From the SelectedWorks of Shuo Jiao

February 2010

# Estimating the Proportion of Equivalently Expressed Genes in Microarray Data Based on Transformed Test Statistics

Contact Author

Start Your Own SelectedWorks

Notify Me of New Work

Available at: http://works.bepress.com/shuo_jiao/3

# Estimating the proportion of equivalently expressed genes in microarray data based on transformed test statistics

Shuo Jiao

Department of Statistics

University of Nebraska Lincoln, Lincoln, NE 68583, USA

jiao.shuo@gmail.com

Shunpu Zhang*

Department of Statistics,

University of Nebraska Lincoln, Lincoln, NE 68583, USA

szhang3@unl.edu

*To whom correspondence should be addressed

**Abstract**

In microarray data analysis, false discovery rate (FDR) is now widely accepted as the control criterion to account for multiple hypothesis testing. The proportion of equivalently expressed genes ($\pi_0$) is a key component to be estimated in the estimation of FDR. Some commonly used $\pi_0$ estimators (BUM, SPLOSH, QVALUE, and LBE ) are all based on $p$-values and they are essentially upper bounds of $\pi_0$. Simulation we carried out show that these four methods significantly overestimate the true $\pi_0$ when differentially expressed genes and equivalently expressed genes are not well separated. To solve this problem, we first introduce a novel way of transforming the test statistics to make them symmetric about 0. Then we propose a $\pi_0$ estimator based on the transformed test statistics using the symmetry assumption. Real data application and simulation both show that the $\pi_0$ estimate from our method is less conservative than BUM, SPLOSH, QVALUE, and LBE in most of the cases. Simulation results also show that our estimator always has the least mean squared error among these five methods.

**Keywords**: Proportion of null hypothesis ($\pi_0$); Gene expression analysis; Microarray; Transformed test statistics.

# 1 Introduction

Microarray technology makes it possible to measure the expression levels of thousands of genes simultaneously. A typical goal of analyzing the gene expression data from this technology is to determine which genes are differentially expressed (DE) between two treatment groups, which is actually a multiple hypothesis testing problem. Controlling family-wise error rate (FWER) is a common practice in regular multiple testing problems. However, due to the large number of genes in microarray data, false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) is now commonly used as the choice of the Type I error rate in microarray studies. It is defined as the expected percentage of false positive genes among the claimed significant genes. It was proved that in many cases controlling FDR is more appropriate compared to controlling FWER since the FDR approaches typically rejects more null hypotheses than the FWER approaches (Yekutieli and Benjamini (1999) and Benjamini and Yekutieli (2001)). Several FDR controlling methods are implemented in the R multtest package (Pollard *et al.* (2004)).

To estimate FDR, the proportion of equivalently expressed (EE) genes ($\pi_0$) needs to be estimated first. A number of methods have been proposed to estimate $\pi_0$ and most of them are based on the distribution of $p$-values under the null hypothesis. For gene $i$, the null hypothesis is that gene $i$ is EE, and a $p$-value ($p_i$) will be computed. Notice that the $p$-values of EE genes are uniformly distributed and denote the distribution of $p$-values of DE genes by $h_1(p)$. It is reasonable to model the overall $p$-values as a mixture distribution with two components (McLachlan and Peel, 2000):

$$h(p) = \pi_0 * 1 + (1 - \pi_0)h_1(p). \tag{1.1}$$

In Pounds and Morris (2003), the authors proposed a method called BUM using a beta-uniform mixture distribution to approximate $h(p)$. Then they estimated $\pi_0$ as $\hat{\pi}_0 = \hat{h}(1)$, which assumed $h_1(1) = 0$ and is an upper bound of the true $\pi_0$. Langaas and Lindqvist (2005) adopted the same assumption but used nonparametric maximum likelihood method to estimate $\hat{h}(p)$. SPLOSH (Pounds and Cheng (2004)) uses a local regression technique (LOESS; Cleveland and Devlin

(1988)) to fit $h(p)$ and gives $\hat{\pi}_0 = min_p \hat{h}(p)$ as the estimator, which is still an upper bound. Storey and Tibshirani (2003) proposed the QVALUE method. Given a tuning parameter $\lambda$, QVALUE estimates $\pi_0$ by $\hat{\pi}_0(\lambda) = \frac{\#(p_i > \lambda)}{n(1-\lambda)}$. It can be proved that $\hat{\pi}_0(\lambda) \to \hat{h}(1)$ as $\lambda \to 1$ (Dalmasso *et al.* (2005)), so QVALUE also overestimates $\pi_0$ like BUM and SPLOSH. All these estimators work well if the following assumption holds: few $p$-values of DE genes are close to 1. Otherwise, if this assumption is strongly violated which will happen when DE and EE genes are not well separated, all of them will tend to overestimate. There are other methods not requiring this assumption. Allison *et al.* (2002) proposed a parametric method to estimate $\pi_0$. Dalmasso *et al.* (2005) proposed the LBE method based on the moments of $p$-values, which also only gives an upper bound of $\pi_0$. More recently, Lai (2007) proposed a moment based method which requires no distribution assumption. Unfortunately, his method only works well when there are enough replicates (>8). As we can see from above, the commonly used $\pi_0$ estimators BUM, SPLOSH, QVALUE and LBE are all actually upper bounds of $\pi_0$.

Most of the current $\pi_0$ estimators are based on $p$-values because a $p$-value is a unified measurement of significance. However, as a result of using $p$-values, we may lose some nice properties, such as the symmetry and unimodality of the original test statistics from which the $p$-values are computed. As we know, the commonly used test statistics are $t$-type statistics, which are generally symmetrically distributed and the use of symmetry can be helpful in estimation of $\pi_0$.

In an interesting paper of Bordes *et al.* (2006), they proposed a nonparametric method to estimate the parameters in a two component mixture model with an unknown component, assuming the unknown distribution is symmetric. The authors also tried to apply this method to microarray data by fitting a similar model as (1.1) to test statistics. Since the $t$-type test statistics (without absolute value) for upregulated and downregulated DE genes obviously have different distributions, they cannot be modeled into one component. Hence, the authors constructed an $F$-type test statistic and assumed that it has a symmetric density. However, assuming an $F$-type test statistic to be symmetrically distributed is obviously not a reasonable assumption.

Although the method in Bordes *et al.* (2006) is not very appropriate to be applied to microarray

data directly, it inspired us to use test statistics instead of $p$-values when estimating $\pi_0$. In the next section, we will first introduce a transformation to make the test statistics symmetric about 0, then we propose a $\pi_0$ estimator based on the transformed test statistics using the symmetry assumption. Some theoretical results are given. Finally, application to real microarray data sets and intensive simulations are conducted to compare the performance of our method with BUM, SPLOSH, QVALUE and LBE.

# 2 Methods

## 2.1 The test statistic and the null statistic

Suppose that $Y_{ij}$ is the expression level of gene *i* in array $j$ ($i = 1, 2, ..., n$; $j = 1, ..., j_1, j_1$+1,..., $j_1$+$j_2$), and the first $j_1$ and last $j_2$ arrays are obtained under the two different conditions. For gene $i$, the null hypothesis is that the mean expression levels under the two conditions are the same.

To test this hypothesis, a possible test statistic would be the standard two sample $t$-statistic. However, it only works well when the normality assumption is not strongly violated, which is not always the case in practice. A class of nonparametric statistical methods (Pan *et al.* (2003), Zhao and Pan (2003), Pan (2003), Zhang (2006)) have been proposed to overcome this problem. The basic idea is to directly estimate the null distribution of the test statistic $Z$ by constructing a null statistic $z$ which has the null distribution of $Z$. In other words, for EE genes, the test and null statistics have the same distribution. Among those methods, we decided to use the test and null statistics in Zhang (2006) because their performance are robust and they have improved power over other methods. The test and null statistics are as following:

$$Z = \frac{\frac{\overline{Y}_{11}+\overline{Y}_{12}}{2} - \frac{\overline{Y}_{21}+\overline{Y}_{22}}{2}}{s_0 + \sqrt{\frac{\frac{1}{j_{11}}+\frac{1}{j_{12}}}{4} s_1^2 + \frac{\frac{1}{j_{21}}+\frac{1}{j_{22}}}{4} s_2^2}}, \tag{2.1}$$

$$z = \frac{\frac{\overline{Y}_{11}-\overline{Y}_{12}}{2} + \frac{\overline{Y}_{21}-\overline{Y}_{22}}{2}}{s_0 + \sqrt{\frac{\frac{1}{j_{11}}+\frac{1}{j_{12}}}{4} s_1^2 + \frac{\frac{1}{j_{21}}+\frac{1}{j_{22}}}{4} s_2^2}}, \tag{2.2}$$

where $j_{11} = j_{12} = j_1/2$ if $j_1$ is even, and $j_{11} = j_{12} - 1 = (j_1 - 1)/2$ if $j_1$ is odd. $j_{21}$ and $j_{22}$ are similarly defined. $\overline{Y}_{11}$,$\overline{Y}_{12}$,$\overline{Y}_{21}$,$\overline{Y}_{22}$ are the sample means of the four partitions of the replicates of each gene under the two experimental conditions. Those four partitions are $(Y_{ij}, j = 1, ..., j_{11})$ and $(Y_{ij}, j = j_{11} + 1, ..., j_1)$ from condition 1; $(Y_{ij}, j = j_1 + 1, ..., j_1 + j_{21})$ and $(Y_{ij}, j = j_1 + j_{21} + 1, ..., j_1 + j_2)$ from condition 2. $s_1^2 = \frac{\sum_{j=1}^{j_{11}}(Y_{ij}-\overline{Y}_{11})^2+\sum_{j=j_{11}+1}^{j_1}(Y_{ij}-\overline{Y}_{12})^2}{j_1-2+I(j_1=2)}$, $s_1^2 = \frac{\sum_{j=j_1+1}^{j_1+j_{21}}(Y_{ij}-\overline{Y}_{21})^2+\sum_{j=j_1+j_{21}+1}^{j_1+j_2}(Y_{ij}-\overline{Y}_{22})^2}{j_2-2+I(j_2=2)}$ are the two pooled sample variances from the replicates under each condition. $s_0$ is a fudge factor invented by SAM (Tusher *et al.* (2001)). In

practice, $B$ sets of null statistics are constructed by permutations of data carried within each experimental condition.

## 2.2 Our method

Now for gene $i$, $i = 1, 2, ..., n$, we have the test statistic $Z_i$ and the null statistic $z_i^b$ from the $bth$ permutation set. Similarly as in (1.1), we try to fit a mixture distribution to the test statistics. As stated in the previous section, the distribution of the test statistics for EE genes is already known - same as the null statistics. However, for DE genes, there are two types - upregulated and downregulated. Unlike the $p$-values, the $t$-type test statistics for those two types of DE genes have means of opposite signs. Hence, instead of two subpopulations (EE and DE) of all genes, it is more appropriate to use three subpopulations (EE, upregulated DE, and downregulated DE). With three components in a mixture model, there are more parameters to be estimated. Fortunately, it is actually less problematic than it seems. We will transform the test statistics as following:

The original test statistic for gene $i$ is $Z_i$. We can create a new set of test statistics $X_i$, where $X_i = -Z_i$ or $Z_i$ with the same probability 0.5. In other words, we randomly keep or flip the sign of each $Z_i$. Now take gene $i$ for example: if gene $i$ is EE, then $E(Z_i) = 0 \Rightarrow E(-Z_i) = 0 \Rightarrow E(X_i) = 0$. Therefore, gene $i$ is still EE after transformation; if gene $i$ is DE, we can see that $E(X_i)$=$E(Z_i)$ or $-E(Z_i)$ with same probability 0.5. In other words, for any DE gene $i$, no matter it is originally upregulated or downregulated, this DE gene has the same chance of being upregulated or downregulated after the transformation, which indicates that the proportion of upregulated and downregulated DE genes will be the same. It also implies that the means of upregulated and downregulated DE genes after transformation will be the opposite of each other regardless of their original means.

Before the transformation, separate proportion and mean parameters need to be estimated for upregulated and downregulated DE genes. After the transformation, upregulated and downregulated DE genes have the same proportion and opposite means, which reduces the number of

parameters by 2. Now we can propose a mixture model for the new test statistic $X_i$:

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0)(g(x + \mu_0)/2 + g(x - \mu_0)/2), \tag{2.3}$$

where $f(x)$ is the density function of test statistics $X_i$, $f_0(x)$ is the density function of the test statistics for EE genes, $g(x + \mu_0)$ and $g(x - \mu_0)$ are densities for downregulated and upregulated DE genes ($\mu_0 > 0$), respectively. $g(x)$ is assumed to be an even and unimodal density function. Since the test statistic is of $t$-type, this assumption is reasonable.

Since it is more convenient to estimate the empirical cumulative distribution function (CDF) than density, we can rewrite (2.3) as:

$$F(x) = \pi_0 F_0(x) + (1 - \pi_0)(G(x + \mu_0)/2 + G(x - \mu_0)/2), \tag{2.4}$$

where $F(x)$, $F_0(x)$, and $G(x)$ are the corresponding CDF's for $f(x)$, $f_0(x)$, and $g(x)$, respectively.

The next step is to estimate $G(x)$ from (2.4). First, we have

$$G(x + \mu_0) + G(x - \mu_0) = 2(F(x) - \pi_0 F_0(x))/(1 - \pi_0),$$

and this implies

$$G(x - 2m\mu_0) + G(x - 2(m + 1)\mu_0) = 2(F(x - (2m + 1)\mu_0) - \pi_0 F_0(x - (2m + 1)\mu_0))/(1 - \pi_0), \tag{2.5}$$

Denote the LHS of (2.5) by $C(m)$, we have

$$\begin{aligned}\sum_{m=0}^{m_1}(-1)^m C(m) &= C(0) - C(1) + C(2)... + (-1)^n C(n) \\ &= G(x) + G(x-2\mu_0) - G(x-2\mu_0) - G(x-4\mu_0) \\ &+ G(x-4\mu_0)... + (-1)^{m_1} G(x-2(m_1+1)\mu_0) \\ &= G(x) + (-1)^{m_1} G(x-2(m_1+1)\mu_0) \to G(x) \end{aligned} \tag{2.6}$$

as $m_1$ tends to infinity since $G(x-2(m_1+1)\mu_0) \to 0$ as $m_1 \to \infty$.

Replace $C(m)$ in (2.6) with the RHS of (2.5), we have

$$G(x) = \sum_{m=0}^{\infty} \frac{F(x-(2m+1)\mu_0) - \pi_0 F_0(x-(2m+1)\mu_0)}{(1-\pi_0)/2}. \tag{2.7}$$

Similar as (2.5), we also have

$$G(x+2m\mu_0) + G(x+2(m+1)\mu_0) = 2(F(x+(2m+1)\mu_0) - \pi_0 F_0(x+(2m+1)\mu_0))/(1-\pi_0).$$

Notice that $G(x+2(m_1+1)\mu_0) \to 1$ as $m_1 \to \infty$. Similar argument as from (2.5) to (2.7) shows that

$$G(x) = \sum_{m=0}^{\infty}(-1 + \frac{F(x+(2m+1)\mu_0) - \pi_0 F_0(x+(2m+1)\mu_0)}{(1-\pi_0)/2}). \tag{2.8}$$

Computing the average of (2.7) and (2.8), we have

$$\begin{aligned} G(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{2}\{ &\frac{F(x-(2m+1)\mu_0) - \pi_0 F_0(x-(2m+1)\mu_0)}{(1-\pi_0)/2} \\ &-1 + \frac{F(x+(2m+1)\mu_0) - \pi_0 F_0(x+(2m+1)\mu_0)}{(1-\pi_0)/2}\}. \end{aligned} \tag{2.9}$$

Consider the RHS of (2.9) as a function of $x$, $p(=\pi_0)$, and $\mu(=\mu_0)$, and denote it by

$$M(x;p,\mu)=\sum_{m=0}^{\infty}\frac{(-1)^m}{2}\{\frac{F(x-(2m+1)\mu)-pF_0(x-(2m+1)\mu)}{(1-p)/2}$$
$$-1+\frac{F(x+(2m+1)\mu)-pF_0(x+(2m+1)\mu)}{(1-p)/2}\},$$

so we have $G(x)=M(x;\pi_0,\mu_0)$. We can also define

$$\hat{M}(x;p,\mu)=\sum_{m=0}^{m_1}\frac{(-1)^m}{2}\{\frac{\hat{F}(x-(2m+1)\mu)-p\hat{F}_0(x-(2m+1)\mu)}{(1-p)/2}$$
$$-1+\frac{\hat{F}(x+(2m+1)\mu)-p\hat{F}_0(x+(2m+1)\mu)}{(1-p)/2}\}. \tag{2.10}$$

as the corresponding estimator for $M(x;p,\mu)$. $\hat{F}(x)=\#(X_i<x)/n$ and $\hat{F}_0(x)=\sum_{b=1}^{B}\#(z_i^b<x)/(Bn)$ are the corresponding empirical CDF's for $F(x)$ and $F_0(x)$; $n$ is the number of genes; $B$ is the number of sets of null statistics; $X_i$ is the test statistic and $z_i^b$ is the $bth$ set of null statistic for gene $i$. $m_1$ is a big integer such that $G(x-2(m_1+1)\mu_0)$ and $1-G(x+2(m_1+1)\mu_0)$ are all very close to 0. In this paper, it was chosen to be 20.

Recall that $g(x)$ is an even function, so $G(x)+G(-x)=1$. Following the idea of Bordes *et al.* (2006) and Hunter *et al.* (2007), we define

$$d(x;p,\mu)=(\frac{1-p}{2})^2(M(x;p,\mu)+M(-x;p,\mu)-1)^2. \tag{2.11}$$

Notice that in $M(x;p,\mu)$, $(1-p)/2$ is in the denominator, when $p$ is close to 1 it can be problematic. That is the reason we multiplied the factor $(\frac{1-p}{2})^2$ in (2.11), and an estimate for $M(x;p,\mu)$ is

$$\hat{d}(x;p,\mu)=(\frac{1-p}{2})^2(\hat{M}(x;p,\mu)+\hat{M}(-x;p,\mu)-1)^2.$$

As we can see, $d(x;p,\mu)=0$ for any $x$ when $p=\pi_0$ and $\mu=\mu_0$, which indicates that when

$p = \pi_0$ and $\mu = \mu_0$,

$$D(p, \mu) = \int_0^\infty d(x; p, \mu)dx = 0. \tag{2.12}$$

Hence, $\pi_0$ can be estimated by minimizing

$$\hat{D}(p, \mu) = \int_0^\infty \hat{d}(x; p, \mu)dx \tag{2.13}$$

with respect to $p$ and $\mu$. The integration starts from 0 because $d(x; p, \mu)$ is a symmetric function of x.

Using the above results, the following procedure is proposed to estimate $\pi_0$:

1. Calculate the test statistics $Z_i$ and the null statistics $z_i^b$ using (2.1) and (2.2), $i = 1, ..., n$, $b = 1, ..B$.

2. Create the new test statistics $X_i$ by randomly keeping or flipping the sign of each $Z_i$.

3. Construct an arithmetic sequence with length $J = 50$. Initial term is 0 and last term is $99_{th}$ percentile of the new test statistic $X_i$. Denote this arithmetic sequence by $x_j$, $j = 1, ..., J$. Then

$$\widehat{MSD}(p, \mu) = \frac{|x_J|}{J} \sum_{j=1}^{J} \hat{d}(x_j; p, \mu) \tag{2.14}$$

   is an approximation for $\hat{D}(p, \mu)$ in (2.13) when $J$ is big enough.

4. Let $p^* = p_{init} \geq \pi_0$. $p_{init}$ can be set to 1 or some known upper bound of $\pi_0$.

5. Let $p^{**} = p^* - \Delta$, where $\Delta > 0$ is a small number. Minimize $\widehat{MSD}(p^*, \mu)$ and $\widehat{MSD}(p^{**}, \mu)$ with respect to $\mu$. $\widehat{MSD}(p, \mu)$ has two local minimums with respect to $\mu$, choose the one at larger $\mu$ (we will explain the reason for doing this). $\Delta$ is set to 0.01 here.

6. If $min_\mu \widehat{MSD}(p^{**}, \mu) < min_\mu \widehat{MSD}(p^*, \mu)$, then let $p^* = p^{**}$ and repeat step 5. If not, $\hat{\pi}_0 = p^*$ will be the estimate of $\pi_0$.

7. Repeat step 2-6 for $R = 20$ times and return the average of all $\hat{\pi}_0$'s, which will be the final estimate of $\pi_0$.

Unlike the standard optimization procedure, our searching algorithm is conducted on a decreasing and discrete parameter space of $\pi_0$ because of two reasons.

1. Let $g'(x) = g(x + \mu_0)/2 + g(x - \mu_0)/2$ from (2.3) and $\mu_0' = 0$. We have

$$\begin{aligned} f(x) &= \pi_0 f_0(x) + (1 - \pi_0)(g(x + \mu_0)/2 + g(x - \mu_0)/2) \\ &= \pi_0 f_0(x) + (1 - \pi_0)(g'(x + \mu_0')/2 + g'(x - \mu_0')/2). \end{aligned}$$

   Hence $f(x)$ is not identifiable - there are two possible $\mu_0$'s. Hence, $\widehat{MSD}(p, \mu)$ is small when $\mu$ is close to $\mu_0' = 0$ for any $p$. If we use the standard optimization procedure and search $(p, \mu)$ on the whole parameter space, we may get very biased results. This is also the reason why we choose the local minimum at the larger $\mu$ in step 5 of our algorithm - the $\mu$ associated with the other local minimum is close to $\mu_0' = 0$.

2. Our algorithm is more computational efficient than the standard optimization procedure since it only searches for $p$'s greater than $\pi_0$. This can be proved by the following theorem.

**Theorem 1.** *Suppose $\Theta$ is the parameter space for $(p, \mu)$, and also assume that when $|x|$ is big enough, $f(x)$ and $g(x + \mu_0)/2 + g(x - \mu_0)/2$ in (2.3) are concave upward, then:*

*(i)* $D(p, \mu) \geq 0$ *is a continuous function on* $\Theta$.

*(ii)* $min_\mu D(p, \mu) = 0$ *when* $p = \pi_0$.

*(iii) There exist a threshold* $\pi_u$ *such that* $min_\mu D(p, \mu)$ *is a strictly increasing function of* $p$ *when* $p > \pi_u$.

*(iv)* $\| \widehat{MSD}(p, \mu) - D(p, \mu) \| \rightarrow 0$ *as* $m_1 \rightarrow \infty$, $J \rightarrow \infty$ *and* $n \rightarrow \infty$

The proof of this theorem is in the Appendix. From Theorem 1 (iii) we can see that as long as $p^* > \pi_u$, $min_\mu D(p^*, \mu)$ will be strictly decreasing as $p^*$ decreases until $p^*$ reaches $\pi_u$. We can also notice that $\pi_u$ cannot be less than $\pi_0$ because from Theorem 1 (ii), we know that $min_\mu D(p, \mu)$ will reach 0 as $p$ reaches $\pi_0$ - it cannot be strictly decreasing anymore. This implies that $\underset{p^* \geq \pi_u}{argmin}(\underset{\mu}{min} D(p, \mu)) = \pi_u \geq \pi_0$. Our algorithm is actually trying to search for $\hat{\pi}_0 = \underset{p^* \geq \pi_u}{argmin}(\underset{\mu}{min} \widehat{MSD}(p, \mu))$, and $\widehat{MSD}(p, \mu)$ converges to $D(p^*, \mu)$ by Theorem 1 (iv). Hence, our estimator $\hat{\pi}_0$ will converge to $\pi_u$, an upper bound of $\pi_0$. Through intensive simulations in the next section, we will show that our $\hat{\pi}_0$ is less conservative than the $\pi_0$ estimates given by BUM, SPLOSH, QVALUE and LBE in most of the cases. In fact, $\pi_u$ is very close to the true $\pi_0$ in some scenarios.

# 3 Results

## 3.1 Real Data

First, we will apply our method, along with BUM, SPLOSH, QVALUE and LBE, to two real microarray data sets. The first data set is the leukemia data from Golub *et al.* (1999). In this study, the purpose was to find differentially expressed genes between acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) samples. The total number of genes is 7129 and there are 27 replicates for the ALL and 11 replicates for the AML. The data was pre-processed by the method in Pan *et al.* (2002).

The second data set is the breast cancer data from Hedenfalk *et al.* (2001). This paper tried to find genes which were differentially expressed in tumors with $BRCA1$ mutations and tumors with $BRCA2$ mutations. In the original data set, there are 3226 genes, 7 replicates for $BRCA1$ and 8 replicates for $BRCA2$. As suggested by Storey and Tibshirani (2003), 56 genes were removed in our study because they have expression level greater than 20, which were considered not reliable.

Recall that the input for BUM, SPLOSH, QVALUE and LBE are $p$-values, and the input for our method are test and null statistics. For every data set, test statistic from (2.1) and null statistic from (2.2) were computed. B=100 sets of null statistics were obtained as in Storey and Tibshirani (2003). Then the $p$-value $p_i$ for every gene $i$ was computed by the method in Storey and Tibshirani (2003) using test and null statistics.

In step 4 of our method, we let $p_{init} = 2\sum_{i=1}^{n} p_i/n$. From (1.1), we have: $h(p) = \pi_0 * 1 + (1-\pi_0)h_1(p) \Rightarrow E(p) \geq 0.5\pi_0 \Rightarrow \pi_0 \geq 2E(p)$. Therefore, the chosen $p_{init}$ satisfies the condition that $p_{init} \geq \pi_0$.

Table 1 summarizes the $\pi_0$ estimate of our method, BUM, SPLOSH, QVALUE and LBE (BUM, SPLOSH, QVALUE and LBE all have R implementations). Among the four methods other than our method, BUM always gives the smallest $\pi_0$ estimates while the estimates from the other three methods are close. On the other hand, $\pi_0$ estimate from our method is always the smallest. Since the $\pi_0$ estimates from all the above methods are essentially upper bounds of $\pi_0$, our upper

Table 1 here

bound is apparently less conservative than the others.

Although the real data application shows some advantage of our method, getting a complete idea about its performance is hard because the true $\pi_0$ is unknown. For this reason, intensive simulations with pre-specified $\pi_0$ are conducted in the next section.

## 3.2 Simulated Data

Data were generated for $n$=10000 genes under two conditions, mimicking the large number of genes in practice. Each condition has four replicates, aiming to study the performance of $\pi_0$ estimators when the number of replicates is small (which is usually the case because of the relatively high cost of microarray experiments). $\pi_0$=0.4, 0.6 or 0.8, representing small, medium and large proportion of EE genes. There are three types of simulation set-ups corresponding to three different situations:

(a). *EE, DE genes well separated*

For $10000\pi_0$ EE genes, the expression levels were generated from $N(0,1)$ under both conditions. For $10000(1-\pi_0)$ DE genes, the expression levels under condition 1 were also generated from $N(0,1)$; the expression levels under condition 2 were generated from either $N(3,1)$ or $N(-3,1)$, representing upregulated or downregulated DE genes. The ratio $\#$(upregulated genes)/ $\#$(downregulated genes) is 1 for $\pi_0$=0.4 and 0.8; when $\pi_0$=0.6, the ratio is 2.

(b). *EE, DE genes not well separated*

All the other configurations are exactly the same as set-up (a) except that the expression levels of DE genes under condition 2 are generated from $N(1,1)$ or $N(-1,1)$.

(c). *Mimic the real data*

For EE gene $i$, the expression levels under both conditions are generated from $N(0,\sigma_i^2)$, where $\sigma_i$ is generated from $Gamma(2,4)$. For DE gene $j$, the expression levels are generated

from $N(\mu_{1j}, \sigma_j^2)$ for condition 1 and from $N(\mu_{2j}, \sigma_j^2)$ for condition 2, where $\mu_{1j}$ and $\mu_{2j}$ are generated from $N(0, 2)$, and $\sigma_j$ is generated from $Gamma(2, 4)$.

Next we estimate $\pi_0$ using our method, BUM, SPLOSH, QVALUE and LBE for the simulated data in the same way as for the real data. We repeated the simulation and estimation process 100 times for each set-up. The mean, standard error (SE), and mean squared error (MSE) for all the estimators are summarized in Table 2 , 3 and 4 , respectively. Table 2, 3, 4 here

For Set-up (a), when DE and EE genes are well separated, there should be few DE genes with $p$-values around 1, which is the assumption of BUM, SPLOSH, QVALUE and LBE. Hence, they are all expected to give accurate $\pi_0$ estimates. In fact, the results confirm that QVALUE, and LBE all give satisfactory results, as well as our method. It is worth noting that these three methods give an underestimation of $\pi_0$ while they are supposed to be conservatively biased. Nevertheless, the underestimation bias are very small so they can be explained by variability. BUM and SPLOSH both give notably biased estimates compared to the other three methods in this set-up. The reason may be that BUM and SPLOSH both need to fit $h(p)$ in (1.1), and the fitted $\hat{h}(p)$ does not approximate the real data well. Except when $\pi_0 = 0.6$ the SE of our method is 0.0001 greater than BUM, the SE and MSE of our method are always the smallest among all the five methods.

DE and EE genes are not well separated in Set-up (b). Therefore, we would expect BUM, SPLOSH, QVALUE and LBE to largely overestimate $\pi_0$, which is confirmed by the results. Our method also overestimates $\pi_0$, but to a much less degree. Our method also has the smallest MSE for all $\pi_0$'s. The SE of our method is relatively big when $\pi_0$=0.4 and 0.6 but they are still within an acceptable range.

Set-up (c) adds more variations in the simulation process to mimic the real data. As we expected, the bias of all $\pi_0$ estimates tend to increase compared to (b) because of the bigger variation. Nevertheless, our method still gives the least biased estimate and has the smallest MSE compared with the other methods.

# 4 Discussion

In this paper, we introduce a way of transforming the test statistics, which may be asymmetrically distributed, to make them symmetric about 0. Then we propose a $\pi_0$ estimator based on the transformed test statistics using the symmetry assumption. The real data application and simulation results show the advantageous performances of the proposed method compared with BUM, SPLOSH, QVALUE and LBE.

There are several important parameters in our estimation procedure, such as $m_1$ in (2.10), $J$ in step 3 of our procedure, $\Delta$ in step 5, and $R$ in step 7. As we can see, the precision of our estimator will increase as $m_1$, $J$, and $R$ increase and as $\Delta$ decreases. However, the computational burden will also increase. Hence, more research are necessary to find the optimized choice of those parameters so that we can achieve a balance between precision and computational efficiency.

Furthermore, since our paper has focused on microarray data with a large number of genes and a small number of replicates, more comprehensive studies are needed to compare the performance of different $\pi_0$ estimators under other data configurations.

It should also be noticed that our method is applicable in situations where the test statistics are $t$-type, hence it is not as general as other methods which are based on $p$-values.

# Acknowledgement

## Disclosure Statement

No competing financial interests exist.

# References


Allison, D.B, Gadbury, G.L., Heo, M., Fernndez, J.R., Lee, C.K., Prolla, T.A., Weindruch, R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39, 1-20.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a pratical and powerful approach to multiple testing. *J. R. Statist. Soc*, 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165-1188.

Bordes,L., Delmas,C., and Vandekerkove,P. (2006) Semiparametric estimation of a two-component mixture model when a component is known, *Scand. J. Statist.*, 33, 733-752.

Bordes,L. and Vandekerkove,P. (2007) Semiparametric two-component mixture model with a known component: A class of asympototically normal estimators, *http://hal.inria.fr/docs/00/17/47/25/PDF/BV-AoS-07.pdf*, working paper.

Cleveland,W.S. and Dev,S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting, *J. Am. Stat. Assoc.*, 83, 596-610.

Dalmasso,C., Broet,P. and Moreau,T. (2005) A simple procedure for estimating the false discovery rate, *Bioinformatics*, 21, 660-668.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. AM Stat Assoc.*, **96**, 1151-1160.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 285, 531-537.

Hedenfalk, I., Duggan, D., Chen, Y.D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esterller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344, 539-544.

Hunter,D.R., Wang,S., and Hettmansperger,T.P. (2006) Inference for mixtures of symmetric distributions, *Ann. Statist.*, 35, 224-251.

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

Langaas,M. and Lindqvist, B.H. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data, *J. R. Statist. Soc.* B, 67, 555-572.

McLachlan, G.J., Peel, D. (2000) *Finite Mixture Models.* Wiley Series in Probability and Statistics.

Pan, W., Lin, J., Le, C. (2003) Model-based cluster analysis of microarray gene expression data. *Genome Biol.*, 3, 1-8.

Pan, W., Lin, J., Le, C. (2003) A mixture model approach to detecting differentially expressed genes with microarray data, *Funct Integr Genomics*, 3, 117-124.

Pan, W. (2003) On the use of permutation in the performance of a class of nonparametric methods to detect differential gene expression, *Bioinformatics*, 19, 1333-1040.

Pollard, K.S., Dudoit, S. and van der Laan, M.J. (2004) Multiple Testing Procedures: R multtest Package and Applications to Genomics. *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper , 164. http://www.bepress.com/ucbbiostat/paper164.

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

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partioning the empirical distribution of $p$-values, *Bioinformatics*, 19, 1236-1242.

Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significant analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98, 5116-5121.

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

Yekutieli, D. and Benjamini, Y. (1999) Resampling based False Discovery Rate controlling multiple testing procedure for correlated test statistics. *J Statist. Plann. Inference*, 82, 171-196.

Zhang, S. (2006) An Improved Nonparametric Approach for Detecting Differentially Expressed Genes with Replicated Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 5, Iss. 1, Article 30.

Zhao, Y., Pan, W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 19, 1046-1054.

# Appendix

First, we need a lemma.

**Lemma.** *Suppose $F(x)$ and $f(x)$ are the CDF and PDF of an even function, respectively; Also assume when $|x|$ is big enough $f(x)$ is concave upward. Define*

$N(x;\mu,F)=\sum_{m=0}^{\infty}(-1)^m(F(x-(2m+1)\mu)+F(-x-(2m+1)\mu))$,

$N_1(x;\mu,F)=\sum_{m=0}^{\infty}(-1)^m(F(x+(2m+1)\mu)+F(-x+(2m+1)\mu))$.

*Then there exist a certain threshold $t>0$ such that when $\mu>t$,*

$$x>\mu \Leftrightarrow N(x;\mu,F)>1/2, N_1(x;\mu,F)>1/2$$
$$x<\mu \Leftrightarrow N(x;\mu,F)<1/2, N_1(x;\mu,F)<1/2.$$

*Proof.* We only need to consider $x>0$ since $N(x;\mu)$ is an even function. When $x=\mu$, it is obvious that

$N(x;\mu,F)=\sum_{m=0}^{\infty}(-1)^m(F(-2m\mu)+F(-(2m+2)\mu))$

$=F(0)-F(-2\mu)+F(-2\mu)-F(-4\mu)+F(-4\mu)-...=F(0)=1/2.$

Suppose the corresponding PDF of $F(x)$ is $f(x)$,

$$\frac{\partial N(x;\mu,F)}{\partial x}=\sum_{m=0}^{\infty}(-1)^m(f(x-(2m+1)\mu)-f(-x-(2m+1)\mu)).$$

Consider one part of the RHS of the above equation

$\sum_{m=2k}^{2k+1}(-1)^m(f(x-(2m+1)\mu)-f(-x-(2m+1)\mu))$

$=f(x-(4k+1)\mu)+f(-x-(4k+3)\mu)-f(-x-(4k+1)\mu)-f(-x-(4k+3)\mu)$,

and from the assumption we know that when $|x|$ is big enough, $f(x)$ is concave. Hence, when $\mu$ is big enough,

$f(x-(4k+1)\mu)+f(-x-(4k+3)\mu)>f(-x-(4k+1)\mu)-f(-x-(4k+3)\mu)$

for any $k$.

Therefore $\sum_{m=0}^{\infty}(-1)^m(f(x-(2m+1)\mu)-f(-x-(2m+1)\mu))>0$, which implies

$$\frac{\partial N(x;\mu,F)}{\partial x}>0.$$

Hence, when $\mu$ is big enough, $x > \mu \Leftrightarrow N(x;\mu,F) > 1/2$ and $x < \mu \Leftrightarrow N(x;\mu,F) < 1/2$. The $N_1(x;\mu,F)$ part can be proved similarly. $\square$

*Proof.* Proof of Theorem 1.

(i) $D(p,\mu) = \int_0^\infty d(x;p,\mu)dx$, and $d(x;p,\mu)$ is continuous and bounded. From Lebesgue dominated convergence Theorem, we can conclude that $D(p,\mu)$ is continuous.

(ii) Since

$$\begin{aligned} d(x;\pi_0,\mu_0) &= (\frac{1-\pi_0}{2})^2(M(x;\pi_0,\mu_0) + M(-x;\pi_0,\mu_0) - 1)^2 \\ &= (\frac{1-\pi_0}{2})^2(G(x) + G(-x) - 1)^2 = 0, \end{aligned}$$

Therefore $D(\pi_0,\mu_0) = \int_0^\infty d(x;\pi_0,\mu_0)dx = 0$.
Also we have $min_\mu D(\pi_0,\mu) \geq 0$ and $min_\mu D(\pi_0,\mu) \leq D(\pi_0,\mu_0) = 0$. Hence, $min_\mu D(\pi_0,\mu) = 0$.

(iii) First, we will prove that if $\mu$ is big enough, $D(p,\mu)$ is a strictly increasing function of $p$. Since $d(x;p,\mu) = (\frac{1-p}{2})^2(M(x;p,\mu) + M(-x;p,\mu) - 1)^2$, and

$$\begin{aligned} M(x;p,\mu) = \sum_{m=0}^{\infty} \frac{(-1)^m}{2}\{&\frac{F(x-(2m+1)\mu) - pF_0(x-(2m+1)\mu)}{(1-p)/2} + \\ &-1 + \frac{F(x+(2m+1)\mu) - pF_0(x+(2m+1)\mu)}{(1-p)/2}\}. \end{aligned}$$

Denote $K(x) = G(x+\mu_0)/2 + G(x-\mu_0)/2$ and plug the RHS of (2.4) into the above equation,

we have

$$\frac{1-p}{2}M(x;p,\mu) = \sum_{m=0}^{\infty}\frac{(-1)^m}{2}\{(1-\pi_0)K(x-(2m+1)\mu)+(\pi_0-p)F_0(x-(2m+1)\mu)$$
$$+(1-p)/2+(1-\pi_0)K(x+(2m+1)\mu)+(\pi_0-p)F_0(x+(2m+1)\mu)\}.$$

Write $d(x;p,\mu)$ in terms of $N(x;\mu,F)$ in the Lemma, we have

$d(x;p,\mu) = ((1-\pi_0)(N(x;\mu,K)+N_1(x;\mu,K))+(\pi_0-p)(N(x;\mu,F_0)+N_1(x;\mu,F_0))-(1-p))^2$.

Denote $A(x;\mu) = N(x;\mu,K)+N_1(x;\mu,K)$ and $B(x;\mu) = N(x;\mu,F_0)+N_1(x;\mu,F_0)$, we have:

$$\frac{\partial d(x;p,\mu)}{\partial p} = 2((1-\pi_0)A(x;\mu)+(\pi_0-p)B(x;\mu)-(1-p))(1-B(x;\mu)).$$

When $\mu$ is big enough and $x > \mu$, we know that $B(x;\mu) > 1/2+1/2 = 1$ and $A(x;\mu) > 1/2+1/2 = 1$ from the Lemma. Hence,

$(1-\pi_0)A(x;\mu)+(\pi_0-p)B(x;\mu)-(1-p) > 1-\pi_0+\pi_0-p-1+p = 0$, and $1-B(x;\mu) < 0$. Therefore $\frac{\partial d(x;p,\mu)}{\partial p} < 0$. When $x < \mu$, it can be similarly proved that $\frac{\partial d(x;p,\mu)}{\partial p} < 0$. Hence $\frac{\partial D(p,\mu)}{\partial p} = \int_0^\infty \frac{\partial d(x;p,\mu)}{\partial p}dx < 0$ and $D(p,\mu)$ is a strictly increasing function of $p$.

We can also see that when $p = 1$, $\underset{\mu}{argmin} D(p,\mu) \to \infty$ and $D(p,\mu)$ is a continuous function. Hence, for any $c > 0$, there exist a $p_c$ such that $\underset{\mu}{argmin} D(p_c,\mu) > c$. Now suppose we have $p^* > p^{**}$, and $p^*$ and $p^{**}$ are big enough such that $\underset{\mu}{argmin} D(p^{**},\mu) > t$, the upper bound in the Lemma, and from the Lemma, we have

$$min_\mu D(p^*,\mu) = D(p^*, \underset{\mu}{argmin} D(p^*,\mu)) < D(p^{**}, \underset{\mu}{argmin} D(p^*,\mu)) \le min_\mu D(p^{**},\mu).$$

(iii) is proved.

(iv) From Lemma 3.2 (iii) in Bordes and Vandekerkove (2007), (iv) is obviously true.

□

Table 1: Comparison of $\pi_0$ Estimates for Real Data.

| Data | our method | BUM | SPLOSH | QVALUE | LBE |
|---|---|---|---|---|---|
| Golub *et al.* | 0.560 | 0.592 | 0.662 | 0.652 | 0.685 |
| Hedenfalk *et al.* | 0.533 | 0.603 | 0.675 | 0.709 | 0.710 |

$\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for the Golub *et al.* (1999) and Hedenfalk *et al.* (2001) data.

Table 2: Comparison of Mean and Bias of Different $\pi_0$ Estimates.

| Set-up | $\pi_0$ | our method | BUM | SPLOSH | QVALUE | LBE |
|---|---|---|---|---|---|---|
| (a) | 0.8 | 0.797 | 0.711 | 0.899 | 0.798 | 0.792 |
| | | (-0.003) | (-0.089) | (0.099) | (-0.002) | (-0.008) |
| | 0.6 | 0.594 | 0.480 | 0.828 | 0.598 | 0.594 |
| | | (-0.006) | (-0.120) | (0.228) | (-0.008) | (-0.006) |
| | 0.4 | 0.391 | 0.274 | 0.725 | 0.397 | 0.398 |
| | | (-0.009) | (-0.126) | (0.325) | (-0.003) | (-0.002) |
| (b) | 0.8 | 0.818 | 0.841 | 0.842 | 0.869 | 0.878 |
| | | (0.018) | (0.041) | (0.042) | (0.069) | (0.078) |
| | 0.6 | 0.636 | 0.693 | 0.717 | 0.739 | 0.747 |
| | | (0.036) | (0.093) | (0.117) | (0.139) | (0.147) |
| | 0.4 | 0.456 | 0.568 | 0.597 | 0.610 | 0.623 |
| | | (0.056) | (0.168) | (0.197) | (0.210) | (0.223) |
| (c) | 0.8 | 0.865 | 0.873 | 0.872 | 0.902 | 0.908 |
| | | (0.065) | (0.073) | (0.072) | (0.102) | (0.108) |
| | 0.6 | 0.728 | 0.749 | 0.764 | 0.792 | 0.799 |
| | | (0.128) | (0.149) | (0.164) | (0.192) | (0.199) |
| | 0.4 | 0.596 | 0.620 | 0.659 | 0.679 | 0.695 |
| | | (0.196) | (0.220) | (0.259) | (0.279) | (0.295) |

Mean and bias of $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data. The values outside and inside parenthesis are mean and bias,respectively.

Table 3: Comparison of Standard Error of Different $\pi_0$ Estimates..

| Set-up | $\pi_0$ | our method | BUM | SPLOSH | QVALUE | LBE |
|---|---|---|---|---|---|---|
| (a) | 0.8 | 0.004 | 0.004 | 0.041 | 0.020 | 0.043 |
| | 0.6 | 0.005 | 0.004 | 0.039 | 0.016 | 0.040 |
| | 0.4 | 0.003 | 0.004 | 0.049 | 0.018 | 0.035 |
| (b) | 0.8 | 0.032 | 0.011 | 0.033 | 0.023 | 0.040 |
| | 0.6 | 0.049 | 0.012 | 0.029 | 0.021 | 0.052 |
| | 0.4 | 0.046 | 0.006 | 0.026 | 0.020 | 0.040 |
| (c) | 0.8 | 0.021 | 0.009 | 0.034 | 0.022 | 0.046 |
| | 0.6 | 0.026 | 0.010 | 0.032 | 0.018 | 0.045 |
| | 0.4 | 0.023 | 0.010 | 0.030 | 0.019 | 0.044 |

Standard error of $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data.

Table 4: Comparison of Mean Squared Error of Different $\pi_0$ Estimates..

| Set-up | $\pi_0$ | our method | BUM | SPLOSH | QVALUE | LBE |
|---|---|---|---|---|---|---|
| (a) | 0.8 | 0.00003 | 0.00793 | 0.01052 | 0.00038 | 0.00193 |
| | 0.6 | 0.00006 | 0.01436 | 0.05377 | 0.00026 | 0.00164 |
| | 0.4 | 0.00010 | 0.01599 | 0.10795 | 0.00034 | 0.00121 |
| (b) | 0.8 | 0.00136 | 0.00181 | 0.00284 | 0.00528 | 0.00759 |
| | 0.6 | 0.00372 | 0.00876 | 0.01459 | 0.01986 | 0.02427 |
| | 0.4 | 0.00528 | 0.02815 | 0.03931 | 0.04463 | 0.05160 |
| (c) | 0.8 | 0.00475 | 0.00537 | 0.00631 | 0.01096 | 0.01367 |
| | 0.6 | 0.01709 | 0.02229 | 0.02774 | 0.03732 | 0.04161 |
| | 0.4 | 0.03891 | 0.04845 | 0.06783 | 0.07811 | 0.08915 |

Mean squared error of $\pi_0$ estimates from our method, BUM, SPLOSH, QVALUE and LBE for set-up (a), EE, DE genes well separated; (b), EE, DE genes not well separated; and (c), Mimic the real data.