Statistics and Its Interface Volume 4 (2011) 37–50

A likelihood ratio test for the proportion of non-differentially expressed genes

Anastasios Markitsis and Yinglei Lai*

The proportion of non-differentially expressed genes (π_0) is an important quantity in microarray data analysis. Although there is a wealthy literature about the estimation of π_0, the issue of hypothesis testing for π_0 has not been well addressed. In this study, we develop a likelihood ratio test for π_0 based on our recently proposed censored beta mixture model, and evaluate its power through a comprehensive simulation study. In order to understand the performance of our method for general experimental data, we simulate gene expression measurements based on a widely used data simulation scheme. The results confirm that a satisfactory power can still be achieved when there is a considerable sample size, a considerable number of genes, or a relatively large proportion of non-differentially expressed genes. Based on two experimental datasets, we illustrate that our method can be particularly useful for testing the hypothesis of no differentially expressed genes and calculating the sample size in an experimental design.

AMS 2000 subject classifications: Primary 62F03, 62F40; secondary 62F30.

Keywords and phrases: Proportion of true null hypothesis, Likelihood ratio test, Mixture model, Censored beta distribution, Power, Microarray gene expression data.

1. INTRODUCTION

With the introduction of microarray technology, large scale gene expression data have been collected for many biological and medical studies. Based on the expression data, a large number of statistical tests are performed to identify differentially expressed genes (Hendenfalk et al., 2001; Mootha et al., 2003). In microarray experiments, a parameter crucial both for controlling false positives and for calculating the appropriate sample size is the proportion of non-differentially expressed genes (π_0). Most papers regarding the proportion of true null hypotheses (π_0) focus on the estimation of π_0 (Pounds and Morris, 2003; Storey and Tibshirani, 2003; Liao et al., 2004; Scheid and Spang, 2004; Langaas et al., 2005; Nettleton et al., 2006; Lai, 2006; Lai, 2007; Guan et al., 2008; Markitsis and Lai, 2010).

To model the marginal distribution of p-values obtained in microarray data analysis, a mixture model is commonly used. [Although some non-parametric methods (e.g. Langaas et al., 2005) have been proposed for estimating π_0, it is not clear to us how to extend these methods for testing π_0.] In the general mixture-model literature there are many papers published to address the issue of testing for homogeneity (for example, Chen et al, 2001; Qin and Smith, 2004). Furthermore, Lo et al. (2001), and Lo (2005) have proposed likelihood ratio tests for the number of components in normal mixtures. However, the issue of a hypothesis test for π_0 has not been well addressed in the literature. To our best knowledge, only Xu and Liu (2008) have proposed a likelihood ratio test for the mixing proportion in a general two-component mixture model. However, the proposed test is based on a generalized pivotal quantity. Furthermore, the statistical inference about π_0 has its own unique features. In Markitsis and Lai (2010), we introduced a new method, which is a modification of the BUM method of Pounds and Morris (2003). The new method utilizes an artificially censored beta mixture model, and has demonstrated a better performance than most existing π_0 estimation methods in both our simulation studies and applications to experimental datasets. In this study, we introduce and discuss the properties of the likelihood ratio test for testing $H_0 : \pi_0 = k_0$ vs. $H_a : \pi_0 \neq k_0$ based on our recently proposed censored beta model (Markitsis and Lai, 2010).

Remark. The method proposed by Xu and Liu (2008) is a different approach that requires the specified forms of component distributions. However, a p-value distribution from a microarray data set can be complicated, and it is difficult to specify an accurate model for this distribution. Additionally, Xu and Liu (2008) proposed this method for estimating a general mixing proportion but not specifically the proportion of non-differentially expressed genes π_0. Furthermore, our method was developed based on the assumption $\pi_0 = f(1)$ as discussed by Langaas et al. (2005), which is another clear difference between our method and the method proposed by Xu and Liu (2008).

2. METHODS

2.1 π_0 and BUM

Suppose that a statistical test is performed to evaluate whether a gene is differentially expressed in two groups. Let μ_1 and μ_2 be the population mean expression levels of the

*Corresponding author.

Table 1. Numbers of true/false null hypotheses and negatives/positives in a typical situation of multiple hypothesis testing

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

gene in groups 1 and 2, respectively. Then, the null and alternative hypothesis are:

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

A positive occurs when H_0 is rejected in favor of H_a, and a negative when H_0 is not rejected.

Since the number of genes m studied in an experiment can be usually up to tens of thousands, and these genes are tested simultaneously, a multiple hypothesis testing setting arises, and four possible outcomes are possible (Benjamini and Hochberg, 1995). The outcomes in the case where m genes are simultaneously tested are shown in Table 1. Genes having a p-value less than a given threshold are declared significantly differentially expressed, and the quantity R is the number of these genes among the m genes being tested. The quantity m_0 is the number of genes that are truly non-differentially expressed. However, m_0 is unknown (and so are U, V, T, S) in general, and only R (and $m - R$) can be observed. The quantity

$$\pi_0 = m_0/m$$

is called the *proportion of true null hypotheses*, or the *proportion of non-differentially expressed genes*, which is important in sample size estimation (Wang and Chen, 2004; Jung, 2005) and false discovery rate (FDR) estimation (Storey and Tibshirani, 2003).

The estimation of π_0 is generally based on the observed m p-values and different models have been proposed for $f(p)$, the marginal distribution of p-values. Without any constraints imposed, π_0 is usually not identifiable due to a lack of degrees of freedom (although the p-value distribution under the null hypothesis generally follows a uniform distribution). Since the probability that a truly differentially expressed gene produces a large p-value should decrease as $p \to 1$, it is common to assume that $f(1) = \pi_0$ (or practically $f(1) \approx \pi_0$). Based on this assumption, a widely used conservative approach is to estimate π_0 by $\hat{f}(1)$ after an estimation of the marginal p-value distribution $f(p)$. Two representative methods based on this approach are the beta uniform model (BUM) proposed by Pounds and Morris (2003) and a nonparametric method proposed by Langaas et al. (2005). The simple BUM assumes:

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

with $0 < \alpha < 1$ and $0 < \gamma < 1$. Note that this is a mixture of a Uniform[0,1] and a Beta$(\alpha, 1)$ distribution. With $\psi = \text{logit}(\alpha)$ and $\phi = \text{logit}(\gamma)$, BUM uses a numerical optimization technique to find $\widehat{\psi}$ and $\widehat{\phi}$ that maximize the log-likelihood $l(\psi, \phi) = \sum \log[f(p|\alpha, \gamma)]$. The estimates of α and γ are $\widehat{\alpha} = \exp(\widehat{\psi})/(1 + \exp(\widehat{\psi}))$, and $\widehat{\gamma} = \exp(\widehat{\phi})/(1 + \exp(\widehat{\phi}))$. Then, the estimate of π_0 is given by

$$\widehat{\pi_0} = \hat{f}(1) = \widehat{\gamma} + (1 - \widehat{\gamma})\widehat{\alpha}.$$

Remark. Notice that the assumption $f(1) = \pi_0$ can also be considered as an upper bound of the true proportion. It has been long argued whether the term conservative estimation or simply the assumption $\pi_0 = f(1)$ should be used. [There was a short discussion about this issue in Langaas et al. (2005).] We eventually chose to use the assumption $\pi_0 = f(1)$ for its simplicity. Furthermore, this assumption is generally valid when the sample size is relatively large or the true value of π_0 is close to one.

2.2 A censored beta mixture model

To improve BUM, we proposed (Markitsis and Lai, 2010) to artifically censor the p-values that are less than a cut-off point λ. 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). 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 $U[0, 1]$ distribution, 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 $Beta(\alpha, 1)$ distribution ($0 < \alpha < 1$). 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 $\Pr(0 \leq p < \lambda) = \gamma\lambda + (1 - \gamma)\lambda^\alpha$. In this study, we set $\lambda = 0.05$, which is conventionally considered small (e.g., a threshold value for declaring statistical significance in practice), and has been shown to give a satisfactory performance in both simulation and application studies (Markitsis and Lai, 2010).

Remark. Notice that our purpose is to use a simple mixture model for the marginal distribution of p-values and achieve a satisfactory statistical inference of π_0 [which is assumed to be $f(1)$ in many published papers (e.g. Langaas et al., 2005) and also in this study]. Based on our experience in the beta-uniform mixture model, small p-values have a considerable impact on the estimation of π_0. This impact can be greatly reduced by the introduction of artificial censoring. This has also been demonstrated in our estimation study (Markitsis and Lai, 2010).

2.3 Estimating model parameters

We can use the Expectation-Maximization (EM) algorithm for a mixture model (McLachlan and Krishnan, 2008) to estimate the parameters γ and α. The latent indicator variables z_i, $1 \leq i \leq m$ (where m is the total number of genes) are 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:

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

With the introduction of $\mathbf{z}$, the E-step and M-step can be easily implemented and then the EM algorithm can be run iteratively. Let $\hat{\gamma}$ and $\hat{\alpha}$ be the MLE estimates of γ and α, respectively, returned by the EM algorithm. Then, the estimate of π_0 is given by:

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

More details about the estimation procedure and the related issues can be found in Markitsis and Lai (2010).

2.4 Test statistic

Suppose that we want to test

$$H_0 : \pi_0 = k_0 \text{ vs. } H_a : \pi_0 \neq k_0,$$

for $k_0 \in (0, 1)$. Under our censored beta model for the marginal distribution of p-values in a two-group microarray experiment, the likelihood is given by

$$\mathcal{L}(\gamma, \alpha|\mathbf{p}) = [\gamma\lambda + (1-\gamma)\lambda^{\alpha}]^{b_\lambda} \prod_{i:\lambda \leq p_i \leq 1} [\gamma + (1-\gamma)\alpha p_i^{\alpha-1}], \tag{1}$$

where $b_\lambda = \#\{i : 0 \leq p_i < \lambda\}$, $1 \leq i \leq m$, and $\mathbf{p} = \{p_1, p_2, \ldots, p_m\}$. Let $\boldsymbol{\theta}$ denote the parameter vector (γ, α). Since λ is fixed at 0.05, the likelihood ratio test statistic for testing H_0 vs. H_a is given by

$$\Lambda(\mathbf{p}) = [\sup_{\boldsymbol{\Theta_0}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{p})]/[\sup_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{p})], \tag{2}$$

where the parameter space $\boldsymbol{\Theta} = (0, 1) \times (0, 1)$, and $\boldsymbol{\Theta_0} = \{(\gamma, \alpha) : \gamma + (1-\gamma)\alpha = k_0, 0 < \alpha < 1, 0 < \gamma < 1\}$ with the details provided below.

2.5 Maximizing the restricted likelihood

Clearly, the likelihood in the denominator of equation (2) is maximized by the (γ, α) pair returned by the EM algorithm in the π_0 estimation procedure described in Section 2.3. To maximize the likelihood in the numerator, we note that under our model, $\pi_0 = f(1)$; under H_0, $\pi_0 = k_0$. Therefore, to find $\sup_{\boldsymbol{\Theta_0}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{p})$ we need to find the (γ, α) pair that maximizes the likelihood, subject to the constraint $f(1) = k_0$. This constraint is equivalent to

$$\gamma + (1-\gamma)\alpha = k_0. \tag{3}$$

To incorporate constraint (3) into an EM algorithm for maximizing the restricted likelihood, we use the method of Lagrange multipliers (see the supplementary materials for a general description of this method).

From Section 2.3, the log-likelihood of our model, given the z_i's is:

$$l = \sum_{i=1}^{m} (1-z_i)\log(\gamma g_1) + \sum_{i=1}^{m} z_i \log[(1-\gamma)g_2].$$

We define the Lagrangian

$$H(\gamma, \alpha, \eta) = l + \eta[\gamma + (1-\gamma)\alpha - k_0],$$

where η is the Lagrange multiplier. Then, we take the partial derivatives of $H(\gamma, \alpha, \eta)$ w.r.t. γ, α, and η, and set each of them equal to zero, obtaining three equations. After some algebra (see the supplementary materials for details), we arrive at the following cubic equation for α:

$$\begin{aligned} B_1\alpha^3 &+ [B_2 - A_1 - (k_0+1)B_1]\alpha^2 \\ &+ [A_1 + k_0 B_1 - (k_0+1)B_2 m(k_0-1)]\alpha + k_0 B_2 = 0, \end{aligned} \tag{4}$$

where $A_1 = \sum_{i=1}^{m} z_i$, $B_2 = \sum_{i:\lambda \leq p_i \leq 1} z_i$, and $B_1 = \log(\lambda)(\sum_{i:0 \leq p_i < \lambda} z_i) + (\sum_{i:\lambda \leq p_i \leq 1} z_i \log(p_i))$.

Using Cardano's method (see the supplementary materials for details), we solve for α. Letting $\tilde{\alpha}$ denote the solution, the estimate for γ is $\tilde{\gamma} = (k_0 - \tilde{\alpha})/(1 - \tilde{\alpha})$ [based on Equation (3)]. The pair $(\tilde{\gamma}, \tilde{\alpha})$ maximizes the likelihood given the z_i's, subject to constraint (3).

In the first iteration of the EM algorithm for maximizing the restricted likelihood, the z_i's are all set equal to 0.5, and equation (4) is solved to obtain $\tilde{\gamma}$ and $\tilde{\alpha}$. Then, $\mathbf{z}$ are updated using $\tilde{\gamma}$ and $\tilde{\alpha}$ in the EM algorithm mentioned in Section 2.3, and the process is iterated. At its convergence, the EM algorithm produces the final pair of $(\tilde{\gamma}, \tilde{\alpha})$, which maximize the restricted likelihood. In practice, the EM algorithm converges numerically when $\pi_0^{(k)}$ (the estimate of π_0 in the current iteration) is within a preset error threshold (say 1×10^{-6}) of $\pi_0^{(k-1)}$ (the estimate of π_0 in the previous iteration).

After the parameter estimation in $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta_0}$, the likelihood ratio test statistic is given by:

(5)

$$\Lambda(\mathbf{p}) = \left[\frac{\hat{\gamma}\lambda + (1-\hat{\gamma})\lambda^{\hat{\alpha}}}{\tilde{\gamma}\lambda + (1-\tilde{\gamma})\lambda^{\tilde{\alpha}}}\right]^{b_\lambda} \cdot \prod_{\lambda \le p_i \le 1} \left[\frac{\hat{\gamma} + (1-\hat{\gamma})\hat{\alpha}p_i^{\hat{\alpha}-1}}{\tilde{\gamma} + (1-\tilde{\gamma})\tilde{\alpha}p_i^{\tilde{\alpha}-1}}\right].$$

2.6 Computing the p-value for the test

In likelihood ratio tests in general, it is well-known that under regularity conditions, the test statistic $-2\log(\Lambda)$ under H_0 has an asymptotic χ^2_q distribution where q is the difference in dimensionality between $\mathbf{\Theta}$ and $\mathbf{\Theta_0}$ in Equation (2) (Lehmann and Romano, 2005). Clearly, in our likelihood ratio test for π_0, we have $q = \dim(\mathbf{\Theta}) - \dim(\mathbf{\Theta_0}) = 2-1 = 1$. However, in our situation, the regularity conditions do not hold since our censored beta model density is not differentiable for $0 \le p < \lambda$.

Therefore, to compute the p-value for the test, we use the parametric bootstrap method proposed by McLachlan (1987). The logic behind the parametric bootstrap is to create an estimate of the distribution of the test statistic under H_0. To do this, we assume that the data follow a censored beta distribution. Then, we estimate the censored beta model parameters from the observed data under H_0, and we generate a large number of samples from the censored beta model with those parameters. The set of the test statistic values computed from the generated samples serves as a sample of observations from the distribution of the test statistic under H_0.

The procedure is as follows:

1. Compute $\tilde{\gamma}$ and $\tilde{\alpha}$ (Section 2.5), the estimates of γ and α under H_0;
2. Compute $\hat{\gamma}$ and $\hat{\alpha}$ (Section 2.3), the estimates of γ and α under H_a;
3. Calculate the value of Λ [Equation (5)] and let $LR_o = -2\log(\Lambda)$;
4. Generate B parametric bootstrap samples by repeating the procedure below B times.
    - Generate an observation from the $Binomial(m, \tilde{\gamma})$; call this $\tilde{m}_0$.
    - Generate $\tilde{m}_0$ observations from the $Uniform[0, 1]$ distribution. These represent p-values from the g_1 component.
    - Generate $\tilde{m}_1$ observations ($\tilde{m}_1 = m - \tilde{m}_0$) from the $Beta(\tilde{\alpha}, 1)$ distribution. These represent p-values from the g_2 component.
5. For each of the B samples, compute the test statistic value Λ_b, $1 \le b \le B$. Let LR_b denote the value of $-2\log(\Lambda_b)$ for the b-th sample.
6. The p-value for the test is given by

$$(6) \qquad p\text{-value} = [\#\{b : LR_b \ge LR_o\} + 1]/(B+1).$$

Notice that the observed LR_o itself is included as an observation from the null distribution, along with the LR_b's from the bootstrap samples. This explains why we have "+1" in both the numerator and denominator of equation (6).

Remark. Based on the requirement of parametric bootstrap (McLachlan, 1987), the simulation configuration (in the parametric bootstrap procedure) should be consistent with the estimation under the null hypothesis. From the modified EM algorithm, the estimate of the mixing proportion for the g_1 component (under the null hypothesis) is $\tilde{\gamma}$. (Note that we do not define g_1 and g_2 as the p-value distributions of nondifferentially and differentially expressed genes, respectively. We actually use the mixture of g_1 and g_2 to model the marginal distribution of p-values.) Therefore, $\tilde{m_0}$, the number of observations from the g_1 component, should be generated from the binomial distribution $Binomial(m, \tilde{\gamma})$. This configuration is also confirmed by our additional simulation result presented in the supplementary materials. When data are generated exactly from the beta-uniform model, the p-values computed based on the null hypothesis scenario follow a uniform distribution.

2.7 Testing $H_0 : \pi_0 = 1$ vs. $H_a : \pi_0 \neq 1$

In the case of the test $H_0 : \pi_0 = 1$ vs. $H_a : \pi_0 \neq 1$, under H_0 we have

$$f(1) = \gamma + (1-\gamma)\alpha = 1.$$

The equation $\gamma + (1-\gamma)\alpha = 1$ can be re-written as $\gamma(1-\alpha) - (1-\alpha) = 0$, or, $(1-\alpha)(\gamma-1) = 0$. The latter equation implies that either γ or α (or both) are equal to 1. Hence, the likelihood equation (1) reduces to

$$\mathcal{L}(\gamma, \alpha|\mathbf{p}) = (\lambda)^{b_\lambda},$$

and equation (5) simplifies to

$$\Lambda(\mathbf{p}) = \left[\frac{\hat{\gamma}\lambda + (1-\hat{\gamma})\lambda^{\hat{\alpha}}}{\lambda}\right]^{b_\lambda} \prod_{\lambda \le p_i \le 1} \left[\hat{\gamma} + (1-\hat{\gamma})\hat{\alpha}p_i^{\hat{\alpha}-1}\right].$$

Therefore, it is clear that Step 1 in the previous p-value calculation procedure can be omitted in this special testing scenario.

Remark. Notice that testing $\pi_0 = 1$ is also equivalent to testing a simple uniform p-value distribution (for all genes) in our study (The one-sample Kolmogorov-Smirnov test can also be considered in this situation). For some microarray data sets (e.g. our first application), the proportion of differentially expressed genes can be small and then π_0 is close to one. Then, testing $\pi_0 = 1$ can be biologically important. An application to such an experimental microarray data set has been presented later for an illustration.

2.8 Power evaluation

The power evaluation with data generated exactly from the beta-uniform models is presented in the supplementary materials. Here, we consider a simulation scenario in which data are not generated from a beta-uniform model. This is intentionally performed so that the robustness of our test statistic can be evaluated.

To evaluate the power of the test

$$H_0 : \pi_0 = k_0 \text{ vs. } H_a : \pi_0 \neq k_0,$$

when the true value of π_0 is c, we first generate K datasets where $\pi_0 = c$, using the gene expression data simulation procedure below. For each of the K datasets, we compute the p-value, as described in Section 2.6. Let p_k be the p-value obtained for the k-th dataset, $1 \leq k \leq K$. An estimate of the power of the test $H_0 : \pi_0 = k_0$ vs. $H_a : \pi_0 \neq k_0$ at significance level α when the true value of π_0 is c is

$$\beta_\alpha = [\#\{k : p_k \leq \alpha\}]/K,$$

where $1 \leq k \leq K$. To obtain a power curve for a given value of k_0, we evaluate the power as described above, for each c (true π_0) alternative in the set $\{k_0 - 0.05, k_0 - 0.04, \ldots, k_0 + 0.04, k_0 + 0.05\}$. For example, for $k_0 = 0.6$, we evaluate the power for $c \in \{0.55, 0.54, \ldots, 0.64, 0.65\}$.

Procedure for Gene Expression Data Simulation

We simulate the expression measurements for m genes based on the widely used scheme below:

1. Generate m expression profiles for two sample groups with n observations per group as follows:
   (a) For non-differentially expressed genes ($m_0 = m \times c$), generate observations from the standard normal distribution $N(0, 1)$ for both sample groups.
   (b) For differentially expressed genes ($m - m_0$), generate observations from the standard normal distribution $N(0, 1)$ for the first sample group, and observations from a normal distribution $N(\mu, 1)$ for the second sample group. (For each differentially expressed gene, its μ is first randomly simulated from a uniform distribution $U[0.5, 1.5]$ and then fixed for the simulation of expression data from $N(\mu, 1)$.)
2. Use the two-sample *Student's* t-test (assuming equal variances) to obtain the p-values for the simulated m genes.

To consider the dependence structure among different genes, we can use a widely used block dependence structure for simulations (Allison et al., 2002; Langaas et al. (2005)). However, based on our additional simulation results presented in the supplementary materials, a satisfactory performance can still be achieved in the situation of general positive dependence structures. Therefore, for simplicity, we present the simulation results based on the independence structure in this study. Notice that the simulated data are not from a censored beta distribution; this has been intentionally done in order to understand the performance of our method when the underlying population is different from our proposed model.

Remark. The marginal p-value (or test statistic) distribution from a microarray data set can be complicated, and it is difficult to propose an accurate model for this distribution. Therefore, we prefer a simple model so that the test statistic for π_0 can be powerful. Based on the widely used block-dependent multivariate normal distributions, we conducted a simulation study to evaluate the robustness of our test. However, we also understand that the data distribution of a microarray data set is usually much more complicated. The simulation configuration used in our study is just a much simplified setting to evaluate the performance of our method when the true model is not a uniform-beta mixture.

3. RESULTS

To study the behavior of our likelihood ratio test for π_0, we generate data and evaluate the test power for different combinations of sample size ($n = 10 + 10$, $15 + 15$ and $20 + 20$) and number of genes ($m = 2,500$, $5,000$, and $10,000$), at different values of k_0.

3.1 Effect of sample size

We first investigate the effect of sample size in the case of the test $H_0 : \pi_0 = 0.6$ vs. $H_a : \pi_0 \neq 0.6$ for $c \in \{0.55, 0.56, \ldots, 0.64, 0.65\}$ (i.e., 11 different cases of true π_0). Using the procedure above, for each value of c, we simulate data for $K = 300$ datasets with $m = 5000$, using sample sizes $n = 10 + 10$, and compute the power as described above. Hence, we obtain power curves for significance levels $\alpha = 0.01, 0.05$, and 0.10. We then repeat this for sample sizes $15 + 15$, and $20 + 20$. (In all cases, to compute p-values, we generate $B = 300$ parametric bootstrap samples). The same procedure is used to obtain results for the test $H_0 : \pi_0 = 0.7$ vs. $H_a : \pi_0 \neq 0.7$, for $c \in \{0.65, 0.66, \ldots, 0.74, 0.75\}$.

The striking feature in Figure 1 is that for $n = 10 + 10$ (first column), the power curves are not centered at $\pi_0 = 0.6$ (0.7). This can be explained as follows: Suppose we are testing $H_0 : \pi_0 = 0.6$ vs. $H_a : \pi_0 \neq 0.6$, and we are evaluating the power at $c = 0.6$. Recall that our method estimates π_0 by $\hat{f}(1)$, and is expected to give a conservative estimate for π_0; i.e., $\hat{f}(1) > \pi_0$. Therefore, each of the $K = 300$ datasets simulated using the "Procedure for Gene Expression Data Simulation" with $c = 0.6$ is expected to give a marginal p-value distribution with $f(1) > 0.6$. Consequently, for $c = 0.6$ the simulated datasets "behave" (under our censored beta model estimation) as having a π_0 value greater than 0.6. For some c slightly less than 0.6, the simulated datasets behave as having a true $\pi_0 = 0.6$. As a result, the test has the lowest power at some c slightly less than 0.6, and the power curve is horizontally shifted to the left. Now, recall that each simulated dataset consists of m p-values generated from two-sample t-tests. When the sample size is increased, in each simulated dataset the t-tests gain power and produce smaller p-values. Hence, in the marginal p-value histogram, density is shifted toward 0, and the minimum of the fitted

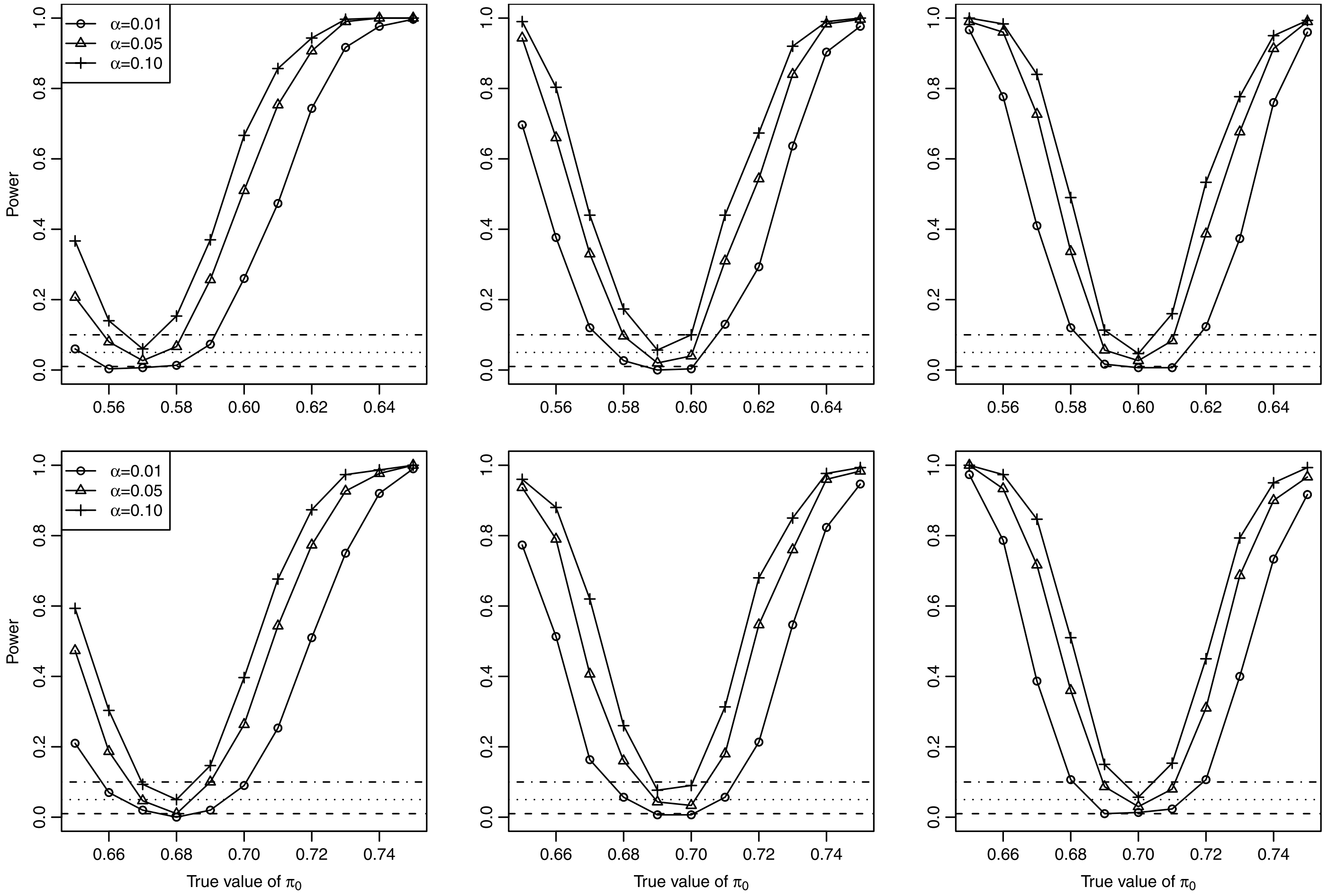

Figure 1. Effect of sample size (n) on power. First row: $k_0 = 0.6$. Second row: $k_0 = 0.7$. First column is for $n = 10 + 10$, the second for $n = 15 + 15$ and the third for $n = 20 + 20$. In all cases $m = 5,000$. The horizontal lines (".-", "..", and "- -") represent significance levels 0.10, 0.05, and 0.01, respectively.

curve is shifted toward the true value of π_0. Therefore, the positive bias is reduced and the power curve becomes centered at $c = 0.6$.

3.2 Effect of number of genes

In Figure 2, we observe that as the number of p-values (number of genes), m, increases (doubles), the power curves become visibly steeper. We can argue that as m increases, so does the power of the test, since the number of observations (p-values) has increased. Note (Figure 2, first row) that increasing the number of genes does not seem to affect the location of minimum of the power curves (unless the sample size is increased; Figure 2, second row), because the value of m does not have any impact on the minimum of the marginal p-value density in the simulated datasets.

3.3 Effect of k_0

In Figure 3, it seems that for larger values of k_0, the power curves become closer and closer to being centered at k_0. The power curves for $k_0 = 0.90$ are perfectly centered at 0.90. In our estimation study (Markitsis and Lai, 2010), we have observed that the bias of the π_0 estimate decreases as the true value of π_0 approaches 1. As explained above, this shifts the power curves toward $\pi_0 = k_0$.

3.4 On the shifts of power curves

In the results above, we have noted that due to the effect of sample size, a horizontal shift in the power curves occurs. Another type of shift present in all the results is a vertical shift. It is clear that, at the power curves' minimum, the power is lower than the nominal α level. This can be explained as follows: the parametric bootstrap samples used to obtain the p-value for the test are from a censored beta distribution, but the simulated dataset on which the test is performed is not. Therefore, the loss of power in our simulations was due to the fact that the parametric bootstrap samples are not from the same distribution as the dataset on which we are conducting the hypothesis test. We have

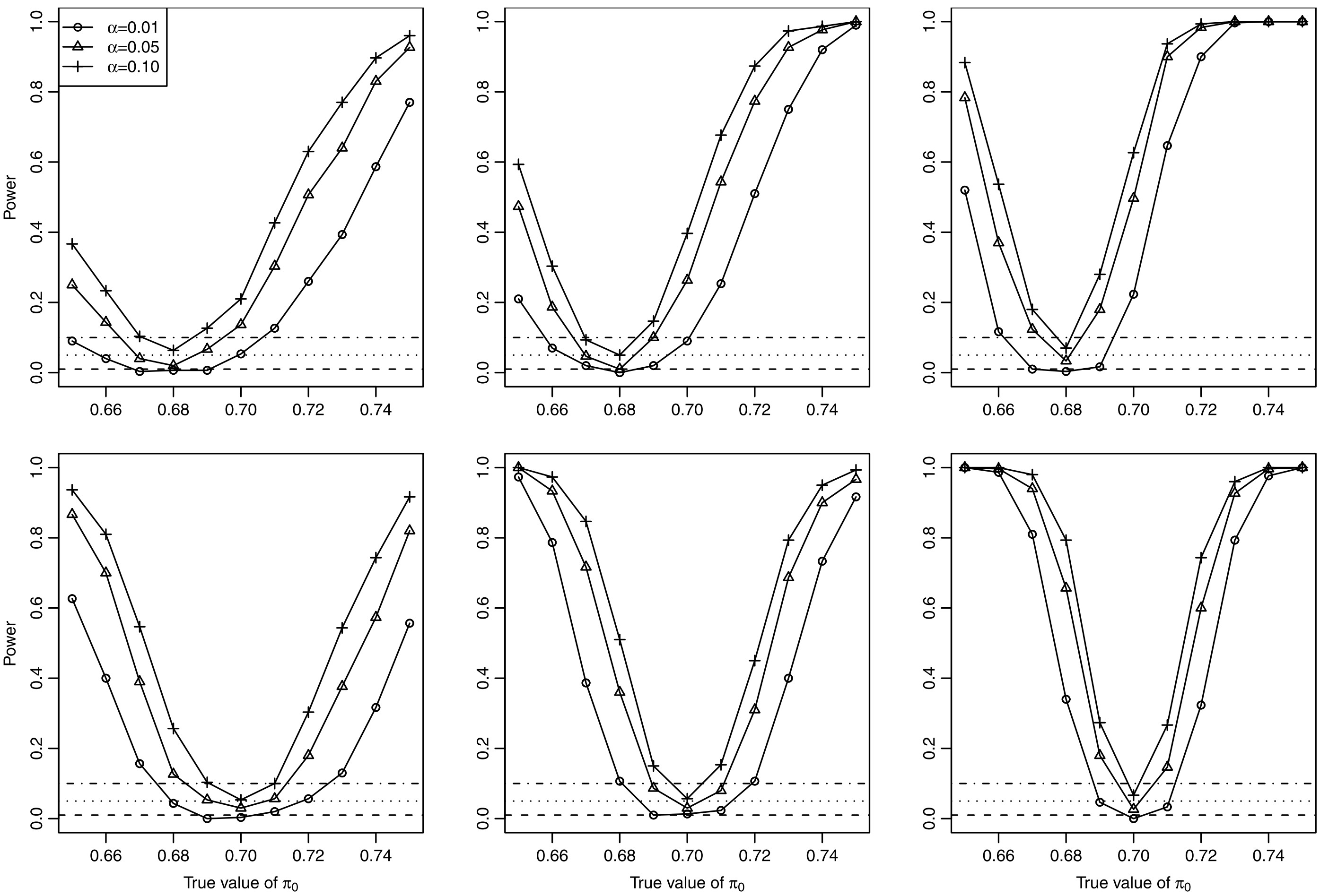

Figure 2. Effect of sample size (n) and number of genes (m) on power. Both rows are for $k_0 = 0.7$. First row: $n = 10 + 10$, and $m = 2,500; 5,000; 10,000$, for the first, the second and the third columns, respectively. Second row: $n = 20 + 20$, and $m = 2,500; 5,000; 10,000$, for the first, the second and the third columns, respectively. The horizontal lines (". -", ". .", and "- -") represent significance levels 0.10, 0.05, and 0.01, respectively.

run some simple simulations that confirm this conclusion (results given in the supplementary materials).

3.5 Application to experimental datasets

In practice, a hypothesis test for π_0 can be important in the specific case of testing whether $\pi_0 = 1$ (i.e., whether all the genes in the experiment are non-differentially expressed). Furthermore, testing π_0 from a pilot study can help researchers to plan an appropriate sample size for a follow-up study. We apply our hypothesis testing procedure to two experimental datasets. For the first dataset, many methods give an estimate of π_0 close to one. It is therefore necessary to conduct the test of $H_0 : \pi_0 = 1$. For the second dataset, π_0 estimates from different methods lie around $0.6 - 0.7$, representing a more general situation for testing π_0. The histogram of the permutation p-values (Storey and Tibshirani, 2003) for each dataset are shown in Figure 4.

3.5.1 Type 2 diabetes data

The gene expression data (Mootha et al., 2003) are from 17 subjects with normal glucose tolerance (NGT) and 18 subjects with Type 2 diabetes (DM2). Expression measurements were collected for 22,283 different genes. As mentioned above, most methods estimate π_0 to be 1, which suggests that none of the 22,283 genes is differentially expressed in the two groups of subjects (NGT vs. DM2). We conduct the test H_0: $\pi_0 = 1$ using the method described in Sections 2.6 and 2.7. The point estimate for π_0 from our method is 0.998. Using $B = 500$ bootstrap samples, the p-value for the test $H_0 : \pi_0 = 1$ is 0.3952. Therefore, there is no strong evidence that any differentially genes exist.

Remark. The one-sample Kolmogorov-Smirnov test can be considered as an alternative. We used the R function "ks.test" and obtained the p-value 0.1476 [R gave a warning that ties exist among the 22,283 permutation p-values. After adding an observation from Normal $N(\mu = 10^{-5}, \sigma =$

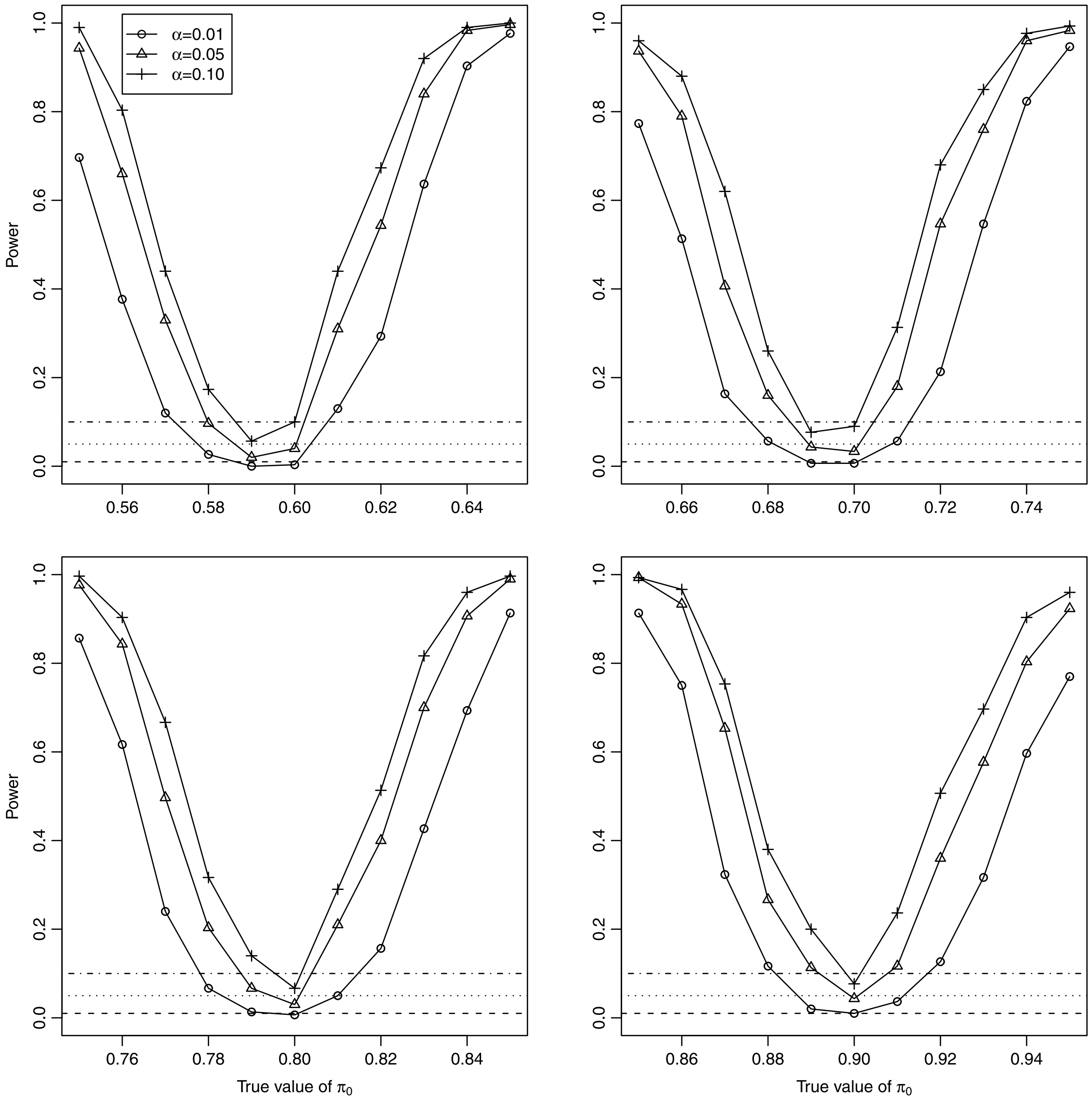

Figure 3. Effect of k_0 on power. Top row: $k_0 = 0.6, 0.7$ (first and second column, respectively). Bottom row: $k_0 = 0.8, 0.9$ (first and second column, respectively). For all graphs $n = 15 + 15$. The horizontal lines (". -", ". .", and "- -") represent significance levels 0.10, 0.05, and 0.01, respectively.

10^{-15}) to each permutation p-value to avoid ties, the p-value of the test was 0.1484.] This p-value does not contradict the p-value of 0.3952 from our test.

3.5.2 Breast cancer data

We use the microarray gene expression data collected by Hedenfalk et al. (2001) for a breast cancer study. The data were obtained from breast cancer patients with tumors involving mutation of either the BRCA1 or the BRCA2 gene. The data consist of expression measurements for 3,226 genes, with 7 samples (patients) for BRCA1 and 8 for BRCA2. The dataset can be accessed at http://research.nhgri.nih.gov/microarray/NEJM_Supplement. Since 56 genes exhibited expression measurements above 20 for one or more of the 15 patients, they were excluded, being regarded as unreliable (Storey and Tibshirani, 2003). Hence, 3170 genes remained in the study.

The point estimate of π_0 from our method is 0.632. Regarding the given data as a pilot study, we conduct the test $H_0 : \pi_0 = k_0$ vs. $H_a : \pi_0 \neq k_0$, for $k_0 = 0.58, 0.59, \ldots, 0.67, 0.68$, and obtain the corresponding p-values. Then, using the sample size computation method by Jung (2005), we calculate the sample size that would be required for a follow-up study to discover 100 of the differentially expressed genes (i.e., $r_1 = 100$ true positives). We control the FDR at $f = 0.01$, and since $n_1/(n_1 + n_2) \approx 0.5$, we set a_1 and a_2 in Jung's formula both equal to 0.5. To estimate the overall effect size δ in Jung's formula, we first transform the p-values to z-scores; i.e., compute $z = \Phi^{-1}(1 - p)$ for each p-value, where Φ^{-1} is the inverse cumulative distribu-

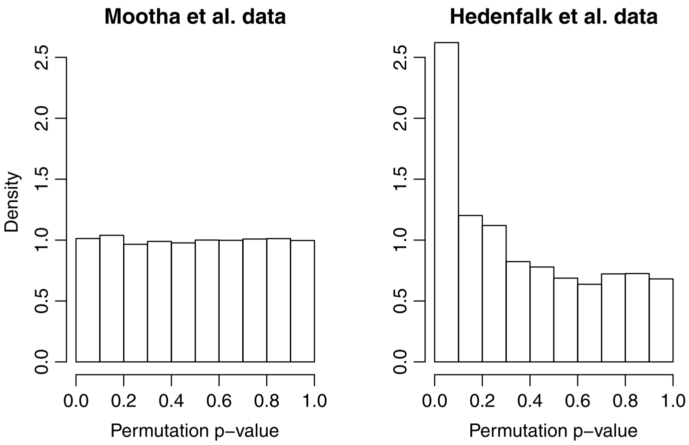

Figure 4. Histograms of permutation p-values for the type 2 diabetes data (left) and the breast cancer data (right).

tion function (inverse c.d.f.) of the standard normal distribution. Then, since the average z-score for the p-values from null genes will be zero, the average of all the z-scores will be approximately $(1-\pi_0)\delta/(n_1^{-1}+n_2^{-1})^{1/2}$. Using the average of the observed z-scores and the estimated value of π_0, we can obtain an estimate of δ. Figure 5 shows the graph of the p-values for the tests H_0: $\pi_0 = k_0$ vs. H_a: $\pi_0 \neq k_0$ ($k_0 \in \{0.58, 0.59, \ldots, 0.68\}$), with the required sample size $(n_1 + n_2)$ printed in the plot. The horizontal line represents the significance level 0.05. The test clearly rejects the null hypothesis of $\pi_0 = 0.58, 0.59$, or 0.68. Therefore, there is strong evidence that π_0 is not equal to one of them. For $\pi_0 = 0.67$, the p-value is 0.05, which equals to the significance level. Therefore, there is no strong evidence that π_0 is not equal to 0.67. For a conservative strategy, we would recommend $n_1 = n_2 = 20$ for a follow-up study.

4. DISCUSSION

The issue of a hypothesis test for the proportion of non-differentially expressed genes, π_0, which is an important parameter in the analysis of microarray data, has not been well addressed in the literature. In general, the major difficulty for hypothesis testing in mixture models is the fact that the behavior of the likelihood ratio test statistic is usually unknown. In this article, we have developed a likelihood ratio test for π_0 based on a parametric bootstrap procedure (McLachlan, 1987). In a comprehensive simulation study, we have evaluated the effects of sample size, number of genes, and null hypothesis value on the power of our test. Overall, our test has shown a satisfactory performance. We have applied our method for testing whether $\pi_0 = 1$ in the type 2 diabetes gene expression data (Mootha et al., 2003), and found no strong evidence that any differentially expressed genes exist. Through an application to the breast cancer gene expression data (Hedenfalk et al., 2001), we have also demonstrated the usefulness of our method for selecting the appropriate sample size for a follow-up study.

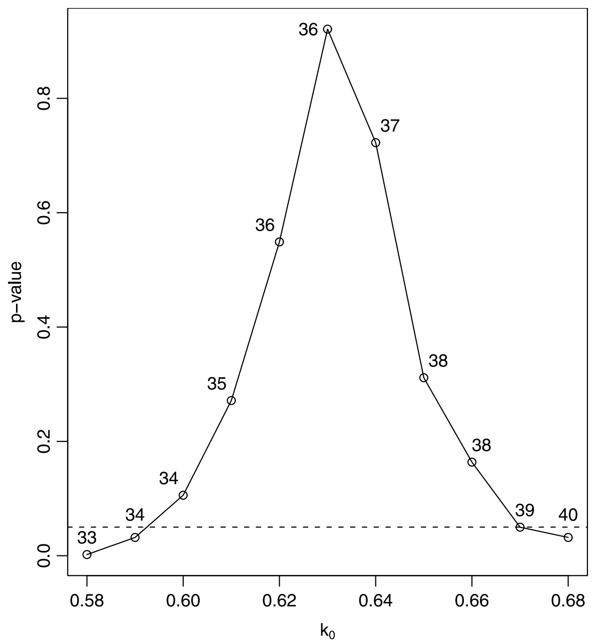

Figure 5. Results for the breast cancer data: p-values for testing $H_0 : \pi_0 = k_0$ for $k_0 \in \{0.58, 0.59, \ldots, 0.67, 0.68\}$ and estimated sample size for each case.

In our simulation study, we have observed that when the sample size is relatively small, the conservative (positively biased) π_0 estimate causes the power curves in our simulations to shift to the left of the null hypothesis value of π_0. However, as the sample size increases to relatively large, the power curves become centered at the null hypothesis value of π_0. Another observation from our simulation study is that the power curves usually have minimums below the nominal significance level. Both observations can be explained by the model misspecification, i.e., that our test is based on the assumption of a censored beta mixture model, but our simulated data are from a widely used data simulation scheme. In our future work, we will investigate possible modifications to our test for improving the power.

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. The R code and the related C programs are freely available at http://home.gwu.edu/~ylai/research/CBpi0/. Mathematical derivations and more simulation results are provided as supplementary materials. This work was supported by the National Institutes of Health [DK-075004 to Y.L.]

SUPPLEMENTARY MATERIAL

APPENDIX: MATHEMATICAL DERIVATIONS

Parameter estimation

The three equations are:

$$\begin{aligned}\frac{\partial H}{\partial \gamma} &= 0 \Rightarrow \frac{m - A_1}{\gamma} - \frac{A1}{1-\gamma} = -\eta(1-\alpha)\\ \frac{\partial H}{\partial \alpha} &= 0 \Rightarrow B_1 + \frac{B_2}{\alpha} = -\eta(1-\gamma)\\ \frac{\partial H}{\partial \eta} &= 0 \Rightarrow \gamma + (1-\gamma)\alpha = k_0.\end{aligned}$$

Multiplying the first equation by $\gamma(1-\gamma)$ (assuming $\gamma \neq 0$ or 1) gives:

$$(1-\gamma)(m - A_1) - \gamma A_1 = -\eta(1-\gamma)\gamma(1-\alpha).$$

Substituting the LHS of the second equation for $-\eta(1-\gamma)$, we have:

$$(1-\gamma)(m - A_1) - \gamma A_1 = \left(B_1 + \frac{B_2}{\alpha}\right)\gamma(1-\alpha). \tag{1}$$

Finally, by the third equation, $\gamma = \frac{k_0 - \alpha}{1-\alpha}$, which we substitute into Equation (1), and obtain the following equation [Equation (5) of the main article].

$$\begin{aligned}&B_1\alpha^3 + [B_2 - A_1 - (k_0+1)B_1]\alpha^2\\ &\quad + [A_1 + k_0B_1 - (k_0+1)B_2m(k_0-1)]\alpha + k_0B_2 = 0,\end{aligned}$$

To solve this cubic equation for α, we first divide it throughout by B_1, to obtain:

$$\alpha^3 + a_1\alpha^2 + b_1\alpha + c_1 = 0,$$

where $a_1 = (B_2 - A1)/B_1 - (k_0+1)$, $b_1 = [m(k_0-1) + A_1 - B_2(k_0+1)]/B_1 + k_0$, and $c_1 = k_0B_2/B_1$. Then we follow Cardano's method to solve for α.

Cardano's method for solving a cubic equation

Suppose we want to solve the equation

$$x^3 + ax^2 + bx + c = 0. \tag{2}$$

(Any cubic equation can be written in this form by dividing throughout by its x^3 coefficient.) First, to eliminate the quadratic term, we make the substitution $x = t - a/3$. Equation (2) then becomes:

$$t^3 + pt + q = 0, \tag{3}$$

where $p = b - a^2/3$ and $q = c + (2a^3 - 9ab)/27$. Now, write t as the sum of two new variables, u and v; i.e., let $t = u + v$. Substituting $u + v$ for t in equation (3) gives

$$u^3 + v^3 + (3uv + p)(u+v) + q = 0. \tag{4}$$

Since we expressed t in terms of two new variables, we need to impose an extra condition on u and v. Cardano suggested the condition

$$3uv + p = 0, \tag{5}$$

which simplifies equation (4) to

$$u^3 + v^3 + q = 0. \tag{6}$$

Solving $3uv + p = 0$ for v and substituting into equation (6) gives

$$u^6 + qu^3 - \frac{p^3}{27} = 0. \tag{7}$$

Equation (7) can be solved as a quadratic in u^3, producing

$$u^3 = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} \tag{8}$$

Choosing the "+" sign in equation (8), and substituting for u^3 into equation (6) yields

$$v^3 = -\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}.$$

Letting $D = q^2/4 + p^3/27$, we have

$$u = \begin{cases}\sqrt[3]{-\frac{q}{2} + \sqrt{D}}\\ (-\frac{1}{2} + i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} + \sqrt{D}}\\ (-\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} + \sqrt{D}},\end{cases}$$

$$v = \begin{cases}\sqrt[3]{-\frac{q}{2} - \sqrt{D}}\\ (-\frac{1}{2} + i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} - \sqrt{D}}\\ (-\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} - \sqrt{D}},\end{cases}$$

where $(-\frac{1}{2} \pm \frac{\sqrt{3}}{2}i)$ are the complex cubic roots of 1. Notice that choosing the "−" in equation (8) does not change the final values of t because t, as well as equations (6) and (5) are symmetric in u and v.

Now, u and v must satisfy equation (5), or, equivalently, $uv = -p/3$. Since p is real, in order to compute $t = u + v$, only certain combinations of the values of u and v are appropriate (1st value of u with 1st value of v; 2nd with 3rd; and 3rd with 2nd). Hence,

$$t = \begin{cases}\sqrt[3]{-\frac{q}{2} + \sqrt{D}} + \sqrt[3]{-\frac{q}{2} - \sqrt{D}}\\ (\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} + \sqrt{D}} + (\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} - \sqrt{D}}\\ (-\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} + \sqrt{D}} + (\frac{1}{2} - i\frac{\sqrt{3}}{2})\sqrt[3]{-\frac{q}{2} - \sqrt{D}},\end{cases}$$

and $x = t - a/3$.

Remarks.

- If $D > 0$, we have a real root and two (conjugate) complex roots.
- If $D = 0$, and $q = 0$ (which imply that also $p = 0$), then $u = v = 0$ and we have the triple root $x = -\frac{a}{3}$. (This case also occurs when $D = 0$ and $p = 0$, which imply that $q = 0$ as well.)
- If $D = 0$ and neither p nor q are 0, then we have three real roots: one single and one double root.
- If $D < 0$, then we have three distinct real roots.

Method of Lagrange multipliers

The method of Lagrange Multipliers is used for maximizing (or minimizing) a function of two or more variables subject to a constraint. Suppose we want to maximize the function $f(x, y)$ subject to the constraint $g(x, y) = k$. First note that $f(x, y)$ represents a surface in 3-D. The *contour lines* of $f(x, y)$ are the curves $f(x, y) = c$ on the xy-plane for different values of $c \in \mathbb{R}$. (The horizontal plane at height c cuts the surface $f(x, y)$ along a curve; the projection of this curve on the xy-plane is the contour line $f(x, y) = c$). Also note that the constraint $g(x, y) = k$ is a contour line of $g(x, y)$.

If we draw some of the contour lines of $f(x, y)$, the point (x_0, y_0) where one of them touches (but does not cross) the curve $g(x, y) = k$ will be the location of the maximum (if a maximum exists). At (x_0, y_0), the contour line of $f(x, y)$ and the contour line of $g(x, y)$ have parallel tangent vectors. Hence, their gradient vectors at (x_0, y_0) are also parallel. We can express this relationship as

$$\nabla f(x, y) = \eta \nabla g(x, y), \tag{9}$$

where η is called the Lagrange multiplier, and

$$\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right).$$

Solving Equation (9) is equivalent to setting the partial derivatives of the *Lagrangian*

$$H(x, y, \eta) = f(x, y) - \eta g(x, y)$$

equal to zero, and solving for x and y to obtain (x_0, y_0).

SIMULATION BASED ON THE CENSORED BETA MIXTURE MODEL

In this section, we present the power evaluation with data generated exactly from the beta-uniform models. The following simulation configuration was considered: $k_0 = 0.7$, $m = 6000$, $\alpha = 0.25$, and $\gamma = 0.5$, 0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.675 and 0.7 [corresponding to $\pi_0 = f(1) =$ 0.625, 0.64375, 0.6625, 0.68125, 0.7, 0.71875, 0.7375, 0.75625 and 0.775]. For each of 500 rounds of simulation, we first generated m_0 from the Binomial distribution $Binomial(m, \gamma)$ and then generated m_0 p-values from the uniform distribution $U[0, 1]$ and $m - m_0$ p-values from the beta distribution $Beta(\alpha, 1)$ (see below “Procedure for generating censored beta mixture data with $\pi_0 = c$”). We then computed the p-values based on $B = 500$ parametric bootstrap samples.

Procedure for generating censored beta mixture data with $\pi_0 = c$

1. Set $\alpha = 0.25$.
2. Let $\gamma = (c - \alpha)/(1 - \alpha)$, using the value of α from the previous step.
3. Let m_0 be an observation generated from a Binomial(m, γ) distribution. This will be the number of non-differentially expressed genes.
4. For p-values corresponding to differentially expressed genes, generate $m - m_0$ observations from a Beta$(\alpha, 1)$ distribution ($\alpha = 0.25$).
5. For p-values corresponding to non-differentially expressed genes, generate m_0 observations from a Uniform[0,1] distribution.

Notice that the value 0.25 is arbitrarily set since there are infinite number of pairs (γ, α) satisfying $\gamma + (1 - \gamma)\alpha = c$.

To address the issue of type I error control more clearly, we chose to present in Figure 1 the curves of (empirical) cumulative distribution functions (ECDF) based on the above simulation scenarios including the null hypothesis ($\pi_0 = 0.7$ or $\gamma = 0.6$). It is clear that the p-values simulated based on the null hypothesis scenario follow a uniform distribution. The two-sided one-sample Kolmogorov-Smirnov test gave a p-value 0.316. (Such p-values were also calculated for other k_0 and π_0 values in similar simulation scenarios. For example, when $k_0 = 0.55$ and 0.85, the p-values were 0.846 and 0.310 for $\pi_0 = 0.55$ and 0.85, respectively.)

SIMULATION BASED ON THE CORRELATED MULTIVARIATE NORMAL DISTRIBUTIONS

We found it difficult to develop a test statistic and/or implement the parametric bootstrap procedure with a dependence structure incorporated. However, the test statistic and the parametric bootstrap procedure proposed in this study should be applicable to a general gene expression data set. To confirm this, we include additional simulation study results so that we can evaluate the power of our test statistic when there is a dependence structure.

Based on our experience in the estimation study (Markitsis and Lai, 2010), the block-dependence structure is a reasonable and widely used one for simulating gene expression data (Allison et al., 2002; Langaas et al., 2005). Furthermore, the correlation within a block is usually not too strong. Therefore, we set 0.3 (for a modest dependence) or

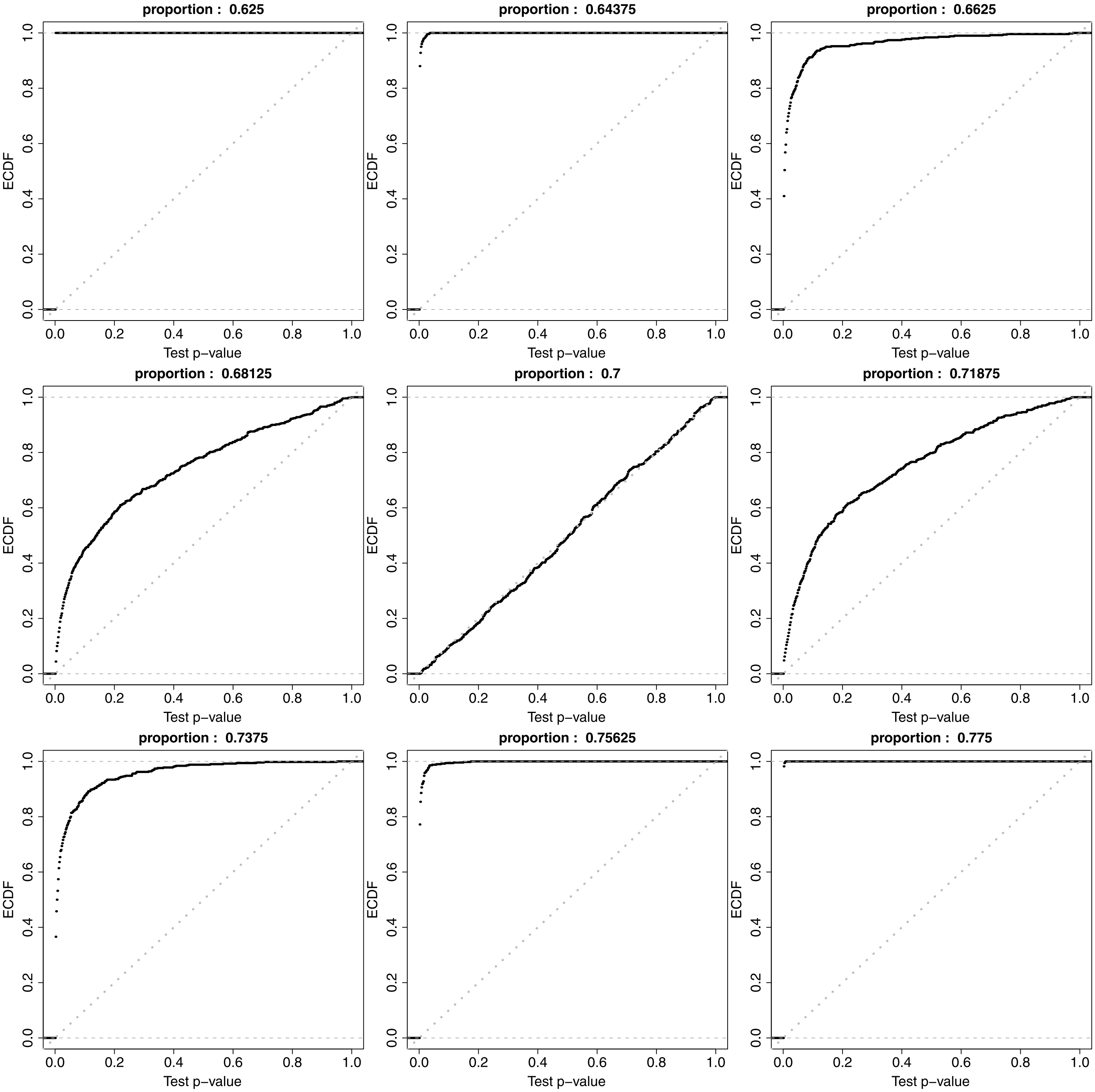

Figure 1. Test p-value distributions (ECDF) based on the data simulated from the beta-uniform models with: $\pi_0 = f(1) =$ *0.625, 0.64375, 0.6625, 0.68125, 0.7, 0.71875, 0.7375, 0.75625, 0.775, and* k_0 *= 0.7 (*γ *= 0.5, 0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.675, 0.7, and* $\alpha = 0.25$ *in all cases).*

0.5 (for a strong dependence) and also 0.0 (for a comparison) as the common correlation for each block with 25 genes (total 6000 genes). Different sample sizes (numbers of arrays) were considered: 10+10, 30+30 and 50+50, and the expression variance of each gene was still fixed at one (no other changes were made to the configuration described in the manuscript).

Based on Figure 2 below (the sample size is 10+10, the true proportion of non-differentially expressed genes is 0.75, i.e. $c = 0.75$, and the null hypothesis of π_0 is 0.7, i.e., $k_0 = 0.7$), there is visible power loss when the common block correlation is increased from 0.0 to 0.3 and 0.5. However, the overall power performance is still quite satisfactory. A similar trend has also been observed for the results based

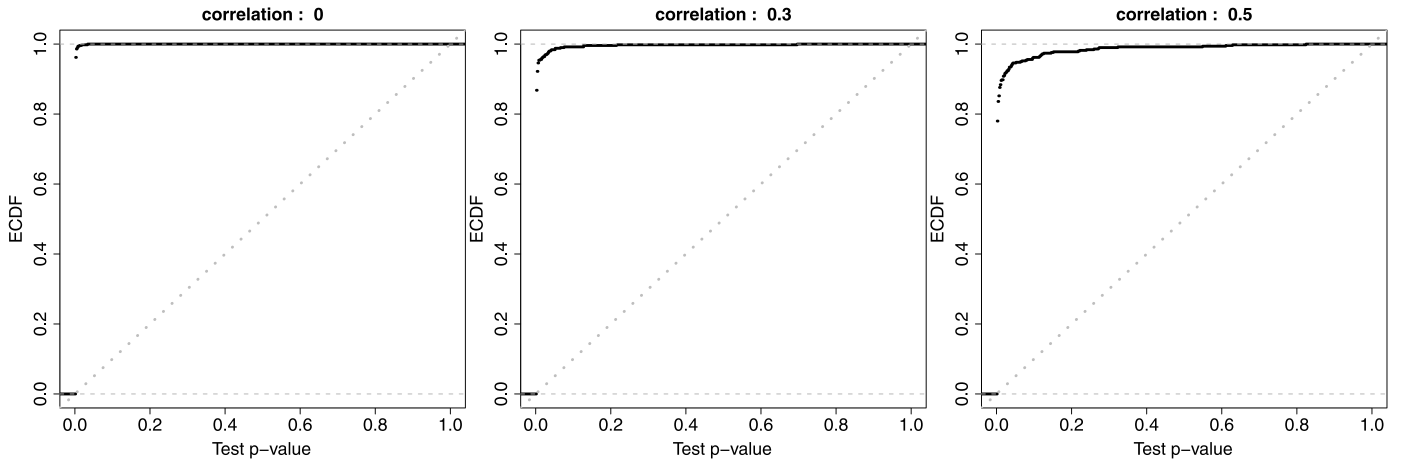

Figure 2. Comparison of the test p-value distributions (ECDF) based the data simulated from (left panel) independence structure with $\rho = 0$; (middle panel) dependence structure with $\rho = 0.3$; and, (right panel) dependence structure with ρ=0.5. In all cases, the true value of proportion of non-differentially expressed genes is c = 0.75, and the null hypothesis of π_0 is k_0=0.7.

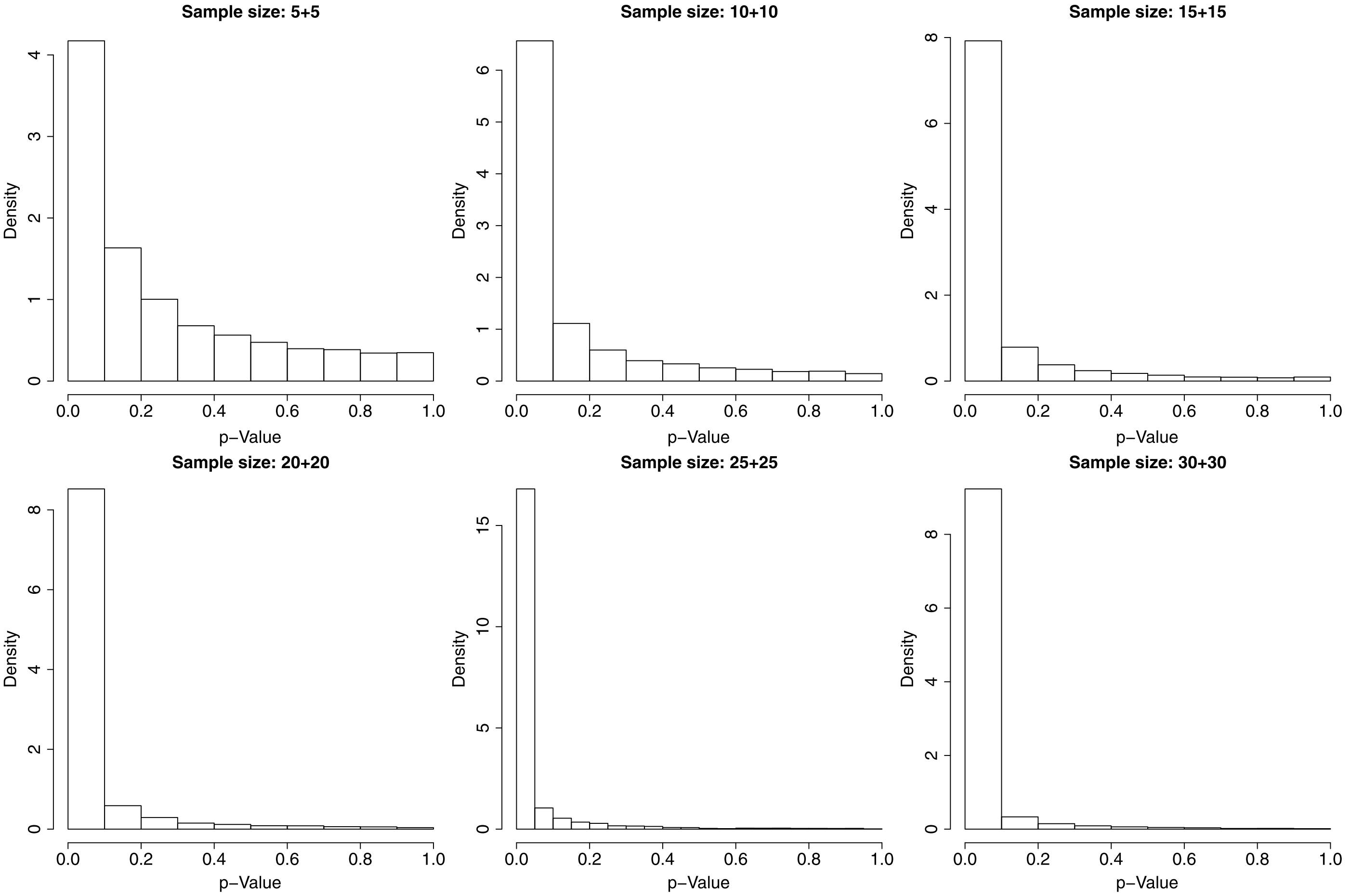

Figure 3. Test p-value histograms based on the simulated non-null data distributions.

on the other sample sizes, 30+30 and 50+50 (results not shown).

Remark. Notice that two-sided p-values have been consistently used in this study. To understand the p-value distribution under the non-null hypothesis more clearly, we simulate expression of 6000 genes all with $\mu \sim U[0.5, 1.5]$ (all differentially expressed). We considered different sample sizes: $n_1 = n_2 = 5, 10, 15, 20, 25, 30$, and calculate some p-value quantiles and generate the histograms. The 75-percentiles of p-values are 0.402, 0.179, 0.069 0.035, 0.015 and 0.007 when the total sample sizes $(n_1 + n_2)$ are 10, 20, 30, 40, 50 and 60 respectively; the corresponding 95-percentiles are 0.857, 0.712, 0.496, 0.353, 0.243 and 0.163. The histograms in Figure 3 confirm further that there are still many large p-values even when the sample size is relatively large.

Received 9 February 2010

REFERENCES

[1] Allison, D. B., Gadbury, G., Heo, M, Fernandez, J, Lee, C-K, Prolla, T. A., and Weindruch, R. (2002). A Mixture Model Approach For The Analysis Of Microarray Gene Expression Data. *Computational Statistics & Data Analysis*, **39**, 1–20. MR1895555

[2] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300. MR1325392

[3] Chen, H., Chen, J. and Kalbfleisch, J.D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, Series B*, **63**, 19–29. MR1811988

[4] Guan, Z., Wu, B. and Zhao, H. (2008) Nonparametric estimator of false discovery rate based on Bernstein polynomials. *Statistica Sinica*, **18**, 905–923. MR2440398

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

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

[7] Lai, Y. (2006) A statistical method for estimating the proportion of differentially expressed genes. *Computational Biology and Chemistry*, **30**, 193–202.

[8] 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.

[9] Langaas, M., Lindqvist, B.H. and Ferkingstad, E. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B*, **67**, 555–572. MR2168204

[10] Lehmann, E.L. and Romano, J.P. (2005) Testing statistical hypotheses. Springer, New York, pp. 513–517. MR2135927

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

[12] Lo, Y., Mendell, N.R. and Rubin, D.B. (2001) Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778. MR1859408

[13] Lo, Y. (2005) Likelihood ratio tests of the number of components in a normal mixture with unequal variances. *Statistics & Probability Letters*, **71**, 225–235. MR2126407

[14] Markistsis, A. and Lai, Y. (2010) A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, **26**, 640–646.

[15] McLachlan, G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society, Series C*, **36**, 318–324.

[16] McLachlan, G. and Krishnan, T. (2008) The EM algorithm and extensions, 2nd edition. John Wiley & Sons, Inc., pp. 18–26. MR2392878

[17] Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L. (2003) PGC-1α-response genes involved in oxidative phos-phorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267–273.

[18] Nettleton, D., Hwang, J.T.G., Caldo, R.A. and Wise R.P. (2006) Estimating the number of true null hypotheses from a histogram of p Values. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 337–356.

[19] 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.

[20] Qin, Y.S. and Smith, B. (2004) Likelihood ratio test for homogeneity in normal mixtures in the presence of a structural parameter. *Statistical Sinica*, **14**, 1165–1177. MR2126346

[21] Scheid, S. and Spang, R. (2004) A Stochastic Downhill Search Algorithm for Estimating the Local False Discovery Rate. *IEEE Transactions on Computational Biology and Bioinformatics*, **1**, 98–108.

[22] Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA*, **100**, 9440–9445. MR1994856

[23] Xu, X. and Liu, F. (2008) Statistical inference on mixing proportion. *Science in China Series A: Mathematics*, **51**, 1593–1608. MR2426056

[24] Wang, S-J. and Chen, J.J. (2004) Sample size for identifying differentially expressed genes in microarray experiments. *Journal of Computational Biology*, **11**, 714–726.

Anastasios Markitsis
Department of Statistics
The George Washington University
Washington D.C., 20052, USA
E-mail address: amarkits@gwu.edu

Yinglei Lai
Department of Statistics and Biostatistics Center
The George Washington University
Washington D.C., 20052, USA
E-mail address: ylai@gwu.edu