Journal of Applied Statistics, 2013
Vol. 40, No. 9, 1949–1964, http://dx.doi.org/10.1080/02664763.2013.800035

Taylor & Francis
Taylor & Francis Group

Estimating the proportion of true null hypotheses using the pattern of observed *p*-values

Tiejun Tong[a,b]*, Zeny Feng[c], Julia S. Hilton[d] and Hongyu Zhao[e]

[a]*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China;* [b]*Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China;* [c]*Department of Mathematics and Statistics, University of Guelph, Guelph, Canada;* [d]*Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, USA;* [e]*Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, USA*

(*Received 13 April 2012; accepted 24 April 2013*)

Estimating the proportion of true null hypotheses, π_0, has attracted much attention in the recent statistical literature. Besides its apparent relevance for a set of specific scientific hypotheses, an accurate estimate of this parameter is key for many multiple testing procedures. Most existing methods for estimating π_0 in the literature are motivated from the independence assumption of test statistics, which is often not true in reality. Simulations indicate that most existing estimators in the presence of the dependence among test statistics can be poor, mainly due to the increase of variation in these estimators. In this paper, we propose several data-driven methods for estimating π_0 by incorporating the distribution pattern of the observed *p*-values as a practical approach to address potential dependence among test statistics. Specifically, we use a linear fit to give a data-driven estimate for the proportion of true-null *p*-values in $(\lambda, 1]$ over the whole range $[0, 1]$ instead of using the expected proportion at $1 - \lambda$. We find that the proposed estimators may substantially decrease the variance of the estimated true null proportion and thus improve the overall performance.

Keywords: gene expression data; multiple testing; proportion of true null hypotheses; *p*-value

1. Introduction

Estimating the proportion of true null hypotheses, π_0, among a set of hypotheses to be tested, has attracted much attention in the recent statistical literature. Besides its apparent relevance in an overall assessment of all the hypotheses to be tested, the estimate of this proportion is needed in many multiple testing procedures. A reliable estimate of π_0 can eliminate the conservative bias

*Corresponding author. Email: tongt@hkbu.edu.hk

of the Benjamini–Hochberg procedure [1] on controlling the false discovery rate and therefore increase the average power [2,19,26]. A good estimate of π_0 can also sharpen the Bonferroni-type familywise error-controlling procedures to improve the power and reduce the false negative rate [7,9,31]. Therefore, it is no surprise that there is an increasing literature on estimating π_0.

Let m be the total number of tests and $p_1, \ldots, p_m$ be the observed p-values. For convenience, we denote the 'true-null p-values' as p-values corresponding to the true null hypotheses, and the 'false-null p-values' as p-values corresponding to the false null hypotheses. Noting that the false-null p-values are more likely to be small, for a reasonably large λ, the majority of p-values in $(\lambda, 1]$ corresponds to true-null p-values. Schweder and Spjøtvoll [23] proposed a graphical method to estimate π_0 by determining λ as the breakdown point in the p-value plot where the false-null p-values tend to be counted in. Storey [26] proposed a similar estimator as $\hat{\pi}_0(\lambda) = W(\lambda)/(m(1-\lambda))$, where $W(\lambda) = \#\{p_i > \lambda\}$ is the total number of p-values greater than λ, and $1-\lambda$ is the expected proportion of true-null p-values in $(\lambda, 1]$ over the whole range $[0, 1]$. This estimator is conservative and a bootstrap procedure on automatically choosing λ was discussed in [28]. Nettleton *et al.* [18] also contributed some work on the choice of λ. In addition, Storey and Tibshirani [27] introduced a smoothing spline estimator for π_0, denoted by $\hat{\pi}_0^s$. Specifically, they first calculate $\hat{\pi}_0(\lambda)$ over a grid of $\lambda \in \mathcal{R}$, for example, $\mathcal{R} = \{0, 0.01, \ldots, 0.95\}$, by some existing method and then estimate π_0 at $f(0.9)$, where $f(\cdot)$ is the fitted natural cubic spline curve of $(\lambda, \hat{\pi}_0(\lambda))$. Langaas *et al.* [14] proposed some new estimators based on a non-parametric maximum-likelihood estimation of the p-value density, subject to the restriction that the density is decreasing or convex decreasing. In general, their estimator $\hat{\pi}_0^c$, based on a convex decreasing density estimation, outperforms other estimators with respect to the mean-squared error (MSE). Other estimators for estimating the proportion of true null hypotheses include Hsueh *et al.* [10] and Ruppert *et al.*'s [21] least-squares methods, Dalmasso *et al.* [6] and Lai's [13] moment-based methods, Wu *et al.*'s [33] non-parametric method, Lu and Perkins's [15] resampling method, Cabrera and Yu's [3] empirical distribution method, Jiang and Doerge's [11] average estimate method, Tamhane and Shi's [29] mixture model method, Celisse and Robin's [4] cross-validation-based method, and Wang *et al.*'s [32] sliding linear model method, among others.

Most existing methods for estimating π_0 in the literature are derived based on the assumption that the test statistics are independent or based on the assumption that the true-null p-values are independently and uniformly distributed on $[0, 1]$ [8]. In either case, they rely on the fact that there is no clear pattern in the histogram of true-null p-values in $(\lambda, 1]$, given a reasonably large λ. However, the independence assumption is often not true in reality, especially in microarray gene expression data, where it is common that genes are correlated with each other due to co-regulation. As seen in Figure 5 in [14], simulations indicate that the performances of most existing estimators become rapidly worse when the dependence among test statistics is stronger. Moreover, we observe that although correlation among test statistics does not affect the estimation bias too much, it increases the variance. The increase in variance contributes to the major increase in the MSE of $\hat{\pi}_0$.

In this paper, we propose several data-driven methods for estimating π_0 by incorporating the distribution pattern of the observed p-values and hope this method may improve the estimation precision in the presence of dependence among hypothesis tests. Specifically, we use a linear fit to give a data-driven estimate for the proportion of true-null p-values in $(\lambda, 1]$ over the whole range $[0, 1]$ instead of using the expected proportion of $1-\lambda$ under the independent assumption. The proposed estimators tend to have smaller variances compared to the existing methods and thus improve the overall performance. In Section 2, we present a motivating example and some explorations and key theorems. We propose new estimators for π_0 in Section 3, where the optimal choice of λ is also discussed. We then evaluate the proposed method through both simulation studies in Section 4 and real data analysis in Section 5. Finally, we conclude the paper with a brief discussion in Section 6.

2. Motivation

In this section, we explore the p-value histogram when the test statistics are correlated with each other.

2.1 *Motivating example*

Consider a microarray experiment that has m genes with n arrays. We simulate each array $\mathbf{x}_j = (x_{1j}, \ldots, x_{mj})^{\mathrm{T}}$, $j = 1, \ldots, n$, from a multivariate normal distribution with zero means and an $m \times m$ compound symmetric covariance matrix

$$\Sigma_{m\times m}(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \cdots & \rho & 1 \end{pmatrix}. \tag{1}$$

Let $m = 500$ and $n = 5$. Let $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$ denote the sample mean and $s_i = [\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2/(n-1)]^{1/2}$ denote the sample standard deviation for $i = 1, \ldots, m$. We use the following one-sample t-test

$$T_i = \frac{\sqrt{n}\bar{x}_i}{s_i},$$

to test $H_{i0} : \mu_i = 0$ versus $H_{i1} : \mu_i \neq 0$. The p-value for this two-sided test is then evaluated at

$$p_i = 2F(-|T_i|),$$

where $F(\cdot)$ is the probability function of Student's t distribution with $n - 1$ degrees of freedom. In this simulation, we are interested in investigating the histogram shape of the generated p-values when $\rho \neq 0$, that is, when the genes are correlated with each other.

Specifically, we use the following chi-square goodness-of-fit test to compare the pattern change in histograms between $\rho \neq 0$ and $\rho = 0$. We divide all the p-values into five equal-length bins and define the test statistic as $\chi^2 = \sum_{i=1}^{5}(O_i - E_i)^2/E_i$, where O_i is the observed frequency for bin i and E_i is the expected frequency for the corresponding bin which is 100 in this example.

We should point out that, regardless of the dependence among p-values, the marginal distribution of the null p-values remains a standard uniform distribution. Thus, the expected frequency in each bin is always at 100. For $\rho = 0$, the null hypothesis that the p-values are a random sample from the uniform distribution on $[0, 1]$ is rejected if $\chi^2 > \chi^2_4(\alpha)$, where $\chi^2_4(\alpha)$ is the upper αth-quantile of a chi-square distribution with four degrees of freedom. For non-zero ρ, we follow the same test procedure as that for $\rho = 0$ to investigate whether the histogram pattern changes along the magnitude of the correlation. We set $\alpha = 0.01$ and report the total number of rejections based on 1000 simulations in Table 1.

Table 1. Total number of rejections in 1000 simulations for various ρ with the expected number of rejections equals to 10, where $\Sigma_{m\times m}(\rho)$ is the compound symmetric covariance matrix defined in Equation (1) and $\hat{\Sigma}_{m\times m}(\rho)$ is Cui *et al.*'s [5] covariance matrix defined in Equation (8).

ρ	0	0.2	0.4	0.6	0.8
$\Sigma_{m\times m}(\rho)$	9	342	667	850	978
$\hat{\Sigma}_{m\times m}(\rho)$	12	96	380	583	734

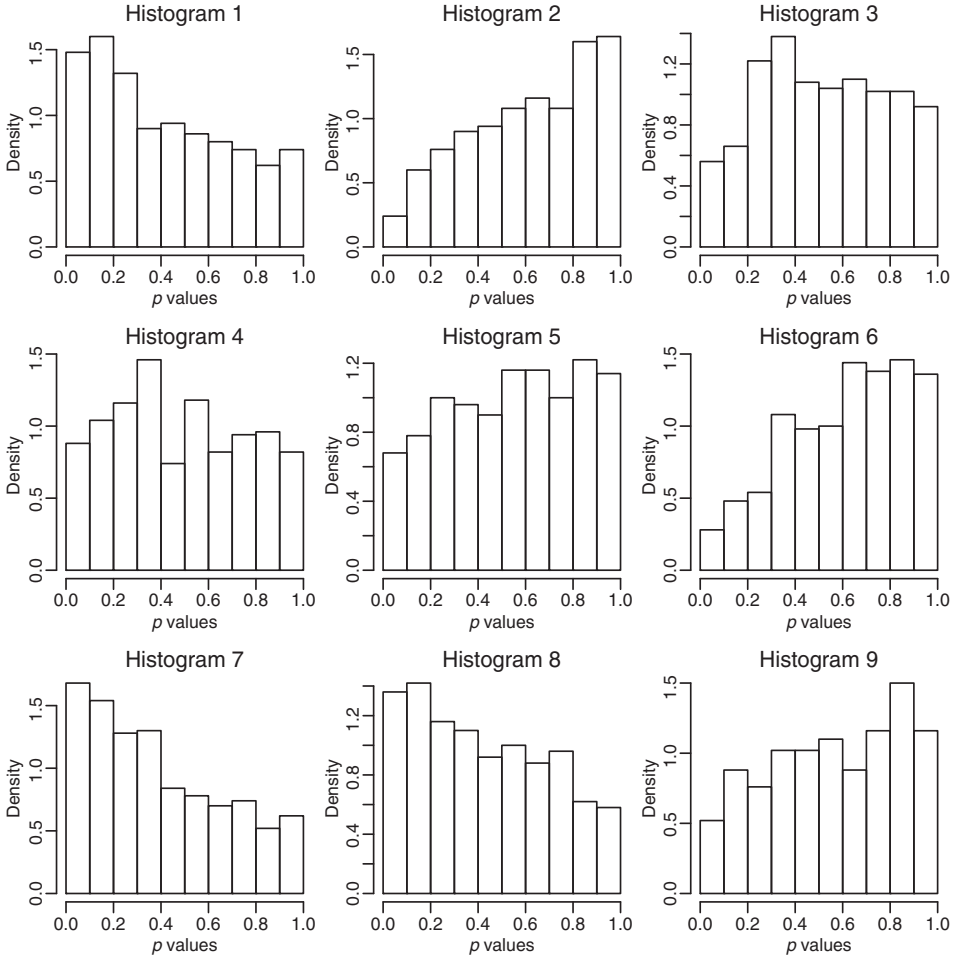

Figure 1. Histograms for nine simulated p-value sets when $\rho = 0.5$.

When all the genes are independent of each other (i.e. $\rho = 0$), the total number of rejections is at the nominal significance level of the goodness-of-fit test. The number of rejections increases with ρ rapidly. For example, when $\rho = 0.4$, about two-thirds of simulations suggest a rejection of p-values being independent and uniformly distributed. Figure 1 displays nine typical histogram patterns for simulated p-values with $\rho = 0.5$, where all are rejected at the significant level $\alpha = 0.01$ except for those in Histograms 3 and 4. Noting that $E(O_i)$ stays the same for any ρ, the greater number of rejections for non-zero ρ is mainly caused by the increased variation in O_i. Consequently, the p-value histogram is less likely to be flat and the pattern is more arbitrary than that coming from a random sample of the uniform distribution. Though arbitrary, we see that most simulated histograms tend to have some trend, either increasing or decreasing along the range (Figure 1). This implies that the estimation of the proportion of true nulls, π_0, might be improved if we utilize the histogram pattern of the observed p-values.

To provide a more realistic example, we now simulate data from a multivariate normal distribution with a covariance matrix generated in a real data set [5]. Let $\hat{\Sigma}_{m\times m}$ be the generated covariance matrix. To mimic difference levels of covariance structure, we further define $\hat{\Sigma}_{m\times m}(\rho) = (1-\rho)\text{diag}(\hat{\Sigma}_{m\times m}) + \rho\hat{\Sigma}_{m\times m}$ that scales the off-diagonal covariance by a

constant ρ. When $\rho = 1$, $\hat{\Sigma}_{m\times m}(\rho)$ is equivalent to $\hat{\Sigma}_{m\times m}$. When $\rho = 0$, $\hat{\Sigma}_{m\times m}(\rho)$ reduces to a diagonal matrix $\text{diag}(\hat{\Sigma}_{m\times m})$. See Section 4 for the detailed description of the data. We then simulate data from a multivariate normal distribution with zero means and covariance matrix $\hat{\Sigma}_{m\times m}(\rho)$. The remaining procedure is the same as before and we also report the total number of rejections in Table 1. Once again, we observe that the total number of rejections increases with ρ rapidly.

2.2 *Further exploration*

In this section, we study the joint density of a pair of p-values. By Equation (1), it is easy to see that for any gene pair i and j, $(x_{ik}, x_{jk})^{\mathrm{T}}, k = 1, \ldots, n$, is a random sample from the bivariate normal distribution with zero means and covariance matrix $\Sigma_{2\times 2} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. We consider the one-sided test: $H_0 : \mu = 0$ against $H_1 : \mu > 0$, for each gene. For simplicity, we assume that the standard error (SE) of sample mean is known ($\text{SE} = 1/\sqrt{n}$) and consider the corresponding z-test, $Z_i = \bar{x}_i/\text{SE} = \sqrt{n}\bar{x}_i$, where $\bar{x}_i$ is the sample mean for gene i. Under the null hypotheses, we have $(Z_i, Z_j)^{\mathrm{T}} \sim \text{MVN}(\mathbf{0}_2, \Sigma_{2\times 2})$, where MVN represents a multivariate normal distribution, and $p_i = 1 - \Phi(Z_i)$ for $Z_i > 0$, where $\Phi(\cdot)$ is the cdf of a standard normal distribution. Instead, when the SE is unknown and $T_i = \sqrt{n}\bar{x}_i/s_i$ is used for testing, the joint null distribution of (T_i, T_j) is a bivariate t distribution [17,25]. Note that the marginal distributions of Z_i and Z_j are $N(0, 1)$ and the corresponding marginal distributions of p_i and p_j are uniformly distributed on $[0, 1]$, respectively. Let z_α be the αth quantile of a standard normal distribution. In the appendix, we prove the following lemma.

LEMMA 1 *Under H_0, for any $|\rho| \neq 1$, the joint distribution of p_1 and p_2 is given as $f(p_1, p_2) = f(p_2 \mid p_1)f(p_1)$, where $f(p_1) = 1$ and*

$$f(p_2|p_1) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{-\frac{\rho^2 z_{p_1}^2 - 2\rho z_{p_1} z_{p_2} + \rho^2 z_{p_2}^2}{2(1-\rho^2)}\right\}. \tag{2}$$

When $\rho = 0, p_1$ and p_2 are independent of each other by noting that $f(p_2 \mid p_1)|_{\rho=0} = f(p_2) = 1$. This result does not hold for non-zero ρ. In general, $f(p_2 \mid p_1)$ is unimodal for $p_1 \in (0, 1)$, and the density becomes sharper when ρ is larger (Figure 2). The conditional density is symmetric only for $p_1 = 0.5$. For $p_1 < 0.5, f(p_2 \mid p_1)$ is right skewed with a mode less than p_1. Conversely, for $p_1 > 0.5$, it becomes left skewed with a mode larger than p_1.

For any interval $(a_l, b_l] \in [0, 1]$, let $I_{i,l} = I_{\{a_l < p_i \le b_l\}}$ denote the indicator function that p_i falls in the interval $(a_l, b_l]$ and $O_l = \sum_{i=1}^{\mathrm{m}} I_{i,l}$ be the total number of p-values in that interval. Then, we have the following.

THEOREM 1 *Suppose that $(a_1, b_1]$ and $(a_2, b_2]$ are any two disjoint intervals on $[0, 1]$. We have $E(O_l) = m(b_l - a_l)$ for $l = 1, 2$, and*

$$\text{var}(O_l) = m(m-1)\int_{a_l}^{b_l}\int_{a_l}^{b_l} f(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2 + m(b_l - a_l) - m^2(b_l - a_l)^2,$$

$$\text{cov}(O_1, O_2) = m(m-1)\int_{a_2}^{b_2}\int_{a_1}^{b_1} f(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2 - m^2(b_1 - a_1)(b_2 - a_2).$$

The proof of Theorem 1 is straightforward by noting that $\text{var}(O_l) = \sum_{i=1}^{\mathrm{m}} \text{var}(I_{i,l}) + \sum_{j\neq i} \text{cov}(I_{i,l}, I_{j,l})$ and $\text{cov}(O_1, O_2) = \sum_{i=1}^{\mathrm{m}} \text{cov}(I_{i,1}, I_{i,2}) + \sum_{j\neq i} \text{cov}(I_{i,1}, I_{j,2})$. When all tests are

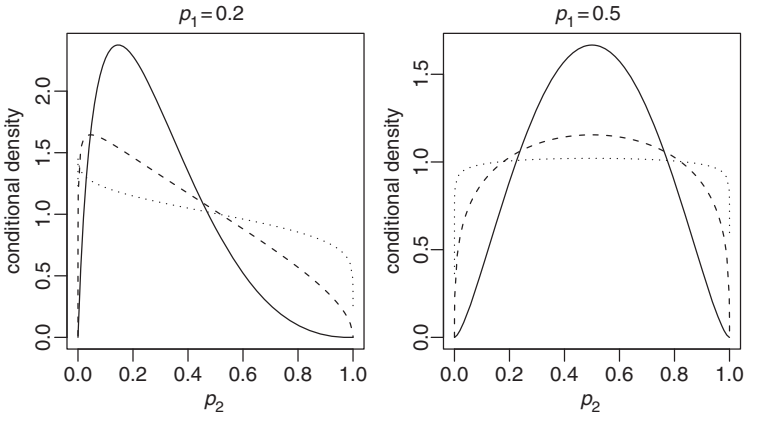

Figure 2. Conditional densities of p_2 given the value of p_1 at 0.2 (the left panel) or 0.5 (the right panel). In both panels, three curves (solid, dashed and dotted) correspond to three different ρ $(0.8, 0.5, 0.2)$, respectively.

independent of each other (i.e. $\rho = 0$), we have $E(O_1) = m(b_1 - a_1)$ and $\text{var}(O_1) = m(b_1 - a_1) - m(b_1 - a_1)^2$. This shows that O_1 follows a binomial distribution with parameters m and $(b_1 - a_1)$. Furthermore, we can assume that the set of null p-values is a random sample from the standard uniform distribution. However, this assumption is violated when $\rho \neq 0$. For example, noting that $\int_{a_1}^{b_1}\int_{a_1}^{b_1} f(p_1, p_2)\mathrm{d}p_1\mathrm{d}p_2 \to (b_1 - a_1)$ as $\rho \to 1$, we have $\lim_{\rho\to 1} \text{var}(O_1) = m^2(b_1 - a_1) - m^2(b_1 - a_1)^2$, which is m times larger than the expected value. This explains why the variations in O_i increase rapidly along with the dependence strength among test statistics in Section 2.1.

3. New estimators for π_0

Our starting point is the same as that in [23,26]. Since the p-values from false nulls are more likely to be small, the majority of p-values in $(\lambda, 1]$ should correspond to true nulls when λ is reasonably large. Denote $W(\lambda) = \#\{p_i > \lambda\}$ as the total number of p-values in $(\lambda, 1]$, as defined in Section 1. Under the assumption that each true-null p-value is uniformly distributed on $[0, 1]$, the expected proportion of true-null p-values in $(\lambda, 1]$ over the whole range of $[0, 1]$ equals $1 - \lambda$. This implies that $E(W(\lambda)) \approx (1 - \lambda)m\pi_0$, and thus a conservative estimate of π_0 is given as

$$\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{(m(1 - \lambda))}. \tag{3}$$

Note that the expectation of $\hat{\pi}_0(\lambda)$, $E(\hat{\pi}_0(\lambda)) = \sum_{i=1}^{\mathrm{m}} EI\{p_i > \lambda\}/(m(1 - \lambda))$, stays the same regardless of the dependence among the test statistics. Therefore, the estimator in Equation (3) is conservative under arbitrary levels of dependence. However, as mentioned in the previous sections, the variation of $\hat{\pi}_0(\lambda)$ may increase rapidly with the dependence among test statistics and thus results in the poor performance of the estimator. In Panel A of Figure 5 in Section 5, we display the histogram of p-values from a study with the objective to identify the targets of the *Arf* gene on the Arf-Mdm2-p53 tumor suppressor pathway. The cDNA array data set was conducted in St Jude Childrens Research Hospital, and the p-values were generated in [20]. We note that the histogram of p-values is monotonically decreasing in its whole range $[0, 1]$. That is, there does not exist a reasonable λ such that $E(W(\lambda)) \approx (1 - \lambda)m\pi_0$ holds. More precisely, for a reasonably large λ such that most p-values in $(\lambda, 1]$ are from true nulls, we expect that $E(W(\lambda)) < (1 - \lambda)m\pi_0$ if the histogram of true-null p-values has a decreasing trend. Motivated

by this problem, we propose several data-driven methods to estimate π_0 by incorporating the distribution pattern of the observed p-values.

For simplicity, we assume that a good choice of λ is available at this moment. Methods on choosing the optimal λ will be discussed later in Section 3.3. Let $W_0(\lambda)$ be the total number of true-null p-values in $(\lambda, 1]$, and

$$f(\lambda) = \frac{W_0(\lambda)}{(m\pi_0)} \tag{4}$$

be the proportion of true-null p-values in $(\lambda, 1]$ over the whole range of $[0, 1]$. Throughout the paper, we take the integer part of $m\pi_0$ whenever necessary. For a well-chosen λ such that only few of false-null p-values are greater than λ, we have $W_0(\lambda) \approx W(\lambda)$. Therefore, by Equation (4),

$$\pi_0 = \frac{W_0(\lambda)}{mf(\lambda)} \approx \frac{W(\lambda)}{mf(\lambda)}. \tag{5}$$

When $f(\lambda)$ is replaced by its expectation $E(f(\lambda)) = 1 - \lambda$, the estimator (5) is simplified to Equation (3), which is a good estimate of π_0 when the true-null p-values are independent of each other. Otherwise, as discussed in Section 2, $1 - \lambda$ can be a poor estimate of $f(\lambda)$ since the pattern of histogram is less likely to be flat. This motivates us to consider new estimators of $f(\lambda)$ for improving the estimation accuracy.

3.1 *Estimation of $f(\lambda)$*

In this section, we propose three estimators for $f(\lambda)$. The first estimator is proposed by assuming that the pattern of true-null p-values is linear (see the solid line in Figure 3). Specifically, we

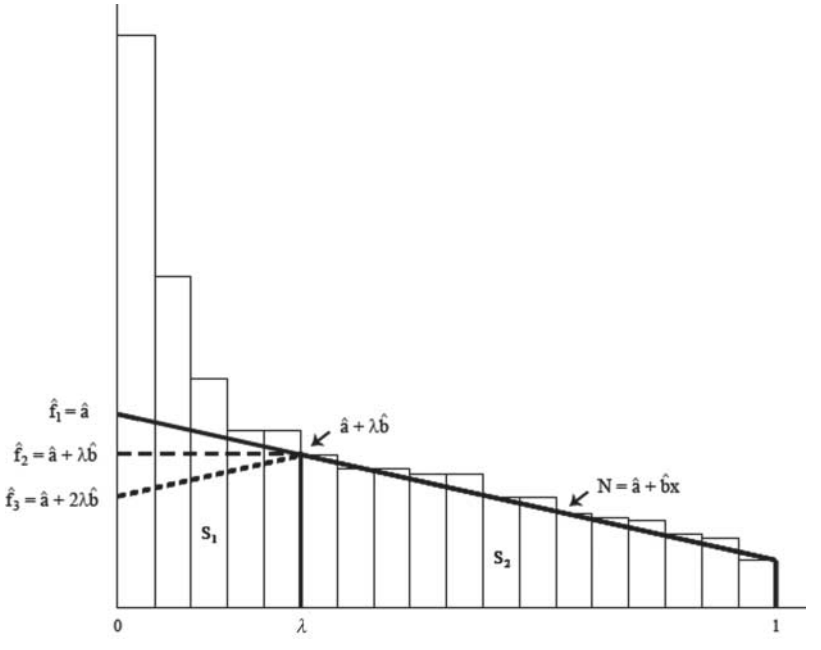

Figure 3. Sketch plot for estimating $f_1(\lambda)$, $f_2(\lambda)$ and $f_3(\lambda)$.

estimate $f(\lambda)$ by

$$\hat{f}_1(\lambda) = \frac{S_2}{(S_1 + S_2)}, \tag{6}$$

where S_1 and S_2 are the estimated numbers of true-null p-values in $[0, \lambda]$ and $(\lambda, 1]$, respectively. The detailed procedure is given as follows.

(a) Partition the interval $(\lambda, 1]$ into K equal-length bins with the ith bin $(\lambda + (i-1)\tau, \lambda + i\tau]$, where $\tau = (1-\lambda)/K$.
(b) Let $x_i = \lambda + (2i-1)\tau/2$ be the middle point of each bin, and let $N_i = \#\{\lambda + (i-1)\tau < p_i \leq \lambda + i\tau\}$ be the number of p-values in the ith bin.
(c) Regress N_i linearly on x_i. We have

$$\hat{a} = \bar{N} - \hat{b}\bar{x},$$
$$\hat{b} = \frac{\sum_{i=1}^{K}(x_i - \bar{x})(N_i - \bar{N})}{\sum_{i=1}^{K}(x_i - \bar{x})^2},$$

where $\bar{N} = \sum_{i=1}^{K} N_i/K$ and $\bar{x} = \sum_{i=1}^{K} x_i/K$.
(d) Estimate $f(\lambda)$ by $\hat{f}_1(\lambda) = S_2/(S_1 + S_2)$. Specifically,

$$\hat{f}_1(\lambda) = \begin{cases} \dfrac{(1-\lambda)(2\hat{a} + (1+\lambda)\hat{b})}{2\hat{a} + \hat{b}}, & \hat{a} > 0 \quad \text{and} \quad \hat{a} + \hat{b} > 0, \\ \dfrac{(\hat{a} + \lambda\hat{b})^2}{\hat{a}^2}, & \hat{a} > 0 \quad \text{and} \quad \hat{a} + \hat{b} \leq 0, \\ 1, & \hat{a} + \lambda\hat{b} < 0 \quad \text{and} \quad \hat{a} + \hat{b} > 0, \\ 1 - \dfrac{(\hat{a} + \lambda\hat{b})^2}{(\hat{a} + \hat{b})^2}, & \text{otherwise.} \end{cases}$$

Note that S_1 and S_2 are non-negative. In step (d), we have replaced the fitted negative values by zero to have more accurate estimates of S_1 and S_2. Now with $\hat{f}_1(x)$, we estimate π_0 by $\hat{\pi}_{01}(\lambda) = W(\lambda)/[m\hat{f}_1(\lambda)]$. In the special case when $\hat{b} = 0$, we have $\hat{f}_1(\lambda) = 1 - \lambda$ and so $\hat{\pi}_{01}(\lambda)$ reduces to $\hat{\pi}_0(\lambda)$ in Equation (3). Note that K serves as a tuning parameter in the estimation. We found that K is not sensitive to the performance of $\hat{\pi}_{01}(\lambda)$. Therefore, for simplicity we set $K = 10$ throughout the simulations.

Note that in practice, as shown in Figure 1, the histogram of true-null p-values can be of other shapes such as unimodal or U-shaped. For this reason, we consider here two other estimators for $f(\lambda)$: (i) treat the histogram of true-null p-values as flat in the unobservable range of $[0, \lambda]$ (the dashed line in Figure 3). Specifically, we estimate $f(\lambda)$ by

$$\hat{f}_2(\lambda) = \begin{cases} \dfrac{(1-\lambda)(2\hat{a} + (1+\lambda)\hat{b})}{2\hat{a} + (1+\lambda^2)\hat{b}}, & \hat{a} + \lambda\hat{b} > 0 \quad \text{and} \quad \hat{a} + \hat{b} > 0, \\ \dfrac{\hat{a} + \lambda\hat{b}}{\hat{a} - \lambda\hat{b}}, & \hat{a} + \lambda\hat{b} > 0 \quad \text{and} \quad \hat{a} + \hat{b} \leq 0, \\ 1, & \hat{a} + \lambda\hat{b} \leq 0 \quad \text{and} \quad \hat{a} + \hat{b} > 0. \end{cases}$$

(ii) Or instead we take a retroflection in the unobservable range of $[0, \lambda]$ when the slope $\hat{b}$ is negative (the dotted line in Figure 3), which leads to

$$\hat{f}_3(\lambda) = \begin{cases} \dfrac{(1-\lambda)(2\hat{a}+\lambda\hat{b}+\hat{b})}{2\hat{a}+\hat{b}(1+2\lambda^2)}, & \hat{b}<0 \quad \text{and} \quad \hat{a}+\hat{b}\geq 0 \quad \text{and} \quad \hat{a}+2\lambda\hat{b}\geq 0, \\ \dfrac{(\hat{a}+\lambda\hat{b})^2}{\hat{a}^2-2\lambda^2\hat{b}^2}, & \hat{b}<0 \quad \text{and} \quad \hat{a}+\hat{b}<0 \quad \text{and} \quad \hat{a}+2\lambda\hat{b}\geq 0, \\ \dfrac{1}{2}, & \hat{b}<0 \quad \text{and} \quad \hat{a}+\hat{b}<0 \text{ and } \hat{a}+2\lambda\hat{b}<0, \\ \dfrac{(1-\lambda)(2\hat{a}+\lambda\hat{b}+\hat{b})}{2\hat{a}+\hat{b}}, & \hat{b}\geq 0 \quad \text{and} \quad \hat{a}\geq 0, \\ 1-\left(\dfrac{\hat{a}+\lambda\hat{b}}{\hat{a}+\hat{b}}\right)^2, & \hat{b}\geq 0 \quad \text{and} \quad \hat{a}<0 \quad \text{and} \quad \hat{a}+\lambda\hat{b}\geq 0, \\ 1, & \hat{b}\geq 0 \quad \text{and} \quad \hat{a}+\lambda\hat{b}<0. \end{cases}$$

3.2 *Estimation of* π_0

In this section, we construct new estimators for π_0 based on the proposed estimates $\hat{f}_i(\lambda), i = 1, 2, 3$. We first use the observed p-values in $(\lambda, 1]$ to test whether the linear pattern has a slope at 0. That is, we test the hypothesis

$$H_0 : b = 0 \quad \text{versus} \quad H_a : b \neq 0.$$

The purpose is to see if there is evidence that the tests are highly correlated with each other. Note that λ plays a role in the above testing problem. When λ is too small, more false null p-values are likely to be included in the left side of the interval $(\lambda, 1]$ so that it may result in a higher rejection rate of H_0. Instead, if λ is too large, the interval $(\lambda, 1]$ will be not wide enough so that the testing result may be unreliable. In the following, we set $\lambda = \min(p_{\text{med}}, \frac{1}{2})$ for testing the null hypothesis that $b = 0$, where p_{med} is the median of observed p-values. More influence on the choice of λ will be discussed in Section 3.3.

If we do not reject H_0 at a significance level of α, we estimate π_0 by an existing method that performs well under the independence assumption. Recall that the convex density estimator of Langaas *et al.* [14], $\hat{\pi}_0^c$, is considered to be a good choice in terms of both bias and variance [11,14,15,18]. For this reason, we adopt $\hat{\pi}_0^c$ if H_0 is not rejected. On the other hand, if we reject H_0, we estimate π_0 by Equation (5) with proposed $\hat{f}_i(\lambda)$ that aims to alleviate the impact of dependence in a certain sense. Specifically, we construct the following estimators for π_0,

$$\hat{\pi}_{0i}^{\text{new}}(\lambda) = \begin{cases} \hat{\pi}_0^c, & \text{if } H_0 \text{ is not rejected at level} \alpha, \\ \hat{\pi}_{0i}(\lambda), & \text{otherwise}, \end{cases} \tag{7}$$

where $\hat{\pi}_{0i}(\lambda) = W(\lambda)/(m\hat{f}_i(\lambda))$ with $i = 1$, 2 or 3. Note that the significance level of α works as a tuning parameter. When $\alpha = 0$, H_0 will never be rejected such that $\hat{\pi}_{0i}^{\text{new}}(\lambda)$ is equivalent to $\hat{\pi}_0^c$. When α is large, H_0 is more likely to be rejected such that π_0 is more likely to be estimated by $\hat{\pi}_{0i}(\lambda)$. We set $\alpha = 0.01$ throughout the simulations.

3.3 *The choice of* λ

As in $\hat{\pi}_0(\lambda)$, there is a bias–variance trade-off in the choice of λ for the proposed estimators. Here, we use $\hat{\pi}_{01}(\lambda)$ for illustration. When $\lambda \to 0, \hat{f}_1(\lambda) \to 1$ and thus $\hat{\pi}_{01}(\lambda) \to W(0)/m = 1$,

which leads to an increase in bias, but a decrease in variance. On the other hand, when $\lambda \to 1$, the variance of $\hat{\pi}_{01}(\lambda)$ will increase rapidly. In practice, to balance the trade-off between the bias and variance, it is desirable to estimate the optimal λ by minimizing the mean square error $\mathrm{MSE}(\lambda) = \mathrm{E}(\hat{\pi}_0(\lambda) - \pi_0)^2$.

Storey *et al.* [28] proposed a bootstrap algorithm for choosing the optimal λ under the assumption of independent test statistics. They first define the search set to be $\mathcal{R} = \{0, 0.05, \ldots, 0.95\}$. Then, for each $\lambda \in \mathcal{R}$, they generate $\hat{\pi}_0^b (b = 1, \ldots, B)$, the bootstrap version of $\hat{\pi}_0(\lambda)$ in Equation (3), from the corresponding bth bootstrap sample of the p-values, respectively. Next, they estimate $\mathrm{MSE}(\lambda)$ by $\widehat{\mathrm{MSE}}_0(\lambda) = \sum_{b=1}^{\mathrm{m}} [\hat{\pi}_0^{*b}(\lambda) - \hat{\pi}_0^p]^2/B$, where the $\hat{\pi}_0^{*b}(\lambda)$ are the bootstrap versions of $\hat{\pi}_0(\lambda)$ and $\hat{\pi}_0^p \triangleq \min_{\lambda \in \mathcal{R}} \hat{\pi}_0(\lambda)$ is a plug-in estimator for π_0. Finally, they estimate $\hat{\pi}_0 = \hat{\pi}_0(\hat{\lambda}_0)$ with $\hat{\lambda}_0 = \mathrm{argmin}_{\lambda \in \mathcal{R}} \widehat{\mathrm{MSE}}_0(\lambda)$. More recently, Nettleton *et al.* [18] proposed another algorithm for choosing the optimal λ by drawing the connection between $\hat{\pi}_0(\lambda)$ and the iterative procedure for estimating π_0 by Mosig *et al.* [16]. The above two procedures have a similar performance in practice (see Nettleton *et al.* [18] for more details).

In this section, we suggest an adaptive bootstrap algorithm for choosing the optimal λ for $\hat{\pi}_{01}(\lambda)$. Similarly as in [28], we estimate MSE by $\widehat{\mathrm{MSE}}_{01}(\lambda) = \sum_{b=1}^{\mathrm{m}} [\hat{\pi}_{01}^{*b}(\lambda) - \hat{\pi}_{01}^p]^2/B$, where the $\hat{\pi}_{01}^{*b}(\lambda)$ are the bootstrap versions of $\hat{\pi}_{01}(\lambda)$ and $\hat{\pi}_{01}^p$ is a plug-in estimator of π_0. Noting that $\hat{\pi}_{01}^{\mathrm{new}}(\lambda)$ may not be conservative in general, we choose $\hat{\pi}_{01}^p$ as $\hat{\pi}_{01}(\min(p_{\mathrm{med}}, \frac{1}{2}))$ instead of $\min_{\lambda \in \mathcal{R}} \hat{\pi}_{01}(\lambda)$, where p_{med} is the median of all observed p-values. Also note that the true-null p-values in $[0, \lambda]$ are unobservable and we only use the information in $(\lambda, 1]$ to estimate $f(\lambda)$. Thus, if λ is too large, the estimator $\hat{\pi}_{01}(\lambda)$ will more likely be unreliable. For this reason, we constrain the search set to be within $[0, \frac{1}{2}]$. Specifically, we set $\mathcal{R}_{01} = \{k/40, k = 1, \ldots, 20\}$. Finally, we estimate the optimal λ by $\hat{\lambda}_{01} \triangleq \mathrm{argmin}_{\lambda \in \mathcal{R}_{01}} \widehat{\mathrm{MSE}}_{01}(\lambda)$, which leads to the estimator (7) as $\hat{\pi}_{01}^{\mathrm{new}}(\hat{\lambda}_{01})$.

We apply the same procedure to choose the optimal $\hat{\lambda}_{02}$ for $\hat{\pi}_{02}^{\mathrm{new}}(\hat{\lambda}_{02})$ and the optimal $\hat{\lambda}_{03}$ for $\hat{\pi}_{03}^{\mathrm{new}}(\hat{\lambda}_{03})$. For simplicity, in what follows we use $\hat{\pi}_{0i}^{\mathrm{new}}$ to represent $\hat{\pi}_{0i}^{\mathrm{new}}(\hat{\lambda}_{0i})$ for each i. The comparative performance of $\hat{\pi}_{01}^{\mathrm{new}}$, $\hat{\pi}_{02}^{\mathrm{new}}$ and $\hat{\pi}_{03}^{\mathrm{new}}$ is evaluated in Sections 4 and 5.

4. Simulation study

In this section, we conduct simulation studies to evaluate the performance of the proposed estimators. As in Section 2.1, we consider an m-gene experiment with n arrays. We simulate each array $\mathbf{x}_j = (x_{1j}, \ldots, x_{mj})^{\mathrm{T}}, j = 1, \ldots, n$, from a multivariate normal distribution $\mathrm{MVN}(\boldsymbol{\mu}_m, \Sigma_{m \times m})$, where the covariance matrix $\Sigma_{m \times m}$ is simulated from a real data set described in [5]. The data set is from a cDNA microarray experiment where the objective is to identify differentially expressed genes in two human colon cancer cell lines, CACO2 and HCT116, and three human ovarian cancer cell lines, ES2, MDAH2774 and OV1063 (see [5] for more details). In total, 9600 genes were tested. After some preliminary work and proper normalization, the authors fitted an analysis of variance model to each gene to account for the multiple sources of variation including array, dye and sample effects. The sample covariance, $\hat{\Sigma}_{9600 \times 9600}$, is then computed from the fitted residuals $\hat{\epsilon}_{9600 \times 20}$, where 20 is the total number of residuals for each gene. We treat these 9600 genes as the population and $\hat{\Sigma}_{9600 \times 9600}$ as the true dependence structure among genes. Then, in each simulation, we draw a random sample of size m from the population to achieve a sub-covariance matrix $\hat{\Sigma}_{m \times m}$, with rows and columns corresponding to the selected genes.

Set $m = 1000$. We consider two different sample sizes ($n = 5$ and 10) and two different proportions of true nulls ($\pi_0 = 0.7$ and 0.9). Let $\boldsymbol{\mu}_m = \{\mu_1, \ldots, \mu_m\}^{\mathrm{T}}$, where the first $m\pi_0$ entries are set to be 0 that correspond to true null hypotheses, while the other entries are non-zero corresponding to false null hypotheses alternatively. We treat $\{\mu_i, i = m\pi_0 + 1, \ldots, m\}$ as a random sample from

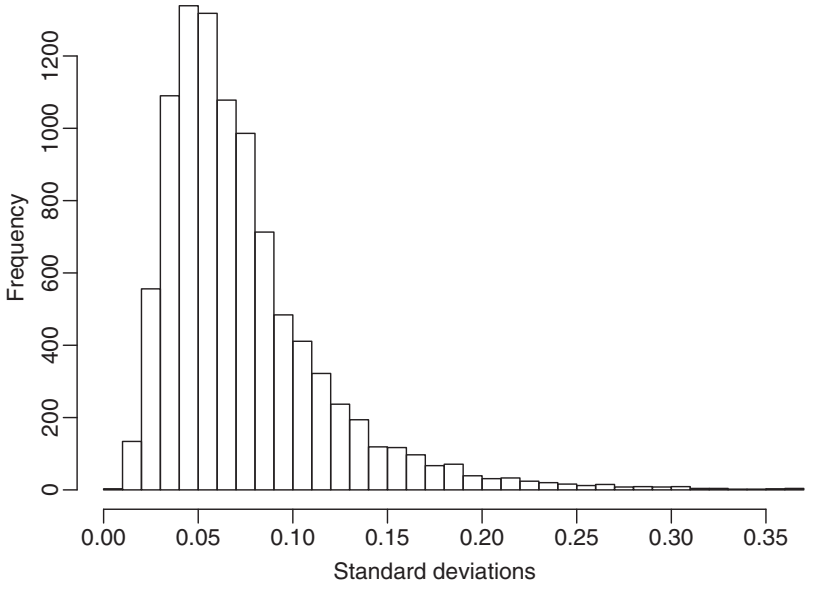

Figure 4. Histogram of the standard deviations for all the genes.

the uniform distribution $U[c_1, c_2]$. Since the mean standard deviation of all the genes is as small as 0.075 (Figure 4), we set $c_1 = 0.05$ and $c_2 = 0.15$ in simulations for having an appropriate range of the effect sizes. In the special case that all standard deviations are set to 1 as in [14] and many other papers, this corresponds to simulating $\{\mu_i,\ i = m\pi_0 + 1, \ldots, m\}$ from $U[\frac{2}{3}, 1]$. In each simulation, we simulate $\mathbf{x}_j, j = 1, \ldots, n$, from the multivariate normal distribution $\mathrm{MVN}(\boldsymbol{\mu}_m, \hat{\Sigma}_{m\times m})$. We then use one-sample t-tests to test $H_{i0} : \mu_i = 0$ against $H_{i1} : \mu_i \neq 0$ as in Section 2.1 and compute the corresponding p-values from Student's t distribution with $n - 1$ degrees of freedom. With the simulated p-values, we use the following seven estimators to estimate π_0:

$\hat{\pi}_0^{\mathrm{b}}$: the bootstrap estimator in [28],
$\hat{\pi}_0^{\mathrm{s}}$: the smooth spline estimator in [27],
$\hat{\pi}_0^{\mathrm{c}}$: the convex density estimator in [14],
$\hat{\pi}_0^{\mathrm{m}}$: the moment-based estimator in [13],
$\hat{\pi}_{0i}^{\mathrm{new}}$: the proposed estimators, $i = 1$, 2, and 3.

The moment-based estimator, $\hat{\pi}_0^{\mathrm{m}}$, is designed for $n \geq 4$ and generally requires $n \geq 8$ to have a comparable performance. For this reason, the performance of $\hat{\pi}_0^{\mathrm{m}}$ is studied only for $n = 10$. To investigate the effect of different levels of dependency among tests on the performance of each considered estimator, we multiply the off-diagonal elements of $\hat{\Sigma}_{m\times m}$ by a coefficient ρ to have

$$\hat{\Sigma}_{m\times m}(\rho) = (1 - \rho)\mathrm{diag}(\hat{\Sigma}_{m\times m}) + \rho\hat{\Sigma}_{m\times m}. \tag{8}$$

Here, we let $\rho = 0, 0.5$ and 1. $\rho = 0$ corresponds to the case of independent test statistics, $\rho = 0.5$ corresponds to the half-level correlation structure among test statistics, and $\rho = 1$ corresponds to the full-level correlation structure among test statistics as those in the original covariance matrix. We keep all other settings the same as before and repeat the simulation 1000 times for each combination of settings. We report their means, SEs, and MSEs in Table 2 for $n = 5$ and in Table 3 for $n = 10$.

Table 2. Simulated mean values (mean), SEs, and MSE for $n = 5$.

π_0	ρ		$\hat{\pi}_0^{\mathrm{b}}$	$\hat{\pi}_0^{\mathrm{s}}$	$\hat{\pi}_0^{\mathrm{c}}$	$\hat{\pi}_{01}^{\mathrm{new}}$	$\hat{\pi}_{02}^{\mathrm{new}}$	$\hat{\pi}_{03}^{\mathrm{new}}$
0.7	0	Mean	0.683	0.722	0.716	0.723	0.717	0.712
		(SE)	(0.066)	(0.075)	(0.031)	(0.033)	(0.032)	(0.032)
		MSE	0.0046	0.0061	0.0012	0.0016	0.0013	0.0012
	0.5	Mean	0.673	0.719	0.699	0.724	0.722	0.719
		(SE)	(0.119)	(0.140)	(0.105)	(0.080)	(0.081)	(0.075)
		MSE	0.015	0.020	0.011	0.007	0.007	0.006
	1	Mean	0.613	0.663	0.643	0.738	0.736	0.731
		(SE)	(0.174)	(0.206)	(0.167)	(0.121)	(0.121)	(0.123)
		MSE	0.038	0.044	0.031	0.016	0.016	0.016
0.9	0	Mean	0.858	0.901	0.897	0.897	0.897	0.896
		(SE)	(0.072)	(0.068)	(0.026)	(0.030)	(0.030)	(0.030)
		MSE	0.0069	0.0046	0.0007	0.0009	0.0007	0.0007
	0.5	Mean	0.824	0.864	0.855	0.872	0.870	0.869
		(SE)	(0.042)	(0.137)	(0.118)	(0.106)	(0.095)	(0.100)
		MSE	0.026	0.020	0.016	0.012	0.010	0.011
	1	Mean	0.751	0.780	0.783	0.857	0.851	0.850
		(SE)	(0.219)	(0.211)	(0.203)	(0.135)	(0.136)	(0.136)
		MSE	0.070	0.059	0.055	0.020	0.021	0.021

Table 3. Simulated mean values (mean), SEs, and MSEs for $n = 10$.

π_0	ρ		$\hat{\pi}_0^{\mathrm{b}}$	$\hat{\pi}_0^{\mathrm{s}}$	$\hat{\pi}_0^{\mathrm{c}}$	$\hat{\pi}_0^{\mathrm{m}}$	$\hat{\pi}_{01}^{\mathrm{new}}$	$\hat{\pi}_{02}^{\mathrm{new}}$	$\hat{\pi}_{03}^{\mathrm{new}}$
0.7	0	Mean	0.666	0.707	0.702	0.722	0.703	0.701	0.700
		(SE)	(0.064)	(0.070)	(0.026)	(0.018)	(0.030)	(0.026)	(0.026)
		MSE	0.0052	0.0050	0.0007	0.0008	0.0009	0.0007	0.0007
	0.5	Mean	0.656	0.710	0.681	0.720	0.697	0.697	0.696
		(SE)	(0.110)	(0.137)	(0.093)	(0.081)	(0.071)	(0.071)	(0.071)
		MSE	0.014	0.019	0.009	0.007	0.005	0.005	0.005
	1	Mean	0.589	0.650	0.616	0.684	0.695	0.694	0.692
		(SE)	(0.154)	(0.204)	(0.148)	(0.160)	(0.109)	(0.114)	(0.114)
		MSE	0.036	0.044	0.029	0.026	0.012	0.013	0.013
0.9	0	Mean	0.851	0.894	0.890	0.907	0.891	0.891	0.891
		(SE)	(0.067)	(0.066)	(0.024)	(0.021)	(0.027)	(0.025)	(0.025)
		MSE	0.0069	0.0044	0.0007	0.0005	0.0008	0.0007	0.0007
	0.5	Mean	0.826	0.868	0.848	0.881	0.857	0.856	0.856
		(SE)	(0.129)	(0.130)	(0.106)	(0.121)	(0.096)	(0.095)	(0.095)
		MSE	0.022	0.018	0.014	0.015	0.011	0.011	0.011
	1	Mean	0.752	0.786	0.776	0.792	0.829	0.824	0.822
		(SE)	(0.195)	(0.200)	(0.178)	(0.242)	(0.134)	(0.139)	(0.138)
		MSE	0.060	0.053	0.047	0.070	0.023	0.025	0.025

Similarly as in [14], we observed that $\hat{\pi}_0^{\mathrm{c}}$ performs better than $\hat{\pi}_0^{\mathrm{b}}$ and $\hat{\pi}_0^{\mathrm{s}}$ in each situation. We also observed that for $n = 10$, $\hat{\pi}_0^{\mathrm{m}}$ outperforms $\hat{\pi}_0^{\mathrm{b}}$ and $\hat{\pi}_0^{\mathrm{s}}$ in most cases and $\hat{\pi}_0^{\mathrm{m}}$ and $\hat{\pi}_0^{\mathrm{c}}$ perform similarly. Note that the performances of $\hat{\pi}_{01}^{\mathrm{new}}$, $\hat{\pi}_{02}^{\mathrm{new}}$ and $\hat{\pi}_{03}^{\mathrm{new}}$ are very comparable. In what follows for simplicity of comparison, we use $\hat{\pi}_{03}^{\mathrm{new}}$ for comparison with other estimators. The comparison between $\hat{\pi}_{03}^{\mathrm{new}}$ and $\hat{\pi}_0^{\mathrm{c}}$ is summarized as follows: (i) when the test statistics are independent of each other, $\hat{\pi}_{03}^{\mathrm{new}}$ performs similarly to $\hat{\pi}_0^{\mathrm{c}}$. To be specific, when $\rho = 0$, only few H_0 are rejected in Equation (7) at level $\alpha = 0.01$ so that $\hat{\pi}_{03}^{\mathrm{new}}$ is almost the same as $\hat{\pi}_0^{\mathrm{c}}$. (ii) When the test statistics are highly correlated, $\hat{\pi}_{03}^{\mathrm{new}}$ has a smaller MSE than $\hat{\pi}_0^{\mathrm{c}}$ in most settings and thus provides a more reliable estimate of π_0. We also observe that the MSEs of both methods increase along with the ρ value. Nevertheless, the impact of ρ on the estimator $\hat{\pi}_{03}(\hat{\lambda}_{03})$ is much less when compared with

the estimator $\hat{\pi}_0^c$. Specifically, when ρ is large, we tend to reject more null hypotheses so that the data-driven estimator $\hat{\pi}_{03}(\hat{\lambda}_{03})$ is more likely to be used.

Finally, it is worth mentioning that if, in the experiment, there are many weakly differentially expressed genes with p-values larger than 0.5, $\hat{b}$ will be a negatively biased estimator of the true slope. Such a situation tends to reject H_0 more likely and thus leads to a more conservative estimate of π_0. For this reason, we recommend the use of $\hat{\pi}_{03}^{\text{new}}$ when π_0 is reasonably large (e.g. $\pi_0 \geq 0.7$) and/or most of the effect sizes are not too small.

5. Real data analysis

We use three real data sets to evaluate our proposed methods. The first data set is from the experiment described by Kuo *et al.* [12]. The objective of the experiment was to identify the targets of the *Arf* gene on the Arf-Mdm2-p53 tumor suppressor pathway [12,24]. In this study, the cDNA microarrays were printed from a murine clone library available at St Jude Children's Research Hospital. Samples from reference and *Arf*-induced cell lines were taken at 0, 2, 4 and 8 h. At each time point, three independent replicates of cDNA microarray were generated. There were 5776 probe spots on each array. Only 2936 spots that passed a quality control of image analysis were used for differential expression analysis. The p-values used in our study were generated by Pounds and Cheng [20] where p-values were computed by permutation tests (see Panel A in Figure 5 for their histogram plot).

The second data set is the Estrogen data from Scholtens *et al.* and is described in the 'Estrogen 2×2 Factorial Design' vignette by Scholtens *et al.* [22]. The objective of the study was to

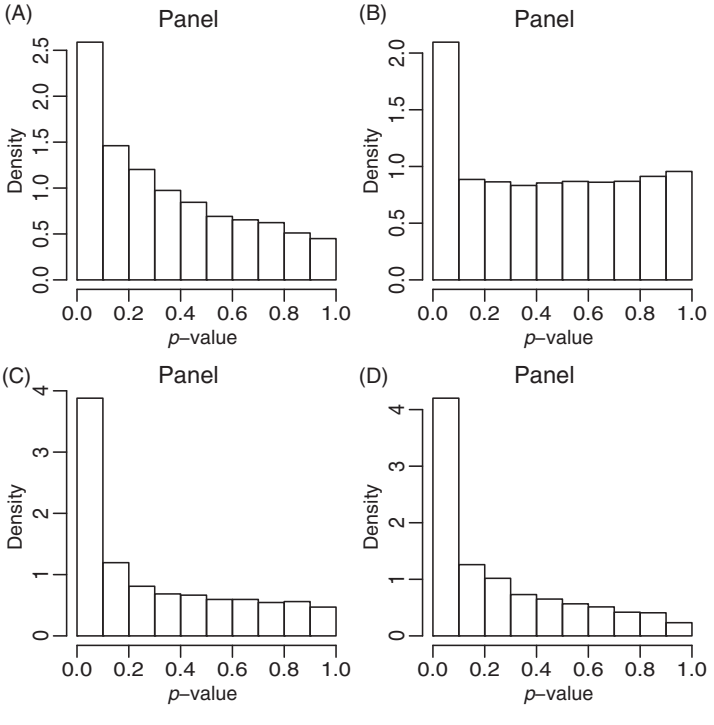

Figure 5. Histograms of p-values for the three data sets, where Panels A–D correspond to the p-values from Kuo *et al.*'s [12] data set, Scholtens *et al.*'s data set, Cui *et al.*'s data set with the CHQBC method, and Cui *et al.*'s data set with the TW method, respectively.

Table 4. Estimation of π_0 and the p-values for testing $H_0 : b = 0$ (inside parentheses) for different data sets.

			Cui *et al.* [5]	
	Kuo *et al.* [12]	Scholtens *et al.* [22]	CHQBC	TW
$\hat{\pi}_0^{\mathrm{b}}$	0.449	0.878	0.223	0.208
$\hat{\pi}_0^{\mathrm{s}}$	0.464	0.952	0.258	0.243
$\hat{\pi}_0^{\mathrm{c}}$	0.463	0.875	0.271	0.252
p-Value	($\sim 10^{-5}$)	(0.057)	($\sim 10^{-6}$)	($\sim 10^{-6}$)
$\hat{\pi}_{03}^{\mathrm{new}}$	0.727	0.875	0.622	0.616

investigate the effect of estrogen on the genes in ER+ breast cancer cells over time. The p-values of testing null hypothesis of no differential expression in the presence and absence of estrogen were used in our study (see Panel B in Figure 5 for their histogram plot).

The third data set is the cancer cell line experiment described by Cui *et al.* [5] (see Section 4 for more details). The p-values of testing differential expression among these cell lines are calculated by two different methods: Cui *et al.*'s [5] shrinkage test (CHQBC) and Tong and Wang's [30] optimal shrinkage test (TW). The p-value histograms for these two methods are also presented in Figure 5 (Panels C and D).

Table 4 reports the estimated value of π_0 for each data set by different methods. Note that the moment-based estimator, $\hat{\pi}_0^{\mathrm{m}}$, is not included due to insufficient sample sizes. For the data set 2, we observe a high degree of agreement among all estimators except that the smooth spline estimator, $\hat{\pi}_0^{\mathrm{s}}$, is somewhat larger. For the data sets 1 and 3, however, $\hat{\pi}_{03}^{\mathrm{new}}$ gives a larger estimate for π_0 than the other three methods, $\hat{\pi}_0^{\mathrm{b}}$, $\hat{\pi}_0^{\mathrm{s}}$ and $\hat{\pi}_0^{\mathrm{c}}$. The small p-values of both data sets for testing $H_0 : b = 0$ indicate that either $\hat{\pi}_{03}^{\mathrm{new}}$ provides an appropriate estimate for π_0, or it cautions that the data sets may contain many weakly differentially expressed genes.

6. Conclusion

The performance of existing methods for estimating π_0 may be unsatisfactory when the dependence among test statistics is strong, mainly due to the increase of variation in the estimation. This motivated us to propose new estimators for π_0 that aims to alleviate the impact of certain dependence. Our proposed method is simple and flexible. Simulation study indicates that our new estimators, especially for $\hat{\pi}_{03}^{\mathrm{new}}$, compare favorably with some of the existing methods under arbitrary dependence among test statistics. Specifically, $\hat{\pi}_{03}^{\mathrm{new}}$ takes advantages of both the convex density estimator by Langaas *et al.* [14] for the independent case and the proposed data-driven estimator for the dependent case. More importantly, our proposed method has a general structure and can be further improved easily. For example, we can replace $\hat{\pi}_0^c$ in Equation (7) by a better estimator developed under the independence assumption, as well as replace $\hat{\pi}_{0i}(\lambda)$ by another better estimator developed under the dependence assumption whenever they are available.

Note that the proposed method fits the observed p-values in $(\lambda, 1]$ by a simple linear curve. Hence, it may be still far from perfect. For instance, as mentioned in Section 4, the proposed estimators may be very conservative when there are many weakly differentially expressed genes with p-values larger than 0.5. Further research is warranted to propose better estimation in this direction. One way to overcome this problem is to fit the model by weighted least squares which assign proper increasing weights to N_i's to reduce the impact from these weakly differentially expressed genes. Other alternatives can be to fit a non-parametric regression model or to fit a natural spline curve to $\hat{\pi}_{0i}(\lambda)$ on the search set $\mathcal{R}$ as in [27].

Acknowledgements

Tiejun Tong's research was supported by Hong Kong RGC grant HKBU202711 and Hong Kong Baptist University FRG grants FRG2/10-11/020 and FRG2/11-12/110. Zeny Feng's research was supported by Natural Sciences and Engineering Research Council of Canada individual discovery grant. Hongyu Zhao's research was supported by NIH grant GM59507 and NSF grant DMS0714817. The authors thank Dr Stan Pounds and Dr Cheng Cheng from St Jude Children's Research Hospital for providing the data set and helpful comments. The authors also thank the editor, the associate editor, and two reviewers for their constructive comments that have substantially improved the paper.

References

[1] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, J. R. Stat. Soc. Ser. B 57 (1995), pp. 289–300.
[2] P. Broberg, *A comparative review of estimates of the proportion unchanged genes and the false discovery rate*, BMC Bioinform. 6 (2005), Article no. 199.
[3] J. Cabrera and C. Yu, *Estimating the proportion of differentially expressed genes in comparative DNA microarray experiments*, IMS Lecture Notes Monogr. Ser. 54 (2007), pp. 92–102.
[4] A. Celisse and S. Robin, *A cross-validation based estimation of the proportion of true null hypotheses*, J. Statist. Plann. Inference 140 (2010), pp. 3132–3147.
[5] X. Cui, J.T.G. Hwang, J. Qiu, N.J. Blades, and G.A. Churchill, *Improved statistical tests for differential gene expression by shrinking variance components estimates*, Biostatistics 6 (2005), pp. 59–75.
[6] C. Dalmasso, P. Broet, and T. Moreau, *A simple procedure for estimating the false discovery rate*, Bioinformatics 21 (2005), pp. 660–668.
[7] H. Finner and V. Gontscharuk, *Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses*, J. R. Stat. Soc. Ser. B 71 (2009), pp. 1031–1048.
[8] C. Friguet and D. Causeur, *Estimation of the proportion of true null hypotheses in high-dimensional data under dependence*, Comput. Statist. Data Anal. 55 (2011), pp. 2665–2676.
[9] Y. Hochberg and Y. Benjamini, *More powerful procedures for multiple significance testing*, Stat. Med. 9 (1990), pp. 811–818.
[10] H.M. Hsueh, J.J. Chen, and R.L. Kodell, *Comparison of methods for estimating the number of true null hypotheses in multiplicity testing*, J. Biopharm. Statist. 13 (2003), pp. 675–689.
[11] H. Jiang and R.W. Doerge, *Estimating the proportion of true null hypotheses for multiple comparisons*, Cancer Inf. 6 (2008), pp. 25–32.
[12] M. Kuo, E.J. Duncavage, R. Mathew, W. den Besten, D. Pei, D. Naeve, T. Yamamoto, C. Cheng, C.J. Sherr, and M.F. Roussel, *Arf induces p53-dependent and -independent antiproliferative genes*, Cancer Res. 63 (2003), pp. 1046–1053.
[13] Y. Lai, *A moment-based method for estimating the proportion of true null hypotheses and its application to microarray gene expression data*, Biostatistics 8 (2007), pp. 744–755.
[14] M. Langaas, B.H. Lindqvist, and E. Ferkingstad, *Estimating the proportion of true null hypotheses, with application to DNA microarray data*, J. R. Stat. Soc. Ser. B 67 (2005), pp. 555–572.
[15] X. Lu and D.L. Perkins, *Re-sampling strategy to improve the estimation of number of null hypotheses in FDR control under strong correlation structures*, BMC Bioinform. 8 (2007), Article no. 157.
[16] M.O. Mosig, E. Lipkin, K. Galina, E. Tchourzyna, M. Soller, and A. Friedmann, *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 (2001), pp. 1683–1698.
[17] S. Nadarajah and S. Kotz, *Multitude of bivariate t distributions*, Statistics 38 (2004), pp. 527–539.
[18] D. Nettleton, J.T.G. Hwang, R.A. Caldo, and R.P. Wise, *Estimating the number of true null hypotheses from a histogram of p-values*, J. Agric. Biol. Environ. Stat. 11 (2006), pp. 337–356.
[19] D.V. Nguyen, *On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies*, Comput. Statist. Data Anal. 47 (2004), pp. 611–637.
[20] S. Pounds and C. Cheng, *Improving false discovery rate estimation*, Bioinformatics 20 (2004), pp. 1737–1745.
[21] D. Ruppert, D. Nettleton, and J.T.G. Hwang, *Exploring the information in p-values for the analysis and planning of multiple-test experiments*, BMC Bioinform. 63 (2007), pp. 483–495.
[22] D. Scholtens, A. Miron, F.M. Merchant, A. Miller, P.L. Miron, J.D. Iglehart, and R. Gentleman, *Analyzing factorial designed microarray experiments*, J. Multivariate Anal. 90 (2004), pp. 19–43.
[23] T. Schweder and E. Spjøtvoll, *Plots of p-values to evaluate many tests simultaneously*, Biometrika 69 (1982), pp. 493–502.
[24] C. Sherr, *Tumor surveillance via the ARF-p53 pathway*, Genes Dev. 12 (1998), pp. 2984–2991.
[25] M.M. Siddiqui, *A bivariate t distribution*, Ann. Math. Stat. 38 (1967), pp. 162–166.
[26] J.D. Storey, *A direct approach to false discovery rates*, J. R. Stat. Soc. Ser. B 64 (2002), pp. 479–498.

[27] J.D. Storey and R. Tibshirani, *SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays*, in *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger, eds., Springer, New York, 2003, pp. 272–290.
[28] J.D. Storey, J.E. Taylor, and D. Siegmund, *Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rate: A unified approach*, J. R. Stat. Soc. Ser. B 66 (2004), pp. 187–205.
[29] A.C. Tamhane and J. Shi, *Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR*, IMS Lecture Notes Monogr. Ser. 57 (2009), pp. 304–325.
[30] T. Tong and Y. Wang, *Optimal shrinkage estimation of variances with applications to microarray data analysis*, J. Amer. Statist. Assoc. 102 (2007), pp. 113–122.
[31] F.E. Turkheimer, C.B. Smith, and K. Schmidt, *Estimation of the number of 'true' null hypotheses in multivariate analysis of neuroimaging data*, NeuroImage 13 (2001), pp. 920–930.
[32] H. Wang, K.L. Tuominen, and C. Tsai, *Slim: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures*, Bioinformatics 27 (2011), pp. 225–231.
[33] B. Wu, Z. Guan, and H. Zhao, *Parametric and nonparametric FDR estimation revisited*, Biometrics 62 (2006), pp. 735–744.

Appendix. Proof of Lemma 1

For a given p_i, we have $Z_i = z_{1-p_i}$ and

$$Z_j|_{Z_i=z_{1-p_i}} \sim N(\rho z_{1-p_i}, 1-\rho^2).$$

Then, for any $0 \le \alpha \le 1$,

$$\begin{aligned} P(p_j \le \alpha|p_i) &= P(Z_j \ge z_{1-\alpha}|Z_i = z_{1-p_i}) \\ &= 1 - P\left(\frac{Z_j - \rho z_{1-p_i}}{\sqrt{1-\rho^2}} \le \frac{z_{1-\alpha} - \rho z_{1-p_i}}{\sqrt{1-\rho^2}} \,\middle|\, Z_i = z_{1-p_i}\right) \\ &= 1 - \Phi\left(\frac{z_{1-\alpha} - \rho z_{1-p_i}}{\sqrt{1-\rho^2}}\right) \\ &= \Phi\left(\frac{z_\alpha + \rho z_{1-p_i}}{\sqrt{1-\rho^2}}\right). \end{aligned}$$

Thus, the conditional probability density function $f(p_j \mid p_i)$ is given as

$$\begin{aligned} f(p_j|p_i) &= \left.\frac{\mathrm{d}P(p_j \le \alpha|p_i)}{\mathrm{d}\alpha}\right|_{\alpha=p_j} \\ &= \phi\left(\frac{z_{p_j} + \rho z_{1-p_i}}{\sqrt{1-\rho^2}}\right)(\sqrt{1-\rho^2}\phi(z_{p_j}))^{-1}, \end{aligned}$$

which leads to Equation (2) with some simple calculation.