)	0·990 (0·002)	0·984 (0·002)	0·976 (0·003)
S-FDR	0·8046 (0·0076)	tFDR	0·243 (0·008)	0·156 (0·007)	0·078 (0·005)	0·038 (0·003)
		Spec.	0·984 (0·001)	0·991 (0·001)	0·996 (0·0)	0·998 (0·0)
		Sens.	0·992 (0·001)	0·989 (0·002)	0·982 (0·002)	0·974 (0·002)
Two-stage	—	tFDR	0·148 (0·006)	0·102 (0·005)	0·056 (0·003)	0·029 (0·002)
		Spec.	0·991 (0·0)	0·994 (0·0)	0·997 (0·0)	0·999 (0·0)
		Sens.	0·987 (0·001)	0·985 (0·002)	0·978 (0·002)	0·970 (0·002)

tFDR, true false discovery rate; Spec., specificity; Sens., sensitivity; SA, the new stochastic approximation method.

In addition to the improper specification of the theoretical null distribution, we suspect that the poor performance of the Storey's and the two-stage methods for this example are partly due to the incorrect p-values. We recalculated the p-values for this example using the permutation t-test, with all possible permutations, and found that the use of the permutation p-values improves the performance of both these methods. For example, the true false discovery rates produced by the two-stage method were 0·223, 0·163, 0·102 and 0·06 for the rejection regions with $q = 0{\cdot}3$, 0·2, 0·1 and 0·05, respectively, and the values produced by Storey's method were similar. The performance of the two methods on this example is still not perfect. For the two-stage method, the true false discovery rates overshoot the nominal levels at low significance levels and undershoot at high significance levels, and the sensitivity is rather low at level 0·05. For Storey's method, the sensitivity is also rather low at the level 0·05.

3·3. *Example III*

In this example, we examine the performance of the stochastic approximation method on pairwise dependence tests. The datasets were generated as follows. First, we generated a 50×30 matrix $X = (x_{ij})$. Secondly, we generated a 3×30 $A = (a_{ij})$. All elements of X and A are independently and identically distributed standard normal random variables. For $j = 1, \ldots, 30$, define

$$
\begin{aligned}
y_{ij} &= {\rho_1}^{1/2} a_{1j} + (1-\rho_1)^{1/2} x_{ij} \quad (i = 1, \ldots, 5),\\
y_{ij} &= \rho_2^{1/2} a_{2j} + (1-\rho_2)^{1/2} x_{ij} \quad (i = 6, \ldots, 10),\\
y_{ij} &= \rho_3^{1/2} a_{3j} + (1-\rho_3)^{1/2} x_{ij} \quad (i = 11, \ldots, 20),\\
y_{ij} &= x_{ij} \quad (i = 31, \ldots, 50),
\end{aligned}
$$

Table 4. *Computational results for Example III. See caption to Table 3 for details*

Method	$\hat{\pi}_0$	Measure	Λ (0·3)	Λ (0·2)	Λ (0·1)	Λ (0·05)
SA	0·9456 (0·0007)	tFDR	0·302 (0·009)	0·204 (0·009)	0·100 (0·008)	0·049 (0·006)
		Spec.	0·976 (0·001)	0·986 (0·001)	0·994 (0·001)	0·997 (0·0)
		Sens.	0·972 (0·006)	0·957 (0·008)	0·916 (0·015)	0·877 (0·020)
E-Spline	0·9312 (0·0069)	tFDR	0·386 (0·017)	0·285 (0·015)	0·162 (0·012)	0·090 (0·009)
		Spec.	0·961 (0·003)	0·976 (0·002)	0·989 (0·001)	0·995 (0·001)
		Sens.	0·978 (0·007)	0·964 (0·009)	0·938 (0·013)	0·910 (0·016)
S-FDR	0·9408 (0·0096)	tFDR	0·333 (0·013)	0·229 (0·011)	0·126 (0·009)	0·065 (0·007)
		Spec.	0·971 (0·002)	0·983 (0·001)	0·992 (0·001)	0·996 (0·0)
		Sens.	0·976 (0·005)	0·966 (0·007)	0·939 (0·010)	0·903 (0·015)
Two-stage	—	tFDR	0·246 (0·011)	0·180 (0·010)	0·106 (0·008)	0·058 (0·007)
		Spec.	0·981 (0·001)	0·988 (0·001)	0·994 (0·001)	0·997 (0·0)
		Sens.	0·968 (0·007)	0·956 (0·008)	0·929 (0·011)	0·896 (0·015)

tFDR, true false discovery rate; Spec., specificity; Sens., sensitivity; SA, the new stochastic approximation method.

where $\rho_1 = 0{\cdot}6$, $\rho_2 = 0{\cdot}8$ and $\rho_3 = 0{\cdot}7$. For each pair of genes (i_1, i_2) $(i_1, i_2 = 1, \ldots, 50)$, we calculate the sample correlation $\hat{\rho}_{i_1,i_2}$ and test the hypotheses $H_0 : \rho_{i_1,i_2} = 0$ versus $H_1 : \rho_{i_1,i_2} \neq 0$ using Fisher's Z-statistic,

$$Z^*_{i_1,i_2} = \frac{(30-3)^{1/2}}{2} \log\left(\frac{1+\hat{\rho}_{i_1,i_2}}{1-\hat{\rho}_{i_1,i_2}}\right),$$

which is approximately a standard normal random variable when the null hypothesis is true. The p-values of the tests, $P_{i_1,i_2} = 2\{1 - \Phi(|Z^*_{i_1,i_2}|)\}$, are calculated and the test scores are generated via the transformation $Z_{i_1,i_2} = \Phi^{-1}(1 - P_{i_1,i_2})$. In this dataset, there are $65 = 2C_5^2 + C_{10}^2$ true instances of H_1, and $1160 = C_{50}^2 - 65$ true instances of H_0, where C_m^k is the binomial coefficient of choosing k from m. As a result of the repeated use of the gene expression data in the calculation of $Z^*_{i_1,i_2}$, the test scores have a rather complicated dependence structure. This example mimics the multiple-comparisons problem encountered in gene network construction.

A total of 50 datasets were generated independently, and the four methods were applied to these. For the stochastic approximation method, we set $M_l = 3$ and $M_u = 5$. The numerical results reported in Table 4 indicate that it works well for this example. The true FDR values from the Storey's and Efron's methods are significantly higher than their nominal levels, as is also the case for the two-stage method at significance level 0·3.

4. Avian pineal gland gene expression data

4·1. *Introduction*

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 mechanism driving the rhythms. For this purpose, Dr. V. Cassone's laboratory at Texas A&M University measured the expression levels of pineal gland genes under light-dark and constant darkness, dark-dark, conditions. Under the light-dark condition, the birds were euthanized at 2, 6, 10, 14, 18 and 22 hours Zeitgeber time to obtain mRNA to produce adequate cDNA libraries. Four microarray chips per time-point were produced, and there were two replicates for each gene in each chip. The experiment was then repeated under the dark-dark

condition. Each chip produced gene expression signals for at least 7400 genes. Throughout the experiment, samples from the light-dark data with Zeitgeber time of 18 hours were used as controls. Gene expression levels relative to the controls were recorded and processed. Our goal is to identify genes that are differentially expressed at different time-points. Mixed-effects analysis with the fixed effect being the different time-points and the random effects corresponding to chips and biological batches were 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 this paper. Under both lighting conditions, p-values, P_i, for testing the existence of different time effects were produced and transformed to test scores using $\Phi^{-1}(1 - P_i)$.

4·2. *Light-dark data*

Our method was applied to the light-dark data. The computational results are summarized in Fig. 2. In our pilot runs, the minimum pseudo-BIC value was often attained at $m = 3$. We thus set $M_l = 2$ and $M_u = 5$ for this dataset. Figure 2(a) shows the histogram of the test scores and the estimated density curve of f_0. It suggests that f_0 can be well estimated by the stochastic approximation method. The run was repeated five times, and almost identical results were yielded in the runs. By averaging over the five runs, we obtained one estimate, $\hat{\pi}_0 = 0{\cdot}865$, for π_0, with standard deviation 0·001. By looking at the interaction of the fitted f_0 and the histogram, we know that the genes with test scores greater than 3·5 are suspiciously differentially expressed. Adding this to the information in Fig. 2(b) and (c), we know that about 1400 genes are suspiciously differentially expressed. Furthermore, among those genes, about $400 \doteq 1400 \times 28\%$ are false positive and about 1000 genes are really differentially expressed.

4·3. *Dark-dark data*

Our method was applied to the dark-dark data. Our pilot runs suggest that setting $M_l = 2$ and $M_u = 5$ is also appropriate for this dataset. Figure 2(d) shows that f_0 can be well estimated by the stochastic approximation method, even though it deviates substantially from normal. The run was repeated five times, yielding five estimates of $f_0(z)$ and an estimate $\hat{\pi}_0 = 0{\cdot}982$ with standard deviation 0·002. Figure 2(f) suggests that, at the test level 0·20, the Bayesian q-value, there are about 80 suspiciously differentially expressed genes, among which only about 65 genes are truly differentially expressed. All of the suspiciously differentially expressed genes have a test score greater than 3·8.

Storey's and Efron's methods were also applied to this dataset. As a result of the significant deviation of the empirical null distribution from normality and the dependence between test scores, both procedures perform less well for this dataset. For example, the estimates of π_0 produced by these two methods were 0·410 and 0·859, respectively. These estimates are not consistent with the histogram of the test scores.

5. DISCUSSION

Prior to preparing this paper, we tried to model $f(z)$ by a mixture of normal distributions. For Example I, the performance of the stochastic approximation method was not significantly affected by this change. However, for Example II, this change increased the variability of the false discovery rate estimates. In Example II, the error distribution of gene expression values deviates from normal. As a consequence, the distribution $f(z)$ needs to be approximated by a mixture with more components, and this usually makes the false discovery rate estimate more variable.

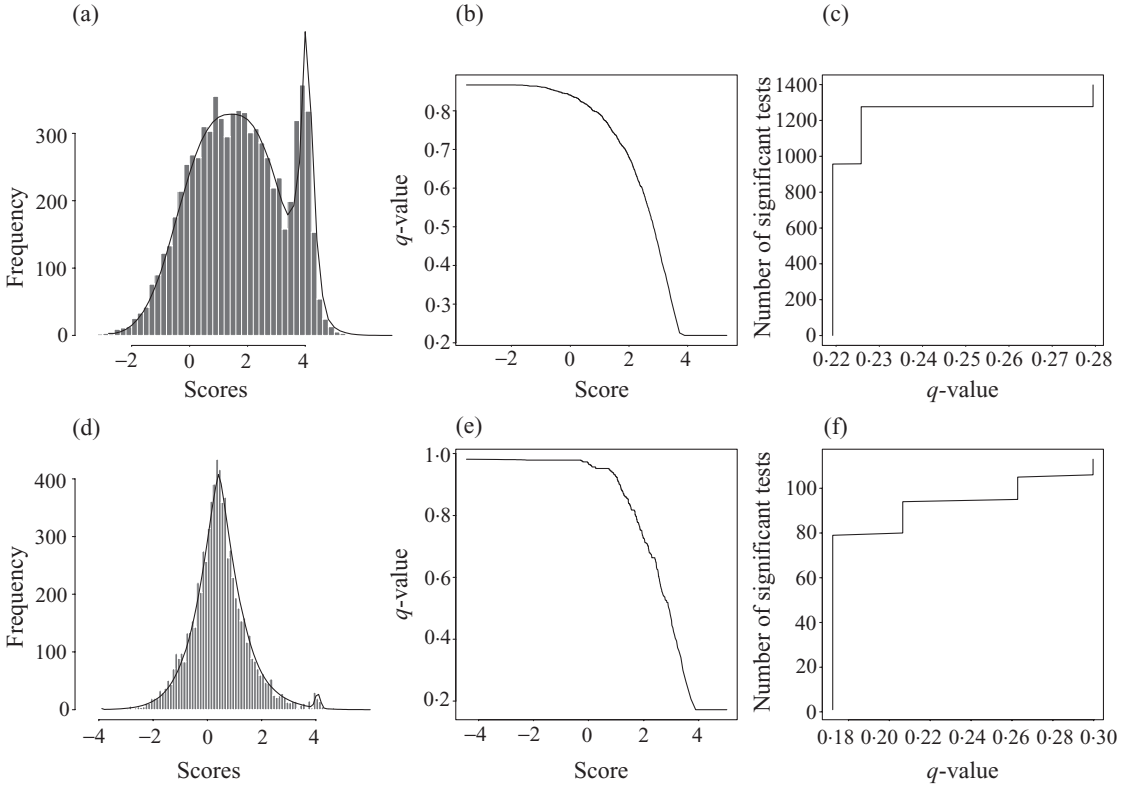

Fig. 2. Computational results for (a)–(c) the light-dark data and (d)–(f) the dark-dark data. Panels (a) and (d) show histograms of the test scores and the fitted density curve of $f_0(z)$ by the stochastic approximation method. Panels (b) and (e) show Bayesian q-values versus test scores. Panels (c) and (f) show numbers of significant genes versus Bayesian q-values.

As an alternative estimator of the Bayesian false discovery rate, we also considered

$$\widetilde{\text{FDR}}(\Lambda) = \frac{\hat{\pi}_0\{1 - \hat{F}_0(z_0)\}}{1 - \hat{F}(z_0)},$$

for a rejection region $\Lambda = \{Z_i \geqslant z_0\}$, where $\hat{\pi}_0$, $\hat{F}_0$ and $\hat{F}$ are estimates of π_0, F_0 and F produced by the stochastic approximation method. Our numerical results show that $\widetilde{\text{FDR}}(\Lambda)$ performs as well as $\widehat{\text{FDR}}(\Lambda)$ for most of our examples, and it is even better than $\widehat{\text{FDR}}(\Lambda)$ for Example II. However, we advocate $\widehat{\text{FDR}}(\Lambda)$ instead of $\widetilde{\text{FDR}}(\Lambda)$ because $\widehat{\text{FDR}}(\Lambda)$ has an intuitive interpretation as the expected proportion of null cases among those with $z_i \geqslant z_0$. In addition, $\widehat{\text{FDR}}(\Lambda)$ is more robust than $\widetilde{\text{FDR}}(\Lambda)$ to the amount and distribution of the differentially expressed genes, as it only involves an empirical estimate of $F(z)$ instead of a density estimate of $f_1(z)$. By nature, $f_1(z)$ tends to have a complicated structure and its estimation is difficult.

Acknowledgement

The authors thank Professor D. M. Titterington, the associate editor and two referees for their constructive comments which have led to a significant improvement of this paper. Liang's research

was partially supported by grants from the U.S. National Science Foundation and the National Cancer Institute.

REFERENCES

ALLISON, D. B., GADBURY, G. L., HEO, M., FERNANDEZ, J. R., Lee, C. K., PROLLA, T. A. & WEINDRUCH, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comp. Statist. Data Anal.* **39**, 1–20.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

BENJAMINI, Y., KRIEGER, A. M. & YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

BENJAMINI, Y. & LIU, W. (1999). A step-down multiple hypothesis procedure that controls the false discovery rate under independence. *J. Statist. Plan. Infer.* **82**, 163–70.

BENJAMINI, Y. & YEKUTIELI, D. (2001). On the control of false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

BENJAMINI, Y. & YEKUTIELI, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters (with Discussion). *J. Am. Statist. Assoc.* **100**, 71–93.

BORDES, L., DELMAS, C. & VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statist.* **33**, 733–53.

CHEN, H. F. (1998). Stochastic approximation with non-additive measurement noise. *J. Appl. Prob.* **35**, 407–17.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

DO, K. A., MÜLLER, P. & TANG, F. (2005). A Bayesian mixture model for differential gene expression. *Appl. Statist.* **54**, 627–44.

DUDOIT, S., SHAFFER, J. P. & 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. Am. Statist. Assoc.* **99**, 96–104.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Assoc.* **102**, 93–103.

EFRON, B., TIBSHIRANI, R. J., STOREY, J. D. & TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.* **96**, 1151–60.

FURRER, R. (2002). M-estimation for dependent random variables. *Statist. Prob. Lett.* **57**, 337–41.

GENOVESE, C. & WASSERMAN, L. (2002). Operating characteristics and extension of the FDR procedure. *J. R. Statist. Soc.* B **64**, 499–517.

HASHEM, S. (1997). Optimal linear combinations of neural networks. *Neural Networks* **10**, 599–614.

HOLZMANN, H., MUNK, A. & GNEITING, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.* **33**, 753–63.

JOHNSON, M. E., TIETJEN, G. L. & BECKMAN, R. J. (1980). A new family of probability distributions with applications to Monte Carlo studies. *J. Am. Statist. Assoc.* **75**, 276–9.

KOOPERBERG, C. & STONE, C. J. (1992). Logspline density estimation for censored data. *J. Comp. Graph. Statist.* **1**, 301–28.

KUSHNER, H. J. (1981). Stochastic approximation with discontinuous dynamics and state dependent noise: w.p.1 and weak convergence. *J. Math. Anal. Appl.* **82**, 527–42.

LIANG, F., LIU, C. & WANG, N. (2007). A robust sequential Bayesian method for identification of differentially expressed genes. *Statist. Sinica* **17**, 571–97.

PAN, W., LIN, J. & LE, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integrat. Genom.* **3**, 117–24.

POUNDS, S. & CHENG, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics* **22**, 1979–87.

QIU, X., KLEBANOV, L. & YAKOVLEV, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statist. Applic. Genet. Molec. Biol.* **4**, article 34.

ROBBINS, H. & MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 400–7.

STONE, C. J., HANSEN, M., KOOPERBERG C. & TRUONG, Y. K. (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371–470.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.

STOREY, J. D., TAYLOR, J. E. & SIEGMUND, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Statist. Soc.* B **66**, 187–205.

TADIĆ, V. (1997). On the convergence of stochastic iterative algorithms and their applications to machine learning. In *Proc. 36th Conf. Decis. Control*, pp. 2281–6, San Diego, CA.
WHITE, H. (1989). Learning in artificial neural networks. *Neural Comp.* **1**, 425–64.
WOLPERT, D. H. (1992). Stacked generalization. *Neural Networks* **5**, 241–59.
YIN, G. & ZHU, Y. M. (1989). Almost sure convergence of stochastic approximation algorithms with nonadditive noise. *Int. J. Contr.* **49**, 1361–76.

[*Received May* 2006. *Revised February* 2008]