BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 5 2010, pages 640–646
doi:10.1093/bioinformatics/btq001

Gene expression

Advance Access publication January 15, 2010

A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes

Anastasios Markitsis[1] and Yinglei Lai[2,*]

[1]Department of Statistics and [2]Department of Statistics and Biostatistics Center, The George Washington University, Washington D.C. 20052, USA

Associate Editor: David Rocke

ABSTRACT

Motivation: The proportion of non-differentially expressed genes (π_0) is an important quantity in microarray data analysis. Although many statistical methods have been proposed for its estimation, it is still necessary to develop more efficient methods.

Methods: Our approach for improving π_0 estimation is to modify an existing simple method by introducing artificial censoring to P-values. In a comprehensive simulation study and the applications to experimental datasets, we compare our method with eight existing estimation methods.

Results: The simulation study confirms that our method can clearly improve the estimation performance. Compared with the existing methods, our method can generally provide a relatively accurate estimate with relatively small variance. Using experimental microarray datasets, we also demonstrate that our method can generally provide satisfactory estimates in practice.

Availability: The R code is freely available at http://home.gwu.edu/~ylai/research/CBpi0/.

Contact: ylai@gwu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 25, 2008; revised on December 23, 2009; accepted on December 30, 2009

1 INTRODUCTION

Microarray technology is a powerful tool for studying complex diseases (Mootha *et al.*, 2003) and for assessing the effects of drugs (Salvatore *et al.*, 2008) at the molecular level. It is an experimental method by which thousands of genes can be printed on a small chip and their expression can be measured simultaneously (Lockhart *et al.*, 1996; Schena *et al.*, 1995). It can be used to detect changes in gene expression between normal and abnormal cells, which enables scientists to detect novel disease-related genes (Singh *et al.*, 2002). Many statistical methods have been developed for this purpose (Cui and Churchill, 2003). Although other advanced genomics technologies, such as RNA sequencing (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008), have been developed, microarrays have been continuously used for broad biomedical studies (Cancer Genome Atlas Research Network, 2008). Furthermore, since the structures of data from different genomics technologies are basically similar, methods for analyzing microarray data can also be useful for analyzing other similar genomics data.

Performing statistical tests for a large number of genes raises the need for an adjustment for multiple hypothesis testing (MHT). A widely used method to address this issue is the false discovery rate (FDR; Benjamini and Hochberg, 1995) that evaluates the proportion of false positives among claimed positives. FDR control is less stringent than the traditional family-wise error rate (FWER) control such as the Bonferroni correction, and provides more power for discovering differentially expressed genes. However, estimating FDR involves the estimation of π_0, the proportion of non-differentially expressed (null) genes [$(1-\pi_0)$ corresponds to the proportion of differentially expressed genes]. A reliable estimate of π_0 is also of great importance to the sample size calculation for microarray experiment design (Jung, 2005; Wang and Chen, 2004).

A variety of methods have been proposed for estimating π_0. Storey and Tibshirani (2003) proposed *qvalue*. This method uses the ordered P-values and a cubic spline, and estimates π_0 as the value of the fitted spline at a value close to 1. Pounds and Morris (2003) suggested BUM, a 'beta-uniform' mixture model with the estimate of π_0 being the value of the fitted model at 1. *convest*, a method introduced by Langaas *et al.* (2005), utilizes a non-parametric convex decreasing density estimation method and gives the value of the density at 1 as an estimate of π_0. A histogram-based method has also been proposed (Mossig *et al.*, 2001; Nettleton *et al.*, 2006). The above methods usually provide conservative estimates of π_0; in other words, they are expected to give positively biased π_0 estimates. This has been considered an advantage, since it protects against overestimating the number of differentially expressed genes.

Many other methods have also been proposed for estimating π_0. Lai (2007) proposed a non-parametric moment-based method coupled with sample-splitting to achieve the identifiability and obtained a closed-form formula for π_0. Scheid and Spang (2004) presented the successive exclusion procedure (SEP), which successively excludes genes until the remaining u-values (transformed P-values) are sufficiently close to a uniform distribution $U[0,1]$. SEP estimates π_0 by J/m, where J is the estimated number of null genes, and m is the total number of genes. Guan *et al.* (2008) estimated the marginal density of P-values using a Bernstein polynomial density estimation, and gave a closed-form expression for their π_0 estimator. Liao *et al.* (2004) obtained an estimate of π_0 through Bayesian inference from a mixture model, which requires the distribution of P-values from non-null genes to be stochastically smaller than that from null genes. In addition to the above methods, there are still many other proposed methods

*To whom correspondence should be addressed.

for estimating π_0 (Broberg, 2005; Dalmasso *et al.*, 2005; Jiang and Doerge, 2008; Lu and Perkins, 2007; Pounds and Cheng, 2004, 2006). Furthermore, π_0 can also be estimated through a normal mixture model based on the z-scores obtained from P-values (McLachlan *et al.*, 2006).

In this study, to improve π_0 estimation, we propose a simple method, which is a modification of BUM. The novelty of our method is the introduction of artificial censoring to P-values so that an improved estimation can be achieved. Our motivation is based on the observation that a well-fitted BUM curve for the empirical P-value distribution may not be optimized for estimating π_0. In the following sections, we first introduce the statistical background and our method; then, we present the evaluation and comparison results from our simulation and application studies. Finally, we give some brief discussion to conclude our study.

2 METHODS

2.1 Detection of differential expression

In a typical microarray experiment, the gene expression in two groups of cells can be compared. On a microarray chip, a large number of genes can be monitored simultaneously, which provides researchers with measurement for each gene in each group. For example, to assess genes' involvement in tumor growth, the expression of tens of thousands of genes can be measured in normal and cancerous cells. Depending on the number of microarray chips available, multiple measurements for the expression of each gene are obtained.

For each gene, let μ_1 and μ_2 be the true mean intensities, in groups 1 and 2, respectively. To determine whether the gene is differentially expressed, the null and alternative hypotheses are:

$$H_0: \mu_1 = \mu_2 \text{ versus } H_a: \mu_1 \neq \mu_2.$$

A commonly used test statistic is the Student's t-test (assuming equal variances). A positive is claimed when H_0 is rejected in favor of H_a, and a negative when H_0 is not rejected. A positive means that the gene is declared differentially expressed; a negative means that the gene is declared non-differentially expressed.

If we knew the true state of each gene (i.e. whether it is truly differentially expressed or not), then the results of testing m genes simultaneously could be classified into four categories (each denoted by the random variable in parentheses): true positives (S), false positives (V), true negatives (U) and false negatives (T). Table 1 gives an illustration. Ideally, one would like to minimize V and T, and maximize S and U.

The probability $\Pr(V > 0)$ is called the FWER. In MHT, strong control is defined as maintaining the FWER below a specified level α. The traditional strong-control method is the Bonferroni procedure; that is, rejecting each H_0 corresponding to a P-value less than α/m. However, in microarray studies, α/m is typically so small that it is unlikely that many null hypotheses will be rejected. A widely used alternative is to control the FDR, the expected proportion of false positives (V) among the claimed positives ($R = V + S$) (Benjamini and Hochberg, 1995):

$$\text{FDR} = E(Q), \text{ where } Q = \frac{V}{R} \text{ when } R > 0, \text{ and } Q = 0 \text{ otherwise.}$$

Table 1. Numbers of true/false null hypotheses and negatives/positives in the situation of MHT

	True null	False null	Total
Negative	U	T	$m-R$
Positive	V	S	R
Total	m_0	$m-m_0$	m

Other versions of FDR have also been proposed: Tsai *et al.* (2003) considers the estimation of four other FDR versions. In general, controlling FDR provides higher statistical power for discovering differentially expressed genes. Let $m_0 = U + V$ denote the total number of true null hypotheses, and $\pi_0 = m_0/m$ denote the proportion of true null hypotheses (i.e. the proportion of non-differentially expressed genes; so the proportion of differentially expressed genes is $1-\pi_0$). Suppose that a researcher rejects H_0 for each gene with a P-value less than a prespecified level α. To estimate the corresponding FDR in this situation, Storey (2002) proposed

$$\widehat{\text{FDR}}(\alpha) = \frac{m\hat{\pi}_0\alpha}{r(\alpha)},$$

where $\hat{\pi}_0$ is an estimate of π_0, and $r(\alpha)$ is the observed number of positives. From this equation, it is clear that the accuracy of an FDR estimate depends on the estimation of π_0, which is the parameter of interest in this study.

2.2 The beta-uniform mixture model

Pounds and Morris (2003) have proposed the beta-uniform mixture (BUM) model. It assumes the following model for the marginal distribution of P-values:

$$f(p) = \gamma + (1-\gamma)\alpha p^{\alpha-1},$$

where $0 < p \leq 1$, $0 < \gamma < 1$ and $0 < \alpha < 1$.

Based on this simple model, Pounds and Morris (2003) have proposed the following estimate of π_0:

$$\hat{f}(1) = \hat{\gamma} + (1-\hat{\gamma})\hat{\alpha},$$

where $\hat{\gamma}$ and $\hat{\alpha}$ are the MLE estimates.

2.3 Our approach

To represent the marginal distribution of P-values, BUM uses a mixture of the uniform distribution $U[0,1]$ (also Beta(1,1)) and a Beta distribution Beta(α,1) with $0 < \alpha < 1$. However, BUM is too simplistic to achieve a robust performance in practice. Let $\mathbf{p} = \{p_1, p_2, \ldots, p_m\}$ be the observed P-values. Under the independence assumption, the log-likelihood is given by

$$L(\gamma, \alpha|\mathbf{p}) = \sum_{i=1}^{m} \log[f(p_i)] = \sum_{i=1}^{m} \log[\gamma + (1-\gamma)\alpha p_i^{\alpha-1}].$$

BUM estimates the fitted model curve by maximizing this log-likelihood. As $p \to 0, f(p) \to \infty$. Clearly, the smaller a p_i is, the larger its contribution will be to the log-likelihood. Therefore, to optimize the fitted curve, BUM places more weight on smaller P-values. However, π_0 is our focus and is estimated by $\hat{f}(1)$, which depends more on the P-values close to 1. To solve this problem, we propose the following censored beta mixture model.

2.3.1 A censored beta mixture model To improve BUM, we artificially censor the P-values that are less than a cut-off point λ. These P-values are considered 'indistinguishable'. In other words, even though the actual P-values less than λ are available, we do not use those values; our model only uses the number of such P-values. (We do not consider P-values $< \lambda$ as missing data). In this way, we aim to reduce the effect of very small P-values. Then, we have the mixture model:

$$f(p) = \gamma g_1(p) + (1-\gamma) g_2(p),$$

where

$$g_1 = \begin{cases} \text{censored} & 0 \leq p < \lambda \\ 1 & \lambda \leq p \leq 1 \end{cases}$$

is a left-censored uniform distribution $U[0,1]$ and,

$$g_2 = \begin{cases} \text{censored} & 0 \leq p < \lambda \\ \alpha p^{\alpha-1} & \lambda \leq p \leq 1 \end{cases}$$

is a left-censored Beta(α,1) distribution ($0 < \alpha < 1$). Figure 1 provides an illustration of this model.

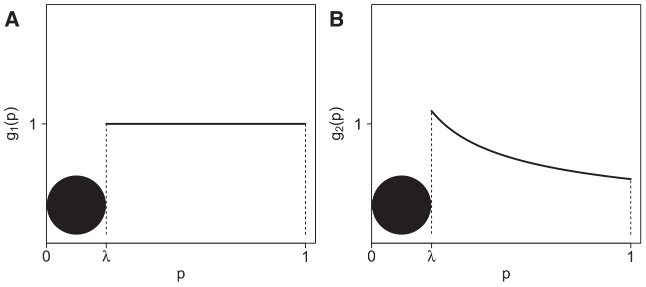

Fig. 1. (**A**) Graph of g_1, a censored uniform distribution $U[0,1]$. (**B**) Graph of g_2, a censored beta distribution Beta$(\alpha,1)$.

Remark 1. *Note that although we do not assume a specific form for the density of $f(p)$ in $[0,\lambda)$, we know that $Pr(0 \leq p < \lambda|g_1) = \lambda$ and $Pr(0 \leq p < \lambda|g_2) = \lambda^{\alpha}$. The marginal probability is $Pr(0 \leq p < \lambda) = \gamma\lambda + (1-\gamma)\lambda^{\alpha}$.*

Remark 2. *In this study, we assume that λ is given as 0.05, which is conventionally considered small (e.g. a threshold value for declaring statistical significance in practice). It is theoretically true that selecting a λ less than the minimum P-value is equivalent to using BUM. Furthermore, as pointed out by a reviewer, selecting a large λ is very similar to using qvalue or the histogram methods.*

2.3.2 Estimating model parameters Our model is a special case of the two-component mixture model in Ji *et al.* (2005). It consists of a censored Beta(1,1) (equivalent to $U[0,1]$), and a censored Beta$(\alpha,1)$. Therefore, we can use the Expectation–Maximization (EM) algorithm (McLachlan and Krishnan, 2008) to estimate the parameters γ and α. Following Ji *et al.* (2005), we augment the data by introducing the latent indicator variables z_i, $1 \leq i \leq m$ (where m is the total number of genes) defined as:

$$z_i = \begin{cases} 0 & \text{if } p_i \text{ belongs to the component } g_1, \\ 1 & \text{if } p_i \text{ belongs to the component } g_2. \end{cases}$$

Let $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$. The log-likelihood of our model given the 'complete' data $\{\mathbf{p}, \mathbf{z}\}$, is:

$$L(\gamma,\alpha|\mathbf{p},\mathbf{z}) = \log\left\{\prod_{i=1}^{m}[(\gamma g_1)^{1-z_i}((1-\gamma)g_2)^{z_i}]\right\}$$

To maximize the log-likelihood with respect to γ and α, given the 'complete' data, we take the partial derivative of the above equation with respect to γ and set it equal to zero, and then do the same for α. Solving these two equations, we obtain the following maximum likelihood estimates of γ and α to be used in the M-step of the EM algorithm:

$$\hat{\gamma} = \frac{\sum_{i=1}^{m}(1-z_i)}{m};$$

$$\hat{\alpha} = -\frac{\sum_{i:\lambda \leq p_i \leq 1} z_i}{\log(\lambda)\sum_{i:0 \leq p_i < \lambda} z_i + \sum_{i:\lambda \leq p_i \leq 1} z_i \log(p_i)}.$$

In the E-step of the EM algorithm, we need to update the expected values of the $\{z_i\}$. Given the current estimates of γ and α, we can compute $z_i^{\star} = \mathbf{E}(z_i|\mathbf{p},\hat{\gamma},\hat{\alpha})$. Since each z_i is an indicator variable, $z_i^{\star}$ is the conditional probability (at each iteration of the algorithm) that p_i belongs to component g_2. Hence, we have the following formulas:

- For each censored P-value, that is, if $0 \leq p_i < \lambda$,

$$z_i^{\star} = \frac{[(1-\hat{\gamma})\lambda^{\hat{\alpha}}]}{[\hat{\gamma}\lambda + (1-\hat{\gamma})\lambda^{\hat{\alpha}}]},$$

- For each non-censored P-value, that is, if $\lambda \leq p_i \leq 1$,

$$z_i^{\star} = \frac{[(1-\hat{\gamma})\hat{\alpha}p_i^{\hat{\alpha}-1}]}{[\hat{\gamma} + (1-\hat{\gamma})\hat{\alpha}p_i^{\hat{\alpha}-1}]},$$

To start the EM algorithm, we select an initial value for γ; in general, we can use $\gamma^{(0)} = 0.5$, unless we have some empirical estimate of π_0 to use instead. Then, we initialize $\{z_i\}$ by setting $z_i^{(0)} = 1 - \gamma^{(0)}$ for $1 \leq i \leq m$. With $\{z_i^{(0)}\}$, we can obtain $\gamma^{(1)}$ and $\alpha^{(1)}$, the estimates of γ and α after the first iteration. The convergence of EM algorithm is declared when

$$|\gamma^{(k)} - \gamma^{(k-1)}| < \delta,$$

where $\gamma^{(k)}$ is the estimate of γ at the end of the k-th iteration, and δ is a prespecified threshold ($\delta = 1 \times 10^{-6}$ in this study). When the EM algorithm converges, let $\hat{\gamma}$ and $\hat{\alpha}$ be the estimates of γ and α, respectively. Then, the estimate of π_0 is given by:

$$\hat{\pi}_0 = \hat{f}(1) = \hat{\gamma} + (1-\hat{\gamma})\hat{\alpha}.$$

Remark 3. *As suggested by a reviewer, it is necessary to consider multiple initial values for BUM. In our simulation study, for BUM's parameters (a and λ), we use a = λ = min(2 × mean of all P-values, 0.9) and 5 pairs of randomly simulated numbers from $U[0,1]$. Although our method is robust to different initial values in this study, we still suggest that multiple initial values may be necessary to achieve a reliable estimate of π_0 in certain situations (e.g. $\pi_0 \approx 1$). Furthermore, although the required computing time of our method is much longer than that of BUM (and several of other methods), it is still affordable with a general computer.*

2.3.3 Confidence interval Since the above EM algorithm does not provide us with any closed formulas of estimates, it is difficult to derive the theoretical confidence interval (CI) for the estimated π_0. Therefore, we use the bootstrap procedure (Efron, 1979) to obtain a CI for π_0 (we set $B = 500$ for both application studies):

(1) Select a random sample of m P-values from $\{p_1, p_2, \ldots, p_m\}$ with replacement and equal probabilities;

(2) Apply the EM algorithm to the sample generated in Step 1 and obtain a resampling estimate of π_0;

(3) Repeat Steps 1 and 2 B times to obtain the resampling distribution of $\hat{\pi}_0$;

(4) For a $100(1-\alpha)\%$ CI for π_0, find the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of the resampling distribution.

Remark 4. *A key assumption for bootstrapping P-values in the construction of CIs is that the observed P-values are independent. However, since genes are correlated in a expression dataset, a bootstrapped CI for π_0 should be considered as an approximation in practice. This issue has been discussed in Allison et al. (2002).*

3 RESULTS

3.1 Simulation studies

3.1.1 Simulation configuration We simulate gene expression data to evaluate the performance of our method. We also select several existing methods for a comparison study. BUM (Pounds and Morris, 2003) has to be included since it is the foundation of our method. Based on the consideration of the popularity and research history of the statistical methods for estimating π_0, the following methods are selected (notation in parentheses): (CB) our method; (BUM) Pounds and Morris (2003); (H) the histogram-based method (Mosig *et al.*, 2001; Nettleton *et al.*, 2006); (Q) *qvalue* (Storey and Tibshirani, 2003); (L) the method proposed by Liao *et al.* (2004); (S) the method proposed by Scheid *et al.* (2004); (C) *convest* (Langaas *et al.* 2005); (RDM) the method proposed by Lai (2007); (G) the method proposed by Guan *et al.* (2008). The notations defined above are used in Figure 2. (We have actually performed a simulation study

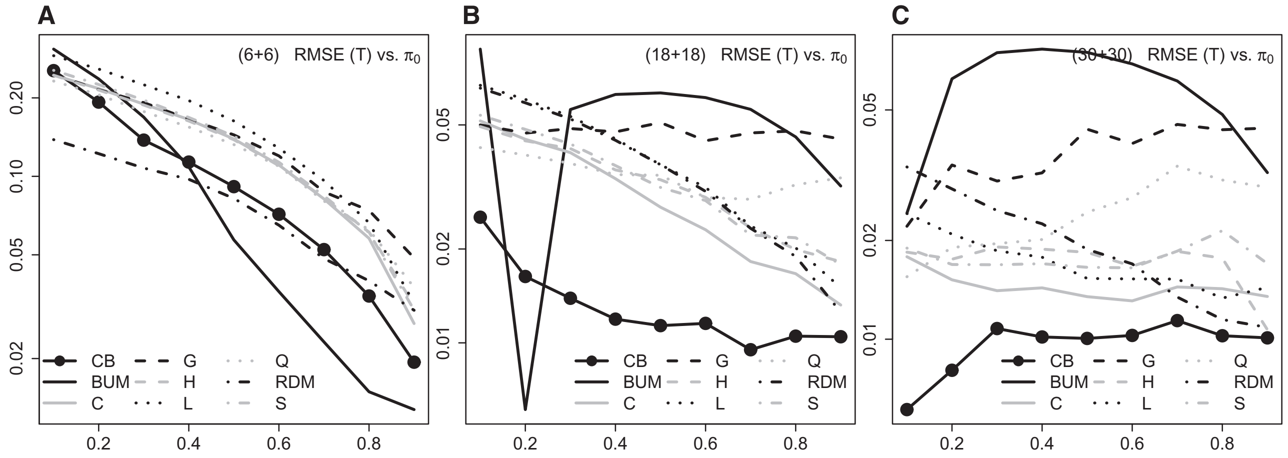

Fig. 2. Simulation results: gene expression data are simulated based on a independence structure. RMSE in log-scale of the estimates from different methods with different sample sizes considered: $n_1 = n_2 = 6$ (**A**), 18 (**B**) and 30 (**C**).

to compare many more methods. However, due to the page limit, it is difficult to present all the results. The exclusion of other methods does not change our conclusion.)

For each dataset, we simulate expression observations for 5000 genes. In reality, genes work together in complicated gene networks. To study the impact of correlation among genes on different methods, we generate data with the assumption that genes interact in blocks ('networks') of equal size. We also assume that within each block, the correlation among any pair of genes is the same, and equal to ρ. We perform simulations for different sample sizes ($n_1, n_2 = 6, 10, 18, 30$ and 50); correlation strength ($\rho = 0, 0.3, 0.5, 0.7$ and 0.9); and number of blocks ($b = 100, 200$ and 500) [or, equivalently, number of genes per block ($g_b = 50, 25$ and 10)].

Remark 5. *It is well-known that the sample size has an important impact on the estimation of π_0. As pointed out by one reviewer, the power of an α level test is the cumulative distribution function of the P-value evaluated at α. Since the power depends on the sample size, so does the distribution of P-values. Therefore, any non-trivial transformation of P-values (including π_0 estimators) depends on the sample size. For example, Pounds and Cheng (2005) have showed that when an estimate of the minimum of the assumed marginal distribution of P-values [e.g. $f(1)$ for BUM] is used for estimating π_0, the estimator can be expressed as a function of the sample size.*

Given a value of π_0 $(0.1, 0.2, \ldots, 0.9)$, a corresponding number of blocks are set to consist entirely of differentially expressed genes, with the remaining blocks consisting entirely of non-differentially expressed genes. For example, to generate a dataset with $\pi_0 = 0.7$ for the $\{b = 100, g_b = 50\}$ configuration, we simulate 30 blocks with differentially expressed genes, and 70 blocks with non-differentially expressed genes. For each block, we use the covariance matrix $\Sigma = (1-\rho)\mathbf{I} + \rho\mathbf{E}$ of size $g_b \times g_b$, where $\mathbf{I}$ is the identity matrix and $\mathbf{E}$ is a matrix of ones. (Note that Σ is also the correlation matrix since all genes have unit variances.) Then, for each configuration mentioned above, we perform the following:

(1) Simulate a gene expression dataset with 5000 genes.

(a) For a block of non-differentially expressed genes, generate observations from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ for both sample groups.

(b) For a block of differentially expressed genes, generate observations from a multivariate normal distribution $N(0, \Sigma)$ for one sample group. Then, generate observations from a multivariate normal distribution $N(\mu, \Sigma)$ for the other group (where μ is a random vector, with elements coming from a uniform distribution $U[0.5, 1.5]$).

(2) Apply the two-sample Student's t-test to the profile of each gene and obtain 5000 theoretical P-values.

(3) Use different methods to estimate π_0.

3.1.2 Criteria for evaluation and comparison We repeat the above steps $B = 100$ times for different values of π_0 $(0.1, 0.2, \ldots, 0.9)$. For each value of π_0 and each method, we compute the bias, standard deviation (SD) and root mean squared error (RMSE) as follows:

- Bias $= \sum_{i=1}^{B} (\hat{\pi}_{0_i} - \pi_0)/B$,
- SD $= \sqrt{\sum_{i=1}^{B} (\hat{\pi}_{0_i} - \sum_{i=1}^{B} \hat{\pi}_{0_i}/B)^2/(B-1)}$,
- RMSE $= \sqrt{\mathrm{Bias}^2 + \mathrm{SD}^2}$,

where $\hat{\pi}_{0_i}$ is the i-th estimate of π_0. These criteria are used to evaluate the estimation performance of different methods and the impact of different λ.

3.1.3 Comparison of different methods In all the results, the patterns in RMSE, bias and SD are very similar for all cases sharing the same sample size and correlation strength. In other words, the block size g_b in our configuration does not substantially affect the patterns in RMSE, bias and SD. In the Supplementary Materials, we present the simulation results based on 200 blocks with 25 genes in each block and different correlation values (ρ). In the following, we discuss the simulation results based on the simple independence structure ($\rho = 0$), which is representative of the other results. The simulation results are presented for samples sizes $6+6$, 18+18 and 30+30. [In order to show a clear comparison among different methods, we use a log-scale for the *y*-axis in RMSE and SD graphs (with the option 'log = "y" ' in the R-function 'plot'), and the cube root of Bias is actually used as the *y*-axis in the Bias graphs. All these comparison plots are given in the Supplementary Materials. However, in the following, we only give the RMSE-based comparison plots due to the page limit.]

When $n_1 = n_2 = 6$ (Fig. 2A), all the π_0 estimation methods show an overall decreasing pattern of RMSE as the value of π_0 increases. BUM gives the lowest RMSE when $\pi_0 > 0.4$; but its RMSE is among the worst when $\pi_0 < 0.3$, where RDM gives the lowest RMSE. Our method always gives a competitive low RMSE when $\pi_0 > 0.1$.

When the sample size increases to $n_1 = n_2 = 18$ (Fig. 2B), BUM shows an unstable performance: it gives a relatively high RMSE for all the values of π_0, except at $\pi_0 = 0.2$ and 0.9. The benefit of using our method is more apparent: its RMSE is lower than those of all the other methods, for all the values of π_0 except for $\pi_0 = 0.2$ (where BUM's RMSE is the lowest).

For $n_1 = n_2 = 30$ (Fig. 2C), BUM's RMSE displays a concave parabola pattern, and is always relatively high. Our method has the lowest RMSE for all π_0.

The figures in the Supplementary Materials also confirm a satisfactory performance in Bias and SD from our method. In general, most methods' bias decreases as the sample sizes and the value of π_0 increase. However, BUM quickly becomes the most negatively biased (which explains the observed large RMSE of BUM although the SD of BUM is among the smallest). A strongly negative bias leads to an undesirable overestimation of the number of truly differentially expressed genes. On the other hand, our method's bias becomes negligible as the sample sizes increase. Most methods' SD increases as the value of π_0 increases and decreases as the sample sizes increase.

In general, when the simulation results based on different dependent structure (independent, weakly/strongly dependent) are compared (Supplementary Materials), the higher the correlation, the higher becomes the SD. (The bias, on the other hand, remains mostly unaffected by the increase in correlation.) However, our results show that the increase in SD induced by positive correlation among test statistics does not render the existing π_0 estimation methods inappropriate.

3.1.4 Choice of λ The above reported simulation configuration can also be used to understand the effect of λ. We simulate data with different sample sizes 6+6, 18+18 and 30+30 and compare the performance of our model for λ in the set $\{0.01, 0.03, 0.05, \ldots, 0.25\}$. The figures in the Supplementary Materials shows that no single value of λ can be identified to minimize RMSE in a wide range of π_0. Furthermore, the RMSE patterns can change significantly when the sample sizes are changed. It is clear that a relatively large λ (e.g. $\lambda = 0.25$) is not a good choice. However, a relatively small λ (e.g. $\lambda = 0.01$) is also not an appropriate choice. In our simulation study, we have observed that $\lambda = 0.05$ is always a reasonable choice to achieve an overall satisfactory performance.

3.2 Applications to experimental data

We first consider the following two published experimental microarray datasets for our applications. The first dataset contains 22 283 gene expression profiles from kidney biopsies of 19 kidney transplant subjects with cyclosporine-based immuno-suppression and 22 kidney transplant subjects with sirolimus-based immuno-suppression. The second dataset consists of 12 488 gene expression profiles from pancreatic T regulatory (three subjects) and T effector cells (five subjects). Both datasets are publicly available in the Gene Expression Omnibus (GEO) database (Barrett *et al.*, 2007) with accession numbers GSE1743 (Flechner *et al.*, 2004) for the first (renal) dataset and GSE1419 (Chen *et al.*, 2005) for the second (T cell) dataset.

Theoretical *P*-values based on the corresponding *t*-distributions are calculated for each dataset (two-sample Student's *t*-test is used for detecting differential expression). The true value of π_0 is unknown in the applications. Therefore, to compare different methods in each application, we obtain $B = 500$ bootstrap estimates of π_0 (see Section 2.3.3 for details) and construct a boxplot for the estimate from each method. Such a boxplot is useful to understand general CIs for an estimate.

Based on our simulation study, *convest* has consistently showed a relatively low RMSE. BUM should be considered since it is the foundation of our method. Therefore, for simplicity, we use boxplots to compare our method with BUM and *convest*. (The exclusion of other methods does not change our conclusion.) Figure 3 shows the *P*-value histograms and the estimates from these three different methods. For both datasets, the *P*-value distribution curves fitted by our method are close to the corresponding *P*-value histograms. Theoretically, π_0 cannot be higher than the marginal *P*-value distribution. This has been briefly discussed in one of our previous publications (Lai, 2007).

For the renal dataset, our method gives an estimate of π_0 0.256 with a relatively tight CI (95% CI: 0.252–0.261). *convest* gives a higher estimate 0.278 with a wider CI (95% CI: 0.263–0.293), whereas BUM gives the highest estimate 0.321 although a slightly tighter CI (95% CI: 0.318–0.325). Notice that only the estimate from our method is under the whole *P*-value histogram. For the T-cell dataset, our method gives an estimate of π_0 0.605 with relatively tight CI (95% CI: 0.586–0.626), whereas BUM gives a slightly lower estimate 0.598 and slightly tighter CI (95% CI: 0.583–0.616). Both estimates are under the whole *P*-value histogram. However, *convest* still gives a higher estimate 0.638 and wider CI (95% CI: 0.609–0.671).

We also use another experimental dataset to illustrate that our method (also BUM) does not always yield satisfactory estimation results. The third application is based on a dataset with 22 283 gene expression profiles from small airway tissues (five non-smokers versus six smokers). This dataset is also publicly available in GEO with accession number GSE3320 (Harvey *et al.*, 2007). The estimation results are also given in Figure 3. A clear 'bumped' shape can be observed in the *P*-value range [0.15, 0.35], which causes the problematic estimation results from our method and BUM (these beta distribution based models do not allow any 'bumped' shapes). Our fitted model curve is not close to the *P*-value histogram. Although the CI from our method (95% CI: 0.899–0.921) and BUM (95% CI: 0.897–0.935) are clearly tighter (the one from our method is the tightest) than that from *convest* (95% CI: 0.850–0.902), both estimates from our method (0.909) and BUM (0.907) are clearly higher than the right end portion of *P*-value histogram. *convest* provides a more reasonable estimate 0.878 for this application, although the difference among the estimates from different methods is quite small.

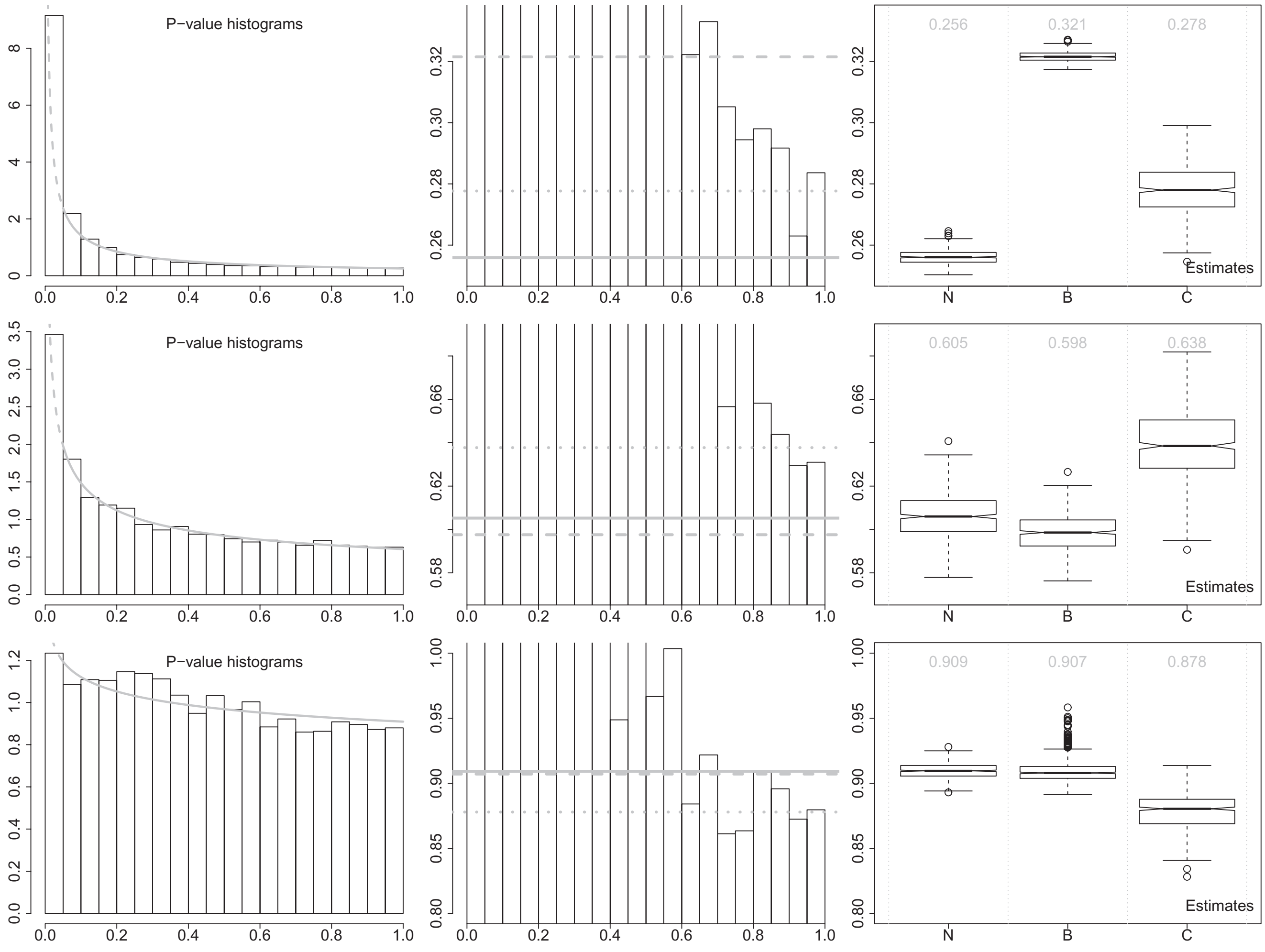

Fig. 3. Application results: histograms of P-values and boxplots of bootstrap estimates of π_0. The P-values are calculated based on three experimental datasets: the renal data (upper panel), the T-cell data (middle panel) and the smoke data (lower panel). In the histograms (left panel), the gray curves represent the fitted censored beta mixture models (the dashed parts are artificially censored). The zoomed-in histograms (middle panel) are also shown to compare the estimate of π_0 from different methods. The gray solid, dashed and dotted lines represent the estimates from our method, BUM and *convest*, respectively. In the boxplots (right panel), N = our method, B = BUM and C = *convest*. The numbers in gray color are the estimates of π_0 based on the original data.

Therefore, in practice, we suggest to check the histogram shape before applying any statistical methods for estimating π_0. If the histogram shape is roughly decreasing, then we expect satisfactory estimation performance from our method (and BUM in certain situations). If the histogram shape is not regular, then we may consider some non-parametric method like *convest* or the moment-based method (Lai, 2007).

4 DISCUSSION

Microarrays have been widely used in biological and medical studies. An accurate estimate of the proportion of differentially expressed genes is important in false positive control and experiment design. Therefore, the improvement of existing estimation methods still remains important.

Our proposed method for estimating π_0 provides an effective solution. Although it is arbitrary, the choice of $\lambda = 0.05$ provides an overall satisfactory performance. In our simulation study, the advantage of using our method is clear in the cases of moderate and large sample size ($18+18$ and $30+30$). In these cases, our method outperforms (w.r.t. RMSE) the other methods considered in this study. In the case of small sample, BUM has a satisfactory performance. Our method may be improved if an efficient method for the automatic selection of λ can be developed. This issue will be pursued in our future research.

Although none of the π_0 estimation methods mentioned above considers the effect of gene networks and interactions, dependence among genes is still a difficult issue in microarray data analysis (Efron, 2007). However, as investigated by Benjamini and Yekutieli (2001), methods that are based on the independence assumption perform quite well in general situations of weak positive dependence, and a positive dependency structure is common in many situations. A satisfactory performance under weak positive dependence has also been confirmed in our simulation studies.

ACKNOWLEDGEMENTS

We thank Professors Joseph Gastwirth and Qing Pan for their helpful comments and suggestions. We also thank the Associate Editor and anonymous reviewers for their insightful and helpful comments and suggestions.

Funding: National Institutes of Health (DK-075004 to Y.L.).

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al*. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, **39**, 1–20.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

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

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

Broberg,P. (2005) A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, **6**, 199.

Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Chen,Z. *et al*. (2005) Where CD4+CD25+ T reg cells impinge on autoimmune diabetes. *J. Exp. Med.*, **202**, 1387–1397.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Dalmasso,C. *et al*. (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Efron,B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

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

Flechner,S.M. *et al*. (2004) De novo kidney transplantation without use of calcineurin inhibitors preserves renal structure and function at two years. *Am. J. Transplant.*, **4**, 1776–1785.

Guan,Z. *et al*. (2008) Nonparametric estimator of false discovery rate based on Bernstein polynomials. *Stat. Sin.*, **18**, 905–923.

Harvey,B.G. *et al.* (2007) Modification of gene expression of the small airway epithelium in response to cigarette smoking. *J. Mol. Med.*, **85**, 39–53.

Jiang,H. and Doerge,R.W. (2008) Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Inform.*, **6**, 25–32.

Ji,Y. *et al*. (2005) Applications of beta-mixture models in bioinformatics. *Bioinformatics*, **2**1, 2118–2122.

Jung,S-H. (2005) Sample size for FDR-control in microarray data analysis. *Bioinformatics*, **21**, 3097–3104.

Lai,Y. (2007) A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data. *Biostatistics*, **8**, 744–755.

Langaas,M. *et al*. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B*, **67**, 555–572.

Liao,J.G. *et al*. (2004) A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**, 2694–2701.

Lockhart,D. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Lu,X. and Perkins,D.L. (2007) Re-sampling strategy to improve the estimation of number of null hypotheses in FDR control under strong correlation structures. *BMC Bioinformatics*, **18**, 157.

McLachlan,G.J. *et al*. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

McLachlan,G.J. and Krishnan,T. (2008) The EM algorithm and extensions, 2nd edn. John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 18–26.

Mootha,V.K. *et al*. (2003) PGC-1α-response genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Mosig,M.O. *et al*. (2001) A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, **157**, 1683–1698.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Nettleton,D. *et al*. (2006) Estimating the number of true null hypotheses from a histogram of *p* values. *J. Agric. Biol. Environ. Stat.*, **11**, 337–356.

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

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1745.

Pounds,S. and Cheng,C. (2005) Sample size determination for the false discovery rate. *Bioinformatics*, **21**, 4263–4271.

Pounds,S. and Cheng,C. (2006) Robust estimation of the false discovery rate. *Bioinformatics*, **22**, 1979–1987.

Salvatore,P. *et al*. (2008) Detrimental effects of Bartonella henselae are counteracted by L-arginine and nitric oxide in human endothelial progenitor cells. *Proc. Natl Acad. Sci. USA*, **105**, 9427–9432.

Scheid,S. and Spang,R. (2004) A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Trans. Comput. Biol. Bioinform.*, **1**, 98–108.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

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

Tsai,C-A. *et al.* (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

Wang,S-J. and Chen,J.J. (2004) Sample size for identifying differentially expressed genes in microarray experiments. *J. Comput. Biol.*, **11**, 714–726.

Wilhelm,B.T. *et al*. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.