Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Dialsingh and Cedeno, J Biom Biostat 2017, 8:2
DOI: 10.4172/2155-6180.1000343

Research Article Open Access

Comparison of Methods for Estimating the Proportion of Null Hypotheses π_0 in High Dimensional Data When the Test Statistics is Continuous

Isaac Dialsingh* and Sherwin P Cedeno

Department of Mathematics and Statistics, University of the West Indies, St. Augustine Campus, Trinidad and Tobago

Abstract

Advances in Genomics have re-energized interest in multiple hypothesis testing procedures but have simultaneously created new methodological and computational challenges. In Genomics for instance, it is now commonplace for experiments to measure expression levels in thousands of genes creating large multiplicity problems when thousands of hypotheses are to be tested simultaneously. Within this context we seek to identify differentially expressed genes, that is, genes whose expression levels are associated with a particular response or covariate of interest. The False Discovery Rate (FDR) is the preferred measure since the Family Wise Error Rates (FWERs) are usually overly restrictive. In the FDR methods, estimation of the proportion of null hypotheses (π0) is an important parameter that needs to be estimated.

In this paper, we compare the effectiveness of 12 methods for estimating π0 when the test statistics are continuous using simulated data with independent, weak dependence, and moderate dependence structures.

Keywords: Family-wise error rate; False discovery rate; High dimensional data; Hypothesis testing; Multiple hypothesis testing; Proportion of null hypotheses

Introduction

A hypothesis is a claim or assertion either about a population parameter. In the case of a single hypothesis, we typically test the null hypothesis H_0 versus an alternative hypothesis H_a based on some test statistic. We reject H_0 in favor of H_a whenever the test statistic lies in the rejection region specified by some rejection rule. The test statistic is a function of the sample data on which the decision to reject or not to reject H_0 will be based. Within the realm of hypothesis testing, there are two possible types of errors that can be committed. A Type I error, is committed if we reject H_0 when H_0 is true while a Type II error occurs when we fail to reject H_0 when H_0 is false, Table 1 summarizes the error possibilities.

The problem associated with single hypothesis testing is compounded in the realm of multiple hypothesis testing. Multiple hypothesis testing is concerned with controlling the rate of false positives when we are testing multiple hypotheses simultaneously. When conducting multiple hypothesis tests, the probability of making at least one Type I error is substantially higher than the nominal level used for each test, particularly when the number of total tests, m, is large.

It is not unusual to have thousands of tests being done simultaneously. This implies that the probability of getting at least one significant result approaches 1. These methods of multiple hypothesis testing require an adjustment of α in some way so that the probability of observing at least one significant event, due largely to chance, remains below the chosen level of statistical significance.

Notation

Tests are typically assumed to be independent, or that the test statistics are independent and identically distributed, but there are methods and procedures that deal with cases of dependence. We assume that we are testing m independent null hypotheses, H_{01}, H_{02}, … ,H_{0m} with corresponding p-values p_1, p_2,.., p_m, and we call the i^{th} hypothesis "significant" if we reject the null hypothesis H_{0i}. Table 2 summarizes the possible configurations when testing m hypotheses simultaneously. In this table, V is the number of false rejections (or false discoveries), U is the number of true non-rejections (or true acceptances), S is the number of true rejections, and T is the number of false non-rejections. Here m_0, the total number of true null hypotheses, is fixed but unknown. Though random variables V, S, U, and T are not observable, the random variables R=S+V and W=U+T, the number of significant and insignificant tests, respectively, are observable. The proportion of false rejections is V/R when R>0 and the proportion of false acceptances is T/W when W>0.

	Declared True	Declared False
H_0 True	Correct Decision (1-α)	Type I Error (α)
H_0 False	Type II Error (β)	Correct Decision (1-β)

Table 1: Possible outcomes for a single hypothesis test.

Using this notation, the FWER rate which is the Probability of making at least one type I error among the m tests is given by:

$$FWER=Prob\ (V \geq 1) \text{ or } FWER=1 - Prob\ (V=0) \quad (1)$$

While the False Discovery Rate (FDR) which is defined as the expected proportion of false rejections is given by:

$$FDR=E(Q) \quad (2)$$

	Significant	Not Significant	Total
True Null	V	U	m_0
False Null	S	T	m_1
Total	R	W	m

Table 2: Possible outcomes for m hypothesis tests.

***Corresponding author:** Dialsingh I, Department of Mathematics and Statistics, University of the West Indies, St. Augustine Campus, Trinidad and Tobago, Tel: +1 876-927-1660; E-mail: isaac.dialsingh@sta.uwi.edu

Received April 1, 2017; **Accepted** April 24, 2017; **Published** April 28, 2017

Citation: Dialsingh I, Cedeno SP (2017) Comparison of Methods for Estimating the Proportion of Null Hypotheses π_0 in High Dimensional Data When the Test Statistics is Continuous. J Biom Biostat 8: 343. doi: 10.4172/2155-6180.1000343

$$\text{where } E(Q) = \begin{cases} \frac{V}{R} & for\ R > 0. \\ 0 & for\ R = 0. \end{cases} \quad (3)$$

Many methods have been created modifying the FWER concepts. A review of these techniques are found in [1,2]. The majority of these methods involves the use of step-up or step-down procedures [3-5]. Controlling FWER results in tests with low power. The FDR is less restrictive and allows a certain amount of false positives [6]. A number of methods have been established for estimation of the proportion of null hypotheses when the test statistics are discrete [7]. In this paper, we compare 12 methods for comparing false discovery rates when the test statistics are continuous under varying dependence structures.

Methods for controlling false discovery rate

Modern approaches in multiple hypothesis testing focus on False Discovery Rate (FDR) control [6], which is a simple step-up procedure using the ordered p-values of the tests. False Discovery Rate originated in two papers that dealt with multiple hypothesis testing. Schweder and Spjotvoll [8] first suggested that the ranked p-values be plotted. The assessment of the number of true null hypotheses m_0 would then be done via an eye fitted straight line starting from the largest p-values. The p-values that deviate (outliers) from this straight line would then correspond to the false null hypotheses. The density of the p-values can be expressed as:

$$f(p)=\pi_0 f_0(p) + (1- \pi_0) f_1(p) \quad (4)$$

Where π_0 and π_1 represent the proportion of p-values under the null density f_0 and under the alternative density f_1 respectively. For continuous tests, p-values are uniformly distributed on the interval (0,1). The distribution of the p-values under the alternative hypothesis is unknown. Methods for estimating π_0 when the test statistics are continuous have been developed by coupling the mixture model with the assumption that either the density of marginal p-values, f(p), or the density of p-values under the alternative, $f_1(p)$ is non-increasing.

The first procedure than controls FDR is the BH Algorithm [6]. To control FDR at level q, reject all null hypotheses where:

$$\left\{ H_{0(i)} : i \le \max\left(k : p_k \le \frac{i \cdot q}{m} \right) \right\} \quad (5)$$

It has been shown that when the test statistics are continuous and independent, this procedure controls the FDR at level $\pi_0 q$ where π_0 is the proportion of true null hypotheses. In this procedure, each of the m hypotheses are treated as null so in this case, π_0=1. There are many other methods that were created. We briefly describe 12 of these methods for which we do our comparisons on.

The Methods

Storey's Smoother Method

The proposed estimator is:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i | p_i \rangle \lambda, i = 1,2,\ldots, m\}}{m(1-\lambda)} \quad (6)$$

Where λ ϵ [0,1) and is called a tuning parameter [9]. Beyond 0.5 the histogram of the p-values is flat and this suggests that mostly null p-values are there. Taking λ=0 gives $\hat{\pi}_0(\lambda) = 1$ which is overly conservative and as λ→ 1 equation (6) above indicates that the variance of $\hat{\pi}_0(\lambda)$ increases and this makes the estimated q-values unreliable.

The general algorithm for estimating q-values from the p-values is:

1. Assume that we have m ordered p-values according to their evidence against the null hypotheses, i.e. $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$.

Use equation (6) and compute $\hat{\pi}_0(\lambda)$ for various values of λ.

Let f be a natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ.

Set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$.

Let t be a threshold where 0 < t ≤ 1 calculate:

$$\hat{q}(p_{(m)}) = \min_{t \ge p_{(m)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \le t\}} = \hat{\pi}_0 \cdot p_{(m)} \quad (7)$$

2. For i=m-1, m-2, …, 1 calculate:

$$\hat{q}(p_{(i)}) = \min_{t \ge p_{(i)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \le t\}} = \min\left(\frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right) \quad (8)$$

3. Equation (8) above gives the estimated q-value for the i^{th} most significant feature.

This method is referred to as "smoother" throughout this paper.

Storey's bootstrap method

Storey's bootstrap method [10] resamples the m ordered p-values $p_{(1)} \le p_{(2)} \le \ldots \le p_{(m)}$ with replacement and creates pseudo-data sets for some number of bootstrap resamples B. The p-values are then calculated for each pseudo-data set and a bootstrap estimate is done for the FDR. The method enables a counter to record whether the minimum p-value from the pseudo-data set is less than or equal to the actual p-value for each base test. For m tests there would be m such counters. This process is repeated a large number of times, say B, and the proportion of resampled data sets where the minimum pseudo p-values is less than or equal to an actual p-value is the adjusted p-value. This method has the ability to incorporate all sources of correlation from both the multiple contrasts and the multivariate structure. The resulting adjusted p-values incorporate all correlations and distributional characteristics. The bootstrap method provides strong control when the joint distribution of the p-values for any subset of the null hypotheses is identical to that under the complete null hypothesis, that is, when the subset pivotality condition holds. It always provides weak control of the FWER. This method is referred to as "st.boot" throughout this paper.

Two-step procedure of Jiang and Doerge

This two-step procedure proposed by Jiang and Doerge [11] increase the power of detecting differentially expressed genes in microarray data. The null hypothesis of equality across the mean expression levels for all treatments is tested in step one. In the second step pairwise comparisons are done on genes for which the treatment means were shown to be statistically significant in step one. This approach estimates the overall FDR used in both steps in such a way that the overall FDR is controlled below a pre-specified FDR significance level. The two-step approach has increased power over a one-step procedure and it controls the FDR at the desired level of significance. The algorithm for this two-step procedure is as follows:

1. Test the null hypothesis that a gene is not differentially expressed across each treatment condition. This can be done using the global F-test from an ANOVA model for instance. For m tests corresponding to m genes an FDR control procedure is applied to control FDR at the α_1 level. Tests that produce p-values ≤ some specified

c_1 are considered to be statistically significant. If we get L significant tests and M genes are declared to have statistically significant treatment effects then we go to step 2. However if L=0 then we stop and conclude that there are no statistically significant pairwise comparisons and therefore no differentially expressed genes.

2. For genes that form collection M we would perform N pairwise comparisons for each gene. We would apply an FDR control procedure at the α_2 level to these L*N tests. Any comparisons found that produce p-values ≤ some specified c_2 are considered to be statistically significant.

It should be noted that this procedure assumes that genes that show no significant treatment effect (step 1) will also show no statistically significant pairwise comparisons (step 2). However, a gene having a significant treatment effect may or may not have statistical significance in pairwise comparisons. This method is referred to as "jiang" throughout this paper.

Nettleton's Histogram method

Nettleton's Method [12] estimates π_0 by estimating the proportion of the observed p-values that follow a uniform distribution. The steps in the algorithm are as follows:

- Partition the interval [0,1] in B bins of equal width.

Assume all null hypotheses are true, and set $m_0^{(0)} = \pi_0^{(0)}(m) = m.$

Calculate the expected number of p-values for each bin given the current estimate of the number of true null hypotheses.

Beginning with the leftmost bin, sum the number of p-values in excess of the expected until a bin with no excess is reached.

Use the excess sum as an updated estimate of m_1, and then use that to update the estimate of m_0=m - m_1.

Return to Step (iii) and repeat the procedure until convergence is reached.

The number of bins is used as a tuning parameter and the authors suggested using B=20. In their simulation study the histogram based estimator with 20 bins tended to be conservative for independent and autoregressive correlation structures in that it tended to overestimate the number of true null hypotheses for certain correlations that was considered in their simulation study. This method is referred to as "histo" throughout this paper.

Convex decreasing p-value estimate (Langaas)

Langass and Landqvist's estimator is based on the non-parametric MLE of the p-value density [13]. The authors restricted their attention to decreasing and convex decreasing densities because of their many mathematically attractive properties. The derived estimator of π_0 assumed independence of the test statistics. These estimators were found to be robust with respect to the independence assumption but also worked well for test statistics that showed moderate levels of dependence. The aim is to derive a NPMLE (non-parametric maximum likelihood estimate) for a convex decreasing density on [0, 1). The mixture density given in (4) is modified by requiring f(p) be a convex decreasing function with f(1)=0. The algorithm is as follows:

Specify a convex decreasing initial function $\hat{f}(0)$.

For j=0, 1, 2, … given the current iterate $\hat{f}_j$ determine $\hat{\theta}$ using:

$$\hat{\theta} = \arg\min_{\theta\epsilon[0,1]}\left\{D_\varphi\left(f_\theta - \hat{f}\right);\hat{f}\right\} = \arg\min_{\theta\epsilon[0,1]}\left\{\sum{}'\left(\frac{\hat{f}(p_i) - f_\theta(p_i)}{\hat{f}(p_i)}\right)\right\} \quad (9)$$

Where $\hat{f}_\cdot$ replaces $\hat{f}_\cdot$

If $\left\{D_{\hat{\theta}}\left(f_{\hat{\theta}} - \hat{f}_j\right);\hat{f}_j\right\} \geq 0$ then the current iterate $\hat{f}_j$ is optimal by $\left\{D_{\hat{\theta}}\left(f_{\hat{\theta}} \quad \hat{f}\right);\hat{f}\right\}$ and we are done.

If this optimality is not found then the next iterate is:

$$\hat{f}_{j+1} = (1+\hat{\varepsilon})\hat{f}_j + \hat{\varepsilon}\hat{f}_{\hat{\theta}} \quad (10)$$

Where

$$\hat{\varepsilon} = \arg\min_{\varepsilon\epsilon[0,1)}\left[\varphi\left\{(1-\varepsilon)\hat{f}_j + \varepsilon\hat{f}_{\hat{\theta}}\right\}\right] = \arg\min_{\varepsilon\epsilon[0,1)}\left[-\sum{}'\log\left\{(1-\varepsilon)\hat{f}_j(p_i) + \varepsilon f_{\hat{\theta}}(p_i)\right\}\right]$$

The procedure employed by this algorithm is similar to the "steepest descent" algorithm used to optimize a function on the Euclidean n-space $\Re^n$. The next iterate in each step of the algorithm is the optimal convex combination of the current iterate and the mixing density $\hat{f}_{\hat{\theta}}$ that corresponds to the most negative directional derivative. This method is referred to as "langaas" throughout this paper.

Robust estimation (Pounds)

Pounds and Cheng's estimator of π_0 [14] when the test statistics are continuous is given by:

$$\hat{\pi}_0 = \begin{cases} \min(1, 2\bar{p}) & \text{for two sided tests} \\ \min(1, 2\bar{t}) & \textit{for one sided tests} \end{cases} \quad (11)$$

Where

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m} p_i \quad (12)$$

is the average p-value for all m hypotheses,

$$\bar{t} = \frac{1}{m}\sum_{i=1}^{m}[2\cdot\min(p_i, 1-p_i] \quad (13)$$

and min(a,b) is the minimum of a and b. This estimator is biased upward but the bias is small when $pf_1(p)$ is small or when π_0 is close to 1. This method is referred to as "pounds" throughout this paper.

Lowest slope method

This stepwise procedure was developed in ref. [15] first estimates the value of m_0 using the data and this estimate is used in the adaptive procedure itemized in the algorithm that follows:

i. Order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ corresponding to null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$.

For level q and i=1, 2, …., k we compare each $p_{(i)}$ with $\frac{iq}{m}$. If no $p_{(i)}$ is found to be smaller we do not reject any null hypotheses and stop.

Compute the slopes $S_i = \frac{1 - p_{(i)}}{m+1-i}$

Starting with i=1 proceed as long as $S_i \geq S_{i-1}$. When we find $S_j < S_{j-1}$ stop.

Set $\hat{m}_0 = \min\left(\left[\frac{1}{S_j}+1\right], m\right)$ (14)

Then starting with the largest p-value $p_{(m)}$, compare each $p_{(i)}$ with $\frac{iq}{\hat{m}_0}$ until we arrive at the first p-value satisfying $p_{(k)} \leq \frac{kq}{\hat{m}_0}$. We

would reject all k hypotheses whose p values are smaller than $p_{(k)}$.

This method is referred to as "abh" throughout this paper.

Sliding linear method

A sliding linear model (SLIM) to estimate π_0 in datasets with dependence structures [16]. Whenever undetected dependence structures exist they distort the distribution of the p-values making any estimate of π_0 less effective. The SLIM method of FDR estimation is based on a linear model transformed from the non-linear λ estimator of Storey. The method functions by partitioning the data into local dependence blocks. This data partitioning and optimization allows SLIM to utilize information from a broader range of p-value distributions for π_0 estimation. The employed optimization scheme exploits the non-static relationship between the p-values and the q-values by minimizing the difference between the fractions of tests called significant by the p-value and q-value methods. SLIM handles the hidden dependence in the data without the need to empirically adjust the null p-value distributions. This method is referred to as "SLIM" throughout this paper.

SPLOSH method

The Spacings LOESS histogram (SPLOSH) estimates the conditional false discovery rate (cFDR) [17]. This cFDR is the expected proportion of false positives conditioned on *k* "significant" findings. SPLOSH is designed to be more stable than Storey's approach in that the q-value is based on an unstable estimator of the positive false discovery rate (pFDR). SPLOSH is applicable in a wider variety of settings than BUM. The SPLOSH algorithm is as follows:

Let $p_{(1)} \le p_{(2)} \le \ldots \le p_{(k)}$ be k ordered p-values. Let $a_{(i)} = \frac{\left(i - \frac{1}{2}\right)}{k}$ be their adjusted ranks to control the type I error rate.

Compute the midpoint of the interval $\left[\tilde{p}_{(j)}, \tilde{p}_{(j+1)}\right]$ using:

$$m_{(j)} = \frac{\tilde{p}_{(j+1)} + \tilde{p}_{(j)}}{2} \tag{14}$$

1. Apply the arc-sine transform for *i*=1,2,…,*k* to the ranked p-values using:

$$x_i = \arcsin\left[2 \times \left(p_{(i)} - \frac{1}{2}\right)\right] \tag{15}$$

Apply LOWESS (LOcally WEighted Scatter-plot Smoother) to $\left(\tilde{x}_{(j)}, \tilde{y}_{(j)}\right)$ for *j*=1,…,*u*-1to obtain an estimated curve $\hat{f}^*(\tilde{p}_{(j)}) = e^{\hat{y}(\tilde{x}_{(j)})}$

For *j*=1,…,*u* to obtain the estimated derivative of the CDF $\hat{f}^*(\tilde{p}_{(j)}) = e^{\hat{y}(\tilde{x}_{(j)})}$ up to a unitizing constant c.

Let $\hat{f}\left(p_{(i)}\right) = 1 \Big/ c\hat{f}^*(p_{(i)})$ be an estimate of the PDF at p_i for *i*=1,2,…,k. The trapezoidal rule is used to estimate c using:

$$c = \frac{1}{2}\sum_{j=l}^{u-1}\left[\hat{f}^*(\tilde{p}_{(j)}) + \hat{f}^*(\tilde{p}_{(j+1)})\right]\Delta j \tag{16}$$

where $\Delta j = \tilde{p}_{(j+1)} - \tilde{p}_{(j)}$ is the sub-interval width..

Let $\hat{F}(\tilde{p}_{(l)}) = 0$ be the CDF and for *k*=*l* + 1,..,*u* let

$$\hat{F}(\tilde{p}_{(k)}) = \frac{1}{2}\sum_{j=l}^{k-1}\left[\hat{f}(\tilde{p}_{(j)}) + \hat{f}(\tilde{p}_{(j+1)})\right]\Delta j \tag{17}$$

be an estimate of $\hat{F}(\tilde{p}_{(j)})$ obtained using the trapezoidal rule in approximate integration.

Take $\hat{\pi} = \min_{l \le i \le k} \hat{f}\left(p_{(i)}\right)$ as the minimum of the pdf.

For i=1, … , k we obtain $r_{(i)} \equiv \hat{r}\left(p_{(i)}\right)$ by substituting $p_{(i)}, \hat{\pi}, \hat{F}\left(p_{(i)}\right)$ into

$$\hat{r}(\alpha) = \frac{\hat{\pi}\alpha}{\hat{F}(\alpha)} \tag{18}$$

For p-values that are equal to zero we use L'Hôspital's Rule to justify $\frac{\hat{\pi}}{\hat{f}(0)}$ as an estimate of the cFDR:

$$\lim_{\alpha \to 0}\hat{r}(\alpha) = \lim_{\alpha \to 0}\frac{\hat{\pi}\alpha}{\hat{F}(\alpha)} = \lim_{\alpha \to 0}\frac{\hat{\pi}}{f(\alpha)} \tag{19}$$

Define a monotone quantity $h_{(i)} = \min_{k \ge i}\left(r_{(k)}\right)$ based on the cFDR estimates $r_{(i)}$ for i=1, … , k.

This method is referred to as "splosh" throughout this paper.

Beta-Uniform Mixture (BUM) Model

Mixture modeling of the p-value distribution was first proposed in ref. [18]. This method models the p-value distribution as a mixture of a Uniform (0,1) distribution, corresponding to the true null hypotheses and a Beta(α,β) distribution corresponding to the false null hypotheses. The Beta-Uniform Mixture (BUM) Model was proposed by Pounds and Morris [19]. This is basically a mixture of a Uniform(0,1) distribution and a Beta(α,1) distribution. The Beta distribution is chosen because of its flexibility in modeling any distribution on the interval [0,1].

This mixture model assumes independence of gene expression levels across genes. Thus, under the null hypothesis the p-values are uniformly distributed on the interval [0,1] regardless of the statistical test being used, as long as the test is valid, and regardless of the size of the sample. Under the alternative hypothesis the distribution of the p-values tends to cluster closer to zero than one. By referring to the entire distribution of the p-values obtained in the sample we can then address whether there is statistically significant evidence that any of the genes under study exhibits a difference in expression across the groups. This is usually done by performing an omnibus test of whether the observed distribution of the p-values is significantly different from a uniform distribution. This method is referred to as "BUM" throughout this paper.

Grenander density estimator

The Grenander density estimator [20] is a non-parametric maximum likelihood estimator (NPMLE) similar to an empirical cumulative density function (ECDF). Unlike the ECDF it has the added constraint of an underlying density. The Grenander estimator uses the ECDF as an estimator of the associated distribution function F(p). This estimator is the decreasing piecewise-constant function equal to the slopes of the least concave majorant (LCM) of the ECDF. Monotone regression with weights is used to compute the LCM of the ECDF in order to obtain this estimator. If x_i and γ_i are used to denote coordinates for the ECDF and $\Delta x_i = x_{i+1} - x_i$ and $\Delta\gamma_i = \gamma_{i+1} - \gamma_i$, then the slopes of the LCM are given by antitonic regression and the raw slopes Δx_i and $\Delta\gamma_i$ with weight Δx_i [21]. This method is referred to as "strimmer" throughout this paper.

Two stage adaptive procedure

Adaptive procedures first estimate the number of null hypotheses

m_0 and then use this estimate to revise a multiple test procedure. Knowledge of m_0 can be used to improve upon the performance of the FDR controlling procedure. The two stage adaptive procedure makes use of a linear step-up procedure making use of m ordered p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Let $k = max\left(i : p_{(i)} \leq \frac{i \cdot q}{m}\right)$. If this k exists then we reject all the k hypotheses associated with $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$, otherwise do not reject any of the hypotheses. If m_0 were known, then the linear set-up procedure with:

$$q' = \frac{qm}{m_0} \tag{20}$$

would control the FDR at precisely the desired level q in the independent and continuous case, and would then be more powerful in rejecting hypotheses for which the alternative holds. The adaptive Benjamini-Hochberg procedure [6] at the level q is as follows:

Compute $\widehat{m_0}$

If $\widehat{m_0} = 0$ reject all hypotheses; otherwise, test the hypotheses using the linear set-up procedure at level $\frac{qm}{\widehat{m_0}}$.

1. Use the linear set-up procedure at level q, and if no hypothesis is rejected stop; otherwise, proceed.

Estimate $m_0(k)$ using $\frac{m+1-k}{1-p_{(k)}}$

Starting with k=2 stop when for the first time $m_0(k) > m_0(k-1)$.

Estimate $\widehat{m_0} = \min\left(m_0(k), m\right)$ rounding up to the next highest integer.

Use the linear step-up procedure with $q^* = \frac{qm}{\widehat{m_0}}$.

This method is referred to as "TSBKY" throughout this paper.

Results and Discussion

The authors conducted a simulation study to compare the twelve (12) methods for estimating the proportion of null hypotheses π_0 under varying levels of dependence. Simulations were drawn from a MVN (μ, Σ) distribution where μ is the mean vector and Σ is the covariance matrix using the R library MASS. The study involved simulating the expression values of 1000 "genes" from 20 samples. The first 10 samples for each "simulated expression value" came from the case subjects and the last 10 samples came from the control subjects. The simulations were done for π_0 values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, and 1.00. The data was simulated for three cases of dependency within the genes:

i. Independent Genes - No dependence structure within or between the genes. In this case the leading diagonal of **Σ** contained all 1's and the off-diagonal entries, denoted as ρ, were all zero. A 1000 x 12 matrix of p-values was generated for each value of π_0. Each column of the matrix would represent 1000 p-values extracted from each of the "m" matrices. This procedure was replicated 1000 times. The p-value matrices were then used as the input test statistics to compute the row means and standard deviations for each π_0 estimation method being compared. This data was then bound into one matrix for each estimation method for plotting.

ii. Weak Dependence - In this case the leading diagonal of Σ contained all 1's and the off-diagonal entries, denoted as ρ, were all 0.1. The simulation was done in the same manner as for independence.

iii. Moderate Dependence - In this case the leading diagonal of Σ contained all 1's and the off-diagonal entries, denoted as ρ, were all 0.5.

All of the simulations were done using R version 3.2.3. Source codes for most of the estimation methods were available freely online.

In Figure 1 the solid black line that runs from bottom left to top right represents "the truth", meaning that if the true value of π_0 is 0.10 then the estimated value $\widehat{\pi_0}$ should correspond to 0.10. Departures either above or below this 45° line would allow one to make graphical inferences about the performance of the proposed estimator. This graphical approach has the advantage of allowing for easy visualization of the data and identification of possible trends and patterns that it contains that might not be readily obvious with tabular approaches. All of the methods compared tended to overestimate $\widehat{\pi_0}$ for low values of π_0. The estimators all performed better as π_0 increased. This is well documented in literature. "SLIM" performed very well for small values of π_0 but tended to underestimate as π_0 increased. It was found that "smoother", "st.boot", "Langaas" and "histo" all performed very well for most values of π_0. "TSBKY" and "abh" overestimated $\widehat{\pi_0}$ for every value of π_0. For example, Table 3 shows that when π_0 was 0.30 "TSBKY" gave an estimate of $\widehat{\pi_0} = 0.805$ with an estimated standard deviation of 0.022 while "abh" gave an estimate of $\widehat{\pi_0} = 0.607$ with an estimated standard deviation of 0.047.

Figure 2 shows the comparison of all twelve methods for weakly dependent test statistics. Most of the estimators overestimated $\widehat{\pi_0}$ for small values of π_0, however, "TSBKY" and "abh" significantly overestimated $\widehat{\pi_0}$ for most values of π_0. For example, Table 4 shows that when π_0 was 0.30 "TSBKY" gave an estimate of $\widehat{\pi_0} = 0.814$ with an estimated standard deviation of 0.125 while "abh" gave an estimate of $\widehat{\pi_0} = 0.619$ with an estimated standard deviation of 0.116. For weak dependent test statistics "BUM" initially overestimated $\widehat{\pi_0}$ but performed better for $\pi_0 \geq 0.60$. With weak dependence in the test statistics "smoother", "st.boot", "Langaas" and "histo" all performed very well for most values of π_0.

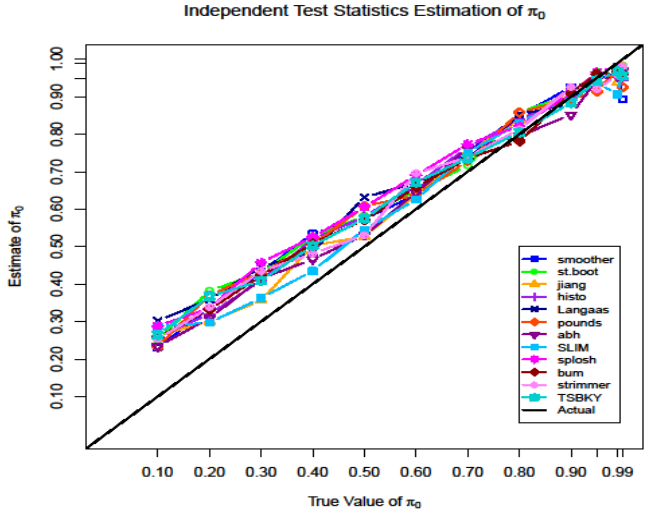

Figure 1: Comparison of π_0 estimation methods using independent test statistics.

π_0	1	2	3	4	5	6	7	8	9	10	11	12
0.1												
a	0.165	0.178	0.184	0.189	0.174	0.311	0.420	0.142	0.222	0.301	0.176	0.695
b	0.032	0.023	0.024	0.028	0.023	0.013	0.063	0.029	0.040	0.050	0.022	0.020
0.2												
a	0.259	0.247	0.275	0.277	0.261	0.387	0.515	0.218	0.319	0.326	0.263	0.754
b	0.047	0.032	0.030	0.033	0.030	0.015	0.058	0.033	0.046	0.074	0.028	0.025
0.3												
a	0.349	0.356	0.366	0.371	0.351	0.464	0.607	0.301	0.403	0.356	0.352	0.805
b	0.056	0.035	0.034	0.037	0.034	0.014	0.047	0.037	0.053	0.007	0.032	0.022
0.4												
a	0.433	0.441	0.455	0.463	0.438	0.541	0.686	0.379	0.481	0.393	0.438	0.858
b	0.069	0.050	0.042	0.039	0.046	0.018	0.039	0.046	0.064	0.010	0.042	0.020
0.5												
a	0.536	0.532	0.553	0.561	0.534	0.617	0.769	0.454	0.548	0.440	0.532	0.900
b	0.076	0.042	0.037	0.031	0.034	0.016	0.039	0.047	0.071	0.014	0.034	0.018
0.6												
a	0.625	0.628	0.644	0.653	0.628	0.696	0.832	0.545	0.620	0.534	0.622	0.939
b	0.072	0.038	0.034	0.029	0.034	0.016	0.031	0.039	0.071	0.020	0.034	0.019
0.7												
a	0.709	0.713	0.736	0.750	0.718	0.774	0.887	0.620	0.682	0.642	0.708	0.972
b	0.074	0.039	0.031	0.026	0.032	0.016	0.022	0.046	0.073	0.022	0.036	0.013
0.8												
a	0.816	0.806	0.829	0.835	0.811	0.851	0.942	0.695	0.771	0.775	0.801	0.989
b	0.080	0.043	0.034	0.026	0.032	0.016	0.015	0.053	0.079	0.021	0.040	0.007
0.9												
a	0.905	0.892	0.914	0.918	0.899	0.925	0.978	0.771	0.834	0.869	0.887	0.997
b	0.071	0.036	0.030	0.018	0.029	0.017	0.009	0.042	0.074	0.028	0.034	0.003
0.95												
a	0.941	0.937	0.958	0.957	0.945	0.963	0.993	0.809	0.872	0.998	0.931	0.999
b	0.062	0.039	0.027	0.018	0.026	0.017	0.005	0.049	0.076	0.014	0.032	0.001
0.99												
a	0.960	0.968	0.984	0.991	0.976	0.989	0.999	0.840	0.877	1.000	0.961	0.999
b	0.059	0.039	0.020	0.013	0.024	0.012	0.001	0.046	0.085	0.000	0.034	0.001
1.00												
a	0.967	0.976	0.990	0.996	0.986	0.994	0.999	0.835	0.864	1.000	0.973	0.999
b	0.051	0.034	0.017	0.009	0.020	0.009	0.001	0.046	0.086	0.000	0.031	0.001

Table 3: Mean (a) and Standard Deviation (b) for compared estimation methods using independent test statistics.

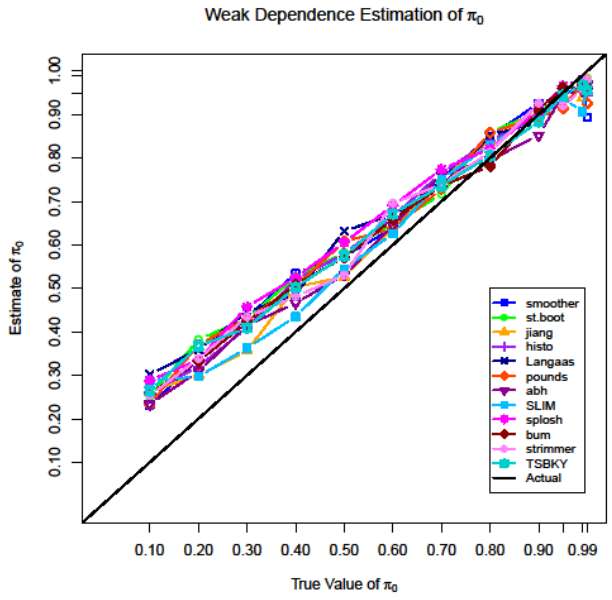

Figure 2: Comparison of π_0 estimation methods using weak dependent test statistics.

π_0	1	2	3	4	5	6	7	8	9	10	11	12
0.1												
a	0.148	0.164	0.173	0.322	0.163	0.301	0.395	0.133	0.203	0.302	0.166	0.685
b	0.055	0.053	0.050	1.411	0.050	0.069	0.133	0.049	0.058	0.043	0.051	0.140
0.2												
a	0.255	0.267	0.275	0.280	0.264	0.391	0.530	0.225	0.319	0.338	0.267	0.760
b	0.070	0.064	0.065	0.068	0.062	0.083	0.156	0.063	0.067	0.061	0.063	0.153
0.3												
a	0.355	0.359	0.370	0.374	0.353	0.468	0.619	0.300	0.404	0.368	0.356	0.814
b	0.069	0.050	0.052	0.055	0.048	0.063	0.116	0.052	0.059	0.055	0.049	0.125
0.4												
a	0.436	0.443	0.457	0.467	0.441	0.540	0.689	0.381	0.479	0.339	0.442	0.852
b	0.074	0.057	0.052	0.050	0.053	0.059	0.104	0.058	0.074	0.058	0.055	0.106
0.5												
a	0.530	0.536	0.553	0.563	0.537	0.622	0.778	0.459	0.556	0.452	0.532	0.903
b	0.076	0.054	0.051	0.050	0.051	0.048	0.088	0.055	0.074	0.057	0.052	0.077
0.6												
a	0.631	0.630	0.650	0.651	0.632	0.699	0.838	0.542	0.637	0.537	0.628	0.943
b	0.079	0.057	0.050	0.049	0.049	0.043	0.064	0.061	0.078	0.050	0.051	0.054
0.7												
a	0.719	0.711	0.732	0.741	0.714	0.766	0.882	0.611	0.694	0.630	0.705	0.959
b	0.087	0.051	0.045	0.043	0.045	0.039	0.056	0.062	0.079	0.048	0.047	0.041
0.8												
a	0.817	0.800	0.824	0.825	0.803	0.847	0.937	0.696	0.775	0.747	0.795	0.984
b	0.100	0.063	0.062	0.053	0.060	0.041	0.032	0.059	0.083	0.052	0.060	0.021
0.9												
a	0.895	0.887	0.912	0.912	0.894	0.926	0.976	0.766	0.823	0.889	0.885	0.996
b	0.109	0.080	0.075	0.062	0.074	0.045	0.019	0.069	0.097	0.084	0.075	0.007
0.95												
a	0.922	0.925	0.943	0.944	0.929	0.957	0.991	0.791	0.847	0.967	0.920	0.999
b	0.094	0.074	0.072	0.063	0.074	0.046	0.010	0.071	0.094	0.080	0.073	0.001
0.99												
a	0.957	0.963	0.973	0.979	0.969	0.982	0.999	0.821	0.837	0.988	0.959	0.999
b	0.074	0.063	0.055	0.050	0.059	0.037	0.005	0.060	0.091	0.055	0.063	0.001
1.00												
a	0.956	0.960	0.975	0.981	0.969	0.983	0.999	0.825	0.786	0.990	0.956	0.999
b	0.075	0.066	0.059	0.055	0.063	0.039	0.004	0.068	0.101	0.053	0.066	0.001

Table 4: Mean (a) and Standard Deviation (b) for compared estimation methods using weak dependent test statistics.

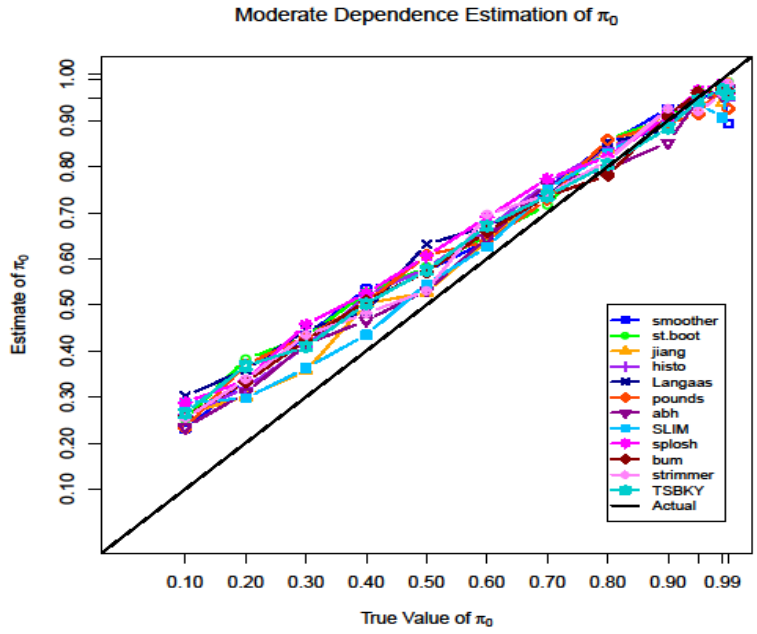

Figure 3: Comparison of π_0 estimation methods using moderately dependent test statistics.

π_0	1	2	3	4	5	6	7	8	9	10	11	12
0.1												
a	0.139	0.147	0.151	0.151	0.141	0.265	0.347	0.109	0.174	0.316	0.145	0.637
b	0.089	0.097	0.094	0.094	0.091	0.152	0.295	0.089	0.096	0.121	0.096	0.317
0.2												
a	0.114	0.116	0.120	5.382	0.114	0.174	0.238	0.331	0.130	0.170	0.116	0.328
b	0.150	0.150	0.155	52.60	0.146	0.225	0.331	0.130	0.170	0.222	0.150	0.416
0.3												
a	0.368	0.372	0.385	0.389	0.366	0.495	0.668	0.303	0.409	0.438	0.370	0.837
b	0.130	0.120	0.127	0.143	0.117	0.146	0.249	0.112	0.115	0.167	0.120	0.228
0.4												
a	0.040	0.041	0.042	0.043	0.040	0.057	0.080	0.036	0.042	0.051	0.042	0.089
b	0.124	0.125	0.128	0.131	0.122	0.180	0.250	0.109	0.130	0.161	0.127	0.273
0.5												
a	0.553	0.529	0.553	0.548	0.525	0.619	0.756	0.432	0.545	0.479	0.526	0.879
b	0.177	0.137	0.146	0.141	0.135	0.121	0.174	0.112	0.132	0.140	0.136	0.164
0.6												
a	0.382	0.368	0.386	0.378	0.368	0.425	0.496	0.304	0.352	0.342	0.370	0.559
b	0.348	0.324	0.341	0.330	0.324	0.352	0.409	0.267	0.305	0.288	0.325	0.452
0.7												
a	0.703	0.776	0.708	0.688	0.669	0.770	0.886	0.548	0.637	0.677	0.673	0.944
b	0.270	0.223	0.243	0.222	0.225	0.169	0.148	0.172	0.194	0.178	0.221	0.103
0.8												
a	0.738	0.724	0.757	0.738	0.717	0.826	0.922	0.567	0.653	0.763	0.723	0.971
b	0.294	0.261	0.279	0.258	0.262	0.193	0.113	0.194	0.202	0.195	0.259	0.071
0.9												
a	0.359	0.367	0.374	0.376	0.360	0.419	0.476	0.275	0.318	0.416	0.366	0.488
b	0.421	0.421	0.429	0.430	0.416	0.446	0.490	0.312	0.357	0.448	0.420	0.500
0.95												
a	0.784	0.797	0.813	0.826	0.796	0.881	0.963	0.591	0.617	0.886	0.796	0.990
b	0.291	0.280	0.280	0.284	0.285	0.192	0.124	0.209	0.212	0.200	0.279	0.062
0.99												
a	0.529	0.531	0.535	0.539	0.531	0.555	0.579	0.393	0.283	0.560	0.530	0.580
b	0.475	0.474	0.477	0.480	0.475	0.480	0.495	0.362	0.284	0.485	0.474	0.496
1.00												
a	0.547	0.554	0.560	0.564	0.554	0.591	0.629	0.381	0.258	0.596	0.553	0.639
b	0.458	0.461	0.463	0.468	0.462	0.462	0.479	0.332	0.245	0.467	0.460	0.482

Table 5: Mean (a) and Standard Deviation (b) for compared estimation methods using moderately dependent test statistics.

Figure 3 shows the comparison of all twelve methods for moderately dependent test statistics. Most of the estimators overestimated $\widehat{\pi_0}$ for small values of π_0, however, "TSBKY" and "abh" significantly overestimated $\widehat{\pi_0}$ for most values of π_0. For example, Table 5 shows that when π_0 was 0.30 "TSBKY" gave an estimate of $\widehat{\pi_0} = 0.837$ with an estimated standard deviation of 0.228 and "abh" gave an estimate of $\widehat{\pi_0} = 0.668$ with an estimated standard deviation of 0.249. In this case "smoother", "st.boot", and "jiang" performed better than most of the other estimators even though they all had a high standard deviation. A small standard deviation is highly desirable because this gives us an indication that the data points are packed very tightly around the mean value that we are estimating, so there is little spread in the data. The table shows that all of the methods appear to perform very badly for π_0=0.40, 0.60, 0.90, 0.99, and 1.00. This might be due to the methods ignoring the correlation effect among the test statistics.

Conclusion

A simulation study to compare 12 methods for estimating the proportion of null hypotheses under different configurations of dependence. Dependence among genes exists. This paper shows which of the methods would be most appropriate under various dependence configurations. The estimation of the proportion of null hypotheses is important especially when it comes to estimation of the false discovery rate. There is no doubt that this field will continue to evolve as datasets becomes larger, the demands on statistics becomes greater, and computing hardware and software improves.

References

1. Austin SR, Dialsingh I, Altman N (2014) Multiple Hypothesis Testing: A review. J Indian Soc of Agricultural Stat 68: 303-314.
2. Shaffer JP (1995) Multiple Hypothesis Testing. Annual Review of Psychology 46: 561-584.
3. Holm S (1979) A simple sequentially rejective multiple testing procedure. Scandanavian Journal of Statistics 6: 65-70.
4. Sidak Z, Sen PK, Hajek J (1999) Theory of Rank Tests.
5. Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751-754.
6. Benjamini Y, Hochberg Y (1995) Controlling the False discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Ser B 57: 289-300.
7. Dialsingh, I, Austin SR, Altman NS (2015) Estimating the proportion of true null hypotheses when the statistics are discrete. Bioinformatics.

8. Schweder T, Spjotvoll E (1982) Plots of p-values to evaluate many tests simultaneously Biometrika 69: 493-502.

9. Storey, John D, Robert T (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100: 9440-9445.

10. Storey, JD (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B 64: 479-498.

11. Jiang, Hongmei, Rebecca W, Doerge (2006) A Two-Step Multiple Comparison Procedure for a Large Number of Tests and Multiple Treatments. Statistical Applications in Genetics and Molecular Biology 5.

12. Nettleton D, JTG, Hwang R, Caldo A (2006) Estimating the number of true null hypotheses from a histogram of p-values. Journal of Agricultural, Biological and Environmental Statistics 11: 337-356.

13. Langaas, Mette, Bo Henry Lindqvist (2005) Estimating the proportion of true null hypotheses with application to DNA microarray data. Journal of the Royal Statistical Society B 67: 555-572.

14. Pounds S , Cheng C (2006) Robust estimation of the false discovery rate. Bioinformatics 22: 1979-1987.

15. Yoav B, Hochberg Y (2000) On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics.Journal of Educational and Behavioural Statistics 25: 60-83.

16. Wang HQ, Tuominen LK, Tsai CJ (2011) SLIM: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. Bioinformatics 27: 225-231.

17. Pounds S, Cheng C (2004) Improving false discovery rate estimation. Bioinformatics 20: 1737-1745.

18. Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee CK, et al. (2002) A mixture model approach to the analysis of microarray gene expression data. Computational Statistics & Data Analysis 39: 1-20.

19. Pounds S, Morris SW (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. Grenander UIF (1956) On the theory of mortality measurement Part II Skan Aktuarietidskr 39: 125-153.

21. Strimmer K (2008) A unified approach to false discovery estimation. Bioinformatics 9: 303-314.