Gene expression

Construction of null statistics in permutation-based multiple testing for multi-factorial microarray experiments

Xin Gao

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada

Received on December 14, 2005; revised on February 28, 2006; accepted on March 19, 2006
Advance Access publication March 30, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The parametric $F$-test has been widely used in the analysis of factorial microarray experiments to assess treatment effects. However, the normality assumption is often untenable for microarray experiments with small replications. Therefore, permutation-based methods are called for help to assess the statistical significance. The distribution of the $F$-statistics across all the genes on the array can be regarded as a mixture distribution with a proportion of statistics generated from the null distribution of no differential gene expression whereas the other proportion of statistics generated from the alternative distribution of genes differentially expressed. This results in the fact that the permutation distribution of the $F$-statistics may not approximate well to the true null distribution of the $F$-statistics. Therefore, the construction of a proper null statistic to better approximate the null distribution of $F$-statistic is of great importance to the permutation-based multiple testing in microarray data analysis.

Results: In this paper, we extend the ideas of constructing null statistics based on pairwise differences to neglect the treatment effects from the two-sample comparison problem to the multifactorial balanced or unbalanced microarray experiments. A null statistic based on a subpartition method is proposed and its distribution is employed to approximate the null distribution of the $F$-statistic. The proposed null statistic is able to accommodate unbalance in the design and is also corrected for the undue correlation between its numerator and denominator. In the simulation studies and real biological data analysis, the number of true positives and the false discovery rate (FDR) of the proposed null statistic are compared with those of the permuted version of the $F$-statistic. It has been shown that our proposed method has a better control of the FDRs and a higher power than the standard permutation method to detect differentially expressed genes because of the better approximated tail probabilities.

Availability: R codes available upon request

Contact: xingao@mathstat.yorku.ca

1 INTRODUCTION

The development of microarray technologies has provided plentiful amount of genome-wide expression data for biologists to explore various biological mechanisms of interest. These data with simultaneous measurements of thousands of genes present challenging statistical analysis problems. As the normality assumption justifying parametric testing is often untenable in microarray studies (Hunter et al., 2001; Gao et al., 2005), a class of non-parametric statistical methods have been proposed, including the empirical Bayes method of Efron et al. (2001), the significance analysis of microarray (SAM) method of Tusher et al. (2001), and the mixture model method (MMM) of Pan et al. (2003). This class of non-parametric methods are very powerful as they make less stringent distributional assumptions. Furthermore, they fully utilize the special feature of microarray data that there is a large number of genes whereas the number of replications is small. Typically a summary statistic $Z$ is constructed for each gene to measure the differential gene expression across the different treatment conditions. The null distribution of the test statistic $Z$ is estimated by constructing a corresponding null statistic $z$. Then the statistical significance is assessed by comparing the observed $Z$-statistic with this estimated null distribution. The most common way of constructing the null statistic $z$ is to use the permuted version of $Z$. However, it has been emphasized in the recent literature including Efron et al. (2001), Zhao and Pan (2003), Pan (2003), Xie et al. (2005) that the permutation distribution of $Z$ may not approximate well to the true null distribution of $Z$ because there may exist a relatively large proportion of genes in the microarray data which are differentially expressed. Therefore, the permutation distribution can be regarded as a mixture distribution with one component corresponding to the permutation distribution under the null hypothesis while the other component corresponding to the permutation distribution under the alternative hypothesis. The resulted mixture distribution has a larger variance and a heavier tail compared to true null distribution, which leads to a potential loss of power. For the mathematical justification of this phenomenon, readers are referred to Xie et al., (2005).

There have been a series of pioneering work in constructing a suitable null statistic $z$ to estimate the unknown null distribution of $Z$. Efron et al. (2001) proposed to construct $z$ based on the pairwise differences which neglect the treatment effects. In the context of MMM approach, Pan et al. (2003) further proposed a new version of $Z$ and $z$. Assume each condition has even number of replications $n_i, i = 1, 2$. Within each condition, the difference between the sums of the first half and second half $d_i = \frac{\sum_{j=1}^{n_i/2} X_{ij}}{n_i} - \frac{\sum_{j=n_i/2+1}^{n_i} X_{ij}}{n_i}$, is formed. The $Z$-statistic has the numerator of the expression $\sum_{i=1}^{k} X_{i1} / n_1 - \sum_{i=1}^{k} X_{i2} / n_2$, whereas the $z$ statistic has the expression $d_1 n_1 - d_2 n_2$. The $Z$ and $z$ statistics share the same denominator of the standard error based on the two sample variances of $X_{ij}$. Zhao and Pan (2003) pointed out a weakness associated with this statistic that the numerator and denominator of the $Z$-statistic are uncorrelated whereas the numerator and denominator of the $z$ are correlated. This undue correlation for $z$ causes its distribution to be...
different from the null distribution of $Z$. Two modifications were proposed by Zhao and Pan (2003) which corrected the aforementioned problem and further extended the method to non-symmetric noise distribution. The limitation of these two modifications is that efficiency suffers because of the loss of degrees of freedom in the variance estimation. Pan (2003) further proposed a new method in which the sample in each condition is decomposed into two approximately equal sized halves, denoted as parts 1 and 2, respectively. The proposed $Z$ has the numerator measuring the difference $X_{11} + X_{12} - X_{21} - X_{22}$, whereas the proposed $z$ has the numerator measuring the difference $X_{11} - X_{12} + X_{31} - X_{32}$. The proposed $Z$ and $z$ share the same numerator of $\sqrt{S_{11}^2/n_{11} + S_{12}^2/n_{12} + S_{21}^2/n_{21} + S_{22}^2/n_{22}}$. This new set of statistics have the advantage that the undue correlation between the numerator and denominator of the $z$ is corrected and furthermore there are much fewer degrees of freedom lost compared with Zhao and Pan’s methods (2003).

All the aforementioned methods have been developed to address the two condition problem in microarray setting. However, a microarray experiment often has a more complicated design than that of two user-defined groups. Besides the treatment effects of interest, there may exist some clinical covariates such as age, gender and certain clinical symptoms, which also influence the gene expression level. For such experiments, a factorial design model is useful to account for the multiple sources of variation. An overall ANOVA model has been proposed to simultaneously consider all the genes on the arrays and incorporate array effect and dye effect (Kerr et al., 2000). A gene specific ANOVA model under the normality assumption was considered in Jin et al. (2001). When the normality assumption is potentially violated, with a limited number of replications, the sensitivity of these parametric approaches could be severely undermined by using the asymptotic distribution to assess the statistical significance of the test statistic. Ideally, it would be very useful to extend the non-parametric approaches discussed above from the two-sample comparison problem to a more complicated multifactorial set up. Thus it is the objective of this article to construct a proper null statistic to approximate the null distribution of the $F$-statistic which is used as the summary statistic in multifactorial data analysis. It is of further interest to examine the power of the proposed method in detecting true positive and its control of false discovery rate (FDR) under the multiple testing scenario.

2 METHOD

2.1 Model

Given a multifactorial microarray experiment, the expression of the $i$-th gene could be modelled as follows:

$$X_{ijks} = \theta_i + \alpha_j + \beta_k + \gamma_s + \epsilon_{ijks},$$

$i = 1, \ldots, n_i; j = 1, \ldots, J; k = 1, \ldots, K; s = 1, \ldots, n_j$ (1)

with $\sum_j \alpha_j = 0$, $\sum_k \beta_k = 0$ and $\sum_s \gamma_s = 0$, where $i$ indexes for the gene number, $j$ indexes for the treatment group, $k$ indexes for the covariate group and $s$ indexes for the replicate number. Here $X_{ijks}$ stands for the expression measurement, $\theta_i$ represents the $i$-th gene specific mean, $\alpha_j$ represents the effect of the $j$-th treatment group (for instance, drug treatments, tissue types, strains of mice, etc.), $\beta_k$ represents the effect of the $k$-th level of a clinical covariate, and $\gamma_s$ represents the interaction effects between the treatment and the clinical covariate. The error terms $\epsilon_{ijks}$ are independently and identically distributed random noise from a continuous distribution function with a common variance $\sigma^2$. It is worthy to note that this model does not make normality assumption on the error terms $\epsilon_{ijks}$.

2.2 Permutation method

Usually the goal of the analysis is to identify genes which are differentially expressed among the treatment groups, which is equivalent to test the following hypotheses for each individual genes:

$$H_{ij} : \alpha_j = \cdots = \alpha_j, \, \text{versus} \, H_{0ij} : \alpha_j \neq \alpha_j, \, j = 1, \ldots, J.$$ (2a)

For simplicity in notation, the index $i$ is suppressed in the expressions hereafter, as all the observations in the formulae are from the specific gene $i$. The marginal and overall averages of the expression values are defined as $\bar{X}_{ik} = \frac{1}{n_i} \sum_{j=1}^{J} X_{ijks}$, $\bar{X}_j = \frac{1}{n} \sum_{i=1}^{I} X_{ijks}$, and $\bar{X} = \frac{1}{n} \sum_{j=1}^{J} \bar{X}_j$. $\bar{X}$ is the overall mean across all genes and treatments, $\bar{X}_j$ is the mean expression level for the $j$-th treatment group, and $\bar{X}_{ik}$ is the mean expression level for the $i$-th gene at the $k$-th level of the $j$-th treatment group. Given a multifactorial microarray experiment, the expression of the $i$-th gene among the treatment groups, which is equivalent to test the following hypothesis for each individual gene:

$$H_i : \alpha_1 = \cdots = \alpha_j, \, \text{versus} \, H_{0i} : \alpha_j \neq \alpha_j, \, j = 1, \ldots, J.$$ (2b)

## Downloaded from https://academic.oup.com/bioinformatics/article-abstract/22/12/1486/207064 by guest on 25 November 2018
by Efron et al. (2001) to construct a null statistic based on pairwise differences to negate the treatment effects so that its empirical distribution will approximate well to the null distribution of the $F$-statistic regardless whether or not the gene is differentially expressed among treatment groups or not. Assume each cell has even number of replications. Within each cell $(j, k)$, we define $\frac{D_{jk}}{C_1}$ pairwise differences
\[
D_{jk}(s) = \frac{X_{jk(s)} - X_{jk(s+1)}}{2},
\]
for $s = 1, \ldots, \frac{2}{C_1}$. Define the marginal and overall averages of the pairwise differences as
\[
\overline{D}_{jk} = \frac{1}{2} \sum_{s=1}^{2/C_1} D_{jk}(s)/C_1 \quad \text{and} \quad \overline{D}_j = \frac{1}{2} \sum_{s=1}^{2/C_1} D_{jk}(s)/C_1.
\]
Then we construct a null statistic
\[
f_1 = \frac{\sum_{j=1}^K \sum_{k=1}^J \sum_{s=1}^{2/C_1} \frac{1}{C_1} X_{jk(s)} - \overline{X}_{jk(s)}^2}{(J - 1)}. \quad (6)
\]
It is noted that the $F$ and $f_1$ share the same denominator. Their numerators bears the same expression with each $\frac{D_{jk}}{C_1}$ exchanged by the corresponding $D_{jk}$. Define function sign$(a) = 1$, if $a \geq 0$, and sign$(a) = -1$, if $a < 0$. It can be shown that
\[
\sum_{j=1}^K \sum_{k=1}^J \sum_{s=1}^{2/C_1} \frac{1}{C_1} X_{jk(s)} = \frac{m_j}{K} \sum_{k=1}^J \sum_{s=1}^{2/C_1} \frac{1}{C_1} Y_{jk(s)} \quad (7)
\]
and
\[
\sum_{j=1}^K \sum_{k=1}^J \sum_{s=1}^{2/C_1} \frac{1}{C_1} Y_{jk(s)} \quad (8)
\]
Define the vector $\mathbf{D}$ as $(D_1, \ldots, D_J)'$. As under the null hypothesis $a_1 = \ldots = a_J = a$, it follows that $E(\mathbf{X}) = \mathbf{m}$ and the mean vector of $\mathbf{X}$ takes the form of $\mu = (\theta_0 + \alpha_1)$. The distributions of $\mathbf{X} - \mu$ and $\mathbf{D}$ are identical under the assumption of noise distribution symmetric about zero. The numerator of $F$ can be expressed as
\[
\mathbf{X}^T \mathbf{A} \mathbf{X} = (\mathbf{X} - \mu)^T \mathbf{A} (\mathbf{X} - \mu), \quad (9)
\]
where $A\mu = 0$. In parallel, the numerator of $f_1$ can be expressed as $\mathbf{D}^T \mathbf{A} \mathbf{D}$. Because the distributions of $\mathbf{X} - \mu$ and $\mathbf{D}$ are identical, the numerator of $F$ under the null hypothesis has the same distribution as the numerator of $f_1$. Furthermore, $F$ and $f_1$ share the same expression for the denominator. Thus the empirical distribution of $f_1$ can be employed to approximate the null distribution of $F$. For each gene we randomly permute the expression values within each cell and then form the pairwise differences and generate the corresponding $f_1$. Suppose we permute the data $B$ times, the empirical distribution of the total $Bn$ null statistic $f_1$ provides the approximated null distribution of $F$. The advantage of this method is that the $\chi^2$-distribution in the numerator of the $f_1$ maintains a zero non-centrality parameter for both differentially expressed or non-differentially expressed genes thus it can accommodate microarray data with a relatively large proportion of genes with differential expression and possible large magnitude of change.

### 2.4 Null statistic based on subpartition

If we further examine the above method of null statistic based on pairwise differences, we notice that there are certain limitations of this null statistic. First of all, this null statistic can only tackle the design with even number of replicates for each cell. Second there is undue correlation between the numerator and denominator of the $f_1$ statistic. This same undue correlation problem was first pointed out by Zhao and Pan (2003) in constructing null statistic for two-sample comparison problem. To see this problem in the multifactorial setting, let us use the similar argument as that of Pan and Zhao (2003). Given two independent observations from the same distribution, $Y_1$ and $Y_2$, it is readily shown that $cov(Y_1 + Y_2, Y_1 - Y_2) = 0$. Namely the pairwise difference will be uncorrelated with the pairwise sum. For the $F$-statistic, the denominator can be expressed as functions of the sample variance $\overline{X}_{jk}$ within cell $(j, k)$ which is based on pairwise differences whereas the numerator can be expressed as functions of the sample mean $\overline{X}_{jk}$, which is based on pairwise sums. Therefore, the numerator and denominator of the $F$-statistic are uncorrelated. In contrast, for the $f_1$ statistic, the denominator is the same as that of $F$ which is based on pairwise differences whereas the numerator is expressed in terms of functions of $\overline{D}_{jk}$, which is also based on pairwise differences. Therefore, unlike $F$, the numerator and denominator of $f_1$ are unduly correlated which causes the distribution of $f_1$ to be different from that of $F$ under null hypothesis.

To overcome the problem of undue correlation and to further accommodate the null statistic to arbitrary odd or even number of cell sizes, we propose the following null statistic based on the idea of subpartition proposed by Pan (2003) in two-sample comparison problem. For each cell $(j, k)$, we partition it into two approximately equal halves indexed as $j_1$, and $j_2$ with the first half containing $\frac{n_{jk}}{2}$ observations and the second half containing $n_{jk} - \frac{n_{jk}}{2}$ observations. Here $[a]$ denotes the integer part of $a$. Now for each of the total $2JK$ subpartitions, we form the sample means and sample variances:
\[
\overline{X}_{jk_1} = \frac{1}{n_{jk_1}} \sum_{j=1}^{n_{jk_1}} X_{jk_1}, \quad \overline{X}_{jk_2} = \frac{1}{n_{jk_2}} \sum_{j=1}^{n_{jk_2}} X_{jk_2},
\]
\[
\overline{S}_{jk_1}^2 = \frac{1}{n_{jk_1}} \sum_{j=1}^{n_{jk_1}} (X_{jk_1} - \overline{X}_{jk_1})^2, \quad \overline{S}_{jk_2}^2 = \frac{1}{n_{jk_2}} \sum_{j=1}^{n_{jk_2}} (X_{jk_2} - \overline{X}_{jk_2})^2.
\]
It is known that $\overline{X}_{jk_1}$ is uncorrelated with $\overline{S}_{jk_1}^2$, and $\overline{X}_{jk_2}$ is uncorrelated with $\overline{S}_{jk_2}^2$.

Define
\[
\overline{X}_{jk} = \frac{1}{K} \sum_{k=1}^J \overline{X}_{jk},
\]
\[
\overline{D}_{jk} = \frac{1}{K} \sum_{k=1}^J \overline{D}_{jk},
\]
\[
\overline{S}_{jk}^2 = \frac{1}{K} \sum_{k=1}^J \overline{S}_{jk}^2.
\]
Denote the vector $\mathbf{X} = (\overline{X}, \ldots, \overline{X})'$ and the vector $\mathbf{D} = (\overline{D}, \ldots, \overline{D})'$. It follows that under the assumption of symmetric noise distribution and under the null hypothesis, the vector $\mathbf{X} - \mu$ and the vector $\mathbf{D}$ follow the same multivariate distribution with the zero mean vector and the covariance matrix to be $\sigma^2diag(\overline{S}_1^2, \ldots, \overline{S}_J^2)$. The proposed $F_2$ and $f_2$ statistic based on subpartition would take the following forms:
\[
F_2 = \left( \sum_{j=1}^K \sum_{k=1}^J \frac{1}{C_1} \overline{X}_{jk} - \overline{X}_{jk'} \right)^2 / (J - 1)
\]
\[
\sum_{j=1}^K \sum_{k=1}^J (n_{jk} - 1) \overline{S}_{jk}^2 \quad (10)
\]
and
\[
f_2 = \left( \sum_{j=1}^K \sum_{k=1}^J \frac{1}{C_1} \overline{D}_{jk} - \overline{D}_{jk'} \right)^2 / (J - 1)
\]
\[
\sum_{j=1}^K \sum_{k=1}^J (n_{jk} - 1) \overline{S}_{jk}^2 \quad (11)
\]
Then following the similar argument applied to $F_1$ and $f_1$, it is readily shown that the numerators of $F_2$ and $f_2$ share the same distribution under the assumption of symmetric noise distribution. Furthermore, the numerator and the denominator of $f_2$ are uncorrelated. Hence the permutation distribution of $f_2$ provides a good approximation to the null distribution of $F_2$. This method is very general as it works for unbalanced or balanced factorial designs with arbitrary even or odd cell sizes and more importantly it is corrected for the undue correlation problem.

### 3 RESULTS ON SIMULATED DATA

This section illustrates the performances of the proposed methods presented in Section 2 through some simulation experiments. There are three competing methods, traditional permutation method ($F$ and $f_1$), pairwise difference method ($F_1$ and $f_1$) and subpartition method ($F_2$ and $f_2$). (Note the test statistic $F_1$ for the pairwise
variation of 7.63 which is very close to the variance of 7.23 for
from the exact
simulated as symmetric normal distribution
provides an almost perfect agreement with the null distribution of
for
a much smaller variance 1.58 compared with the variance of 4.25
The empirical null distributions of
20 permutations which results in 200 000 null statistics values.
empirical distributions of the null statistics were obtained from
20 permutations which results in 200 000 null statistics values.
The empirical null distributions of
F, F1 and F2 were generated from the exact $F^{1,12}$, $F^{1,12}$ and $F^{1,8}$ respectively.
According to the simulation, it is noticed that the variance of
f-statistics is 22.45 whereas the variance of F-statistics under the null
is 4.25, which are substantially different. This result also agrees with
the findings obtained for two-sample comparison problem in Xie
et al. (2005). The larger variation of f is attributed to the contribution
of the genes which are DE. On the contrary, the distribution of f1 has
a much smaller variance 1.58 compared with the variance of 4.25
for F1 owing to the undue correlation between the numerator and
denominator of f1. Best of the three competing methods, f2 has a
variation of 7.63 which is very close to the variance of 7.23 for F2
reflecting a much better approximation to F2.

Furthermore all the empirical distributions of the null statistics
are plotted against the theoretical density curves of the null
distributions. From Figure 1a, it is observed that f has a much heavier tail
compared with the null distribution of F which is because of the
mixture proportion of DE genes on the array. The heavier tail behavior also explains the larger variation of f listed above. In
Figure 1b, the distribution of f1 has a much shorter tail compared
to the null distribution of F1, which explains the smaller variation of
f1 discussed above. This is because of the undue correlation between
the numerator and denominator of f1. In Figure 1c, it is demonstrated
that the subpartition method yields the null statistic f2 which pro-
vides an almost perfect agreement with the null distribution of F2 for
even extreme tail regions.

3.1 Comparison of the null distribution
The performances of these methods highly depend on how accurately by the proposed null statistics approximate the null distributions of the test statistics. To investigate this aspect, we performed the following simulations. We simulated a $2 \times 2$ design with $n_{11} = n_{12} = n_{21} = n_{22} = 4$. The arrays were simulated to contain 10000 genes with 50% genes non-differentially expressed (NDE) whereas the other 50% genes differentially expressed (DE). The gene specific mean was simulated as $0 \sim U(0, 1)$, the covariate effect $\beta_1 = -\beta_2 \sim N(0, 3)$, the gene-covariate interaction effect $\gamma_{11} = -\gamma_{12} = -\gamma_{21} = \gamma_{22} \sim N(0, 1)$. For those NDE genes, the treatment effect was set to be $\alpha_1 = \alpha_2 = 0$, whereas for those DE genes, the treatment effect was set to be $\alpha_1 = \alpha_2 \sim U(0, 1, 4)$. The noise distribution was simulated as symmetric normal distribution $\epsilon_{ijk} \sim N(0, 1)$. The empirical distributions of the null statistics were obtained from 20 permutations which results in 200000 null statistics values. The empirical null distributions of $F, F_1$ and $F_2$ were generated from the exact $F^{1,12}, F^{1,12}$ and $F^{1,8}$ respectively.

According to the simulation, it is noticed that the variance of
f-statistics is 22.45 whereas the variance of F-statistics under the null
is 4.25, which are substantially different. This result also agrees with
the findings obtained for two-sample comparison problem in Xie
et al. (2005). The larger variation of f is attributed to the contribution
of the genes which are DE. On the contrary, the distribution of f1 has
a much smaller variance 1.58 compared with the variance of 4.25
for F1 owing to the undue correlation between the numerator and
denominator of f1. Best of the three competing methods, f2 has a
variation of 7.63 which is very close to the variance of 7.23 for F2
reflecting a much better approximation to F2.

Furthermore all the empirical distributions of the null statistics
are plotted against the theoretical density curves of the null
distributions. From Figure 1a, it is observed that f has a much heavier tail
compared with the null distribution of F which is because of the
mixture proportion of DE genes on the array. The heavier tail behavior also explains the larger variation of f listed above. In
Figure 1b, the distribution of f1 has a much shorter tail compared
to the null distribution of F1, which explains the smaller variation of
f1 discussed above. This is because of the undue correlation between
the numerator and denominator of f1. In Figure 1c, it is demonstrated
that the subpartition method yields the null statistic f2 which pro-
vides an almost perfect agreement with the null distribution of F2 for
even extreme tail regions.

3.2 Comparison of power and false discovery rates
The above discussion focuses on the approximation to the null
distributions. In multifactorial microarray analysis, the data analysis can proceed by first computing the summary $F$-statistic for each
gene. Then the significance of each individual gene is assessed by
comparing the observed statistic with the distributions of the null
statistics to obtain the empirical $p$-value. Subsequently a threshold for the $p$-values or equivalently a threshold for the test statistics
themselves is determined such that all the genes above the thresholds are declared as differentially expressed genes.

FDR introduced by Benjamini and Hochberg (1995) has been
widely adopted to adjust for the multiple testing problem in microarray
data analysis (Storey and Tibshirani, 2003; Hu et al., 2005, etc). FDR is defined as the expected proportion of the false positive
discoveries among all the declared positive discoveries. First let us
denote the null statistics generated from the $B$ permutations by $F^b, m = 1, \ldots, Bn$. Also denote the $F$-statistics from the $n$
genes by $F_i, i = 1, \ldots, n$. Given a statistic $F^b$, the corresponding
$p$-value is defined as $p$-value($F^b$) = $\frac{1}{Bn} \sum_{m=1}^{Bn} I(f^m \geq F^b)$. 

Fig. 1. (a–c) The comparison of the tail density curves of the empirical
distributions of the null statistics versus the theoretical null distributions
under the normal noise. The solid black curves are the theoretical null dis-
tributions and the dashed curves are the null statistics.
Given a fixed cut-off value $\alpha$ for $p$-value, we can obtain the realized FDR and its estimates (Storey and Tibshirani, 2003), $\text{FDR} = (\pi F_P)/(T_P)$, and $\hat{\text{FDR}} = (\hat{\pi} F_P)/(T_P)$, where $\pi$ is the proportion of NDE genes and $\hat{\pi}$ is its estimator, $F_P$ is the number of false positive genes, i.e. the number of genes which are NDE genes but claimed as DE genes. $\text{FP}$ is the estimated number of false positive genes, $T_P$ is the total number of genes claimed as DE genes. It is proposed to be based on the testing on all genes under the null hypothesis. Similar adjustment can be found in Pan (2003). It is worthy to point out that if permutation method is employed to estimate the $\text{FP}$, the accuracy of the approximation to the null distribution of the $p$-values. As a result, the traditional permutation method will lead to overestimation of the $p$-values, whereas the paired-difference method will lead to under-estimation of the $p$-values. There also exist non-parametric approaches to obtain an estimate of the $\pi$, however, only upper bounds of $\pi$ is readily available (Efron et al., 2001; Dalmasso et al., 2005). As it is beyond the scope of this article to compare different estimation methods for $\pi$, we will use the realized $\text{FDR}$ for the three competing methods.

In practice, if $\pi$ is to be estimated, there are methods available based on the distributions of $p$-values (Storey, 2002; Allison et al., 2002; Pounds and Morris, 2003; Pounds and Cheng, 2004; Guan et al., 2004, http://www.biostat.umn.edu/rsr.php; Wu et al., 2004). It is worthy to point out that if permutation method is employed to estimate the $p$-values, the accuracy of the approximation to the null distribution is also very crucial to the precision of the $p$-values. As a result, the traditional permutation method will lead to overestimation of the $p$-values, whereas the paired-difference method will lead to under-estimation of the $p$-values. Therefore, by selecting the significance level of the non-parametric approaches, we are controlling the estimated number of false positive discoveries. When yielding the same number of $\text{FP}$, the most desirable method should yield the largest number of $T_P$, and in the meantime, the estimated $\text{FP}$ should be as close to the true $\text{FP}$ as possible.

In the simulation, a microarray of 2000 genes was simulated with two levels of the treatment ($J = 2$) and two levels of the clinical covariate ($K = 2$). An unbalanced design of $n_{11} = 4, n_{12} = 6, n_{21} = 6, n_{22} = 4$, was considered. The gene specific mean was simulated as $\theta_i \sim U(0, 1)$, the covariate effect $\beta_1 = -\beta_2 \sim N(0, 3)$, the gene-covariate interaction effect $\gamma_{11} = -\gamma_{12} = -\gamma_{21} = \gamma_{22} \sim N(0, 1)$. For those NDE genes, the treatment effect was set to be $\alpha_1 = \alpha_2 = 0$, whereas for those DE genes, the treatment effect was set to be $\alpha_1 = \alpha_2 \sim U(0.1, 4)$. The noise distribution was simulated as symmetric normal distribution $\epsilon_{(i)k} \sim N(0, 1)$. Other symmetric noise distributions including uniform, double exponential and $t$-distributions were also considered. As the comparison results are similar to the normal distribution, only the result under normal noise is presented in this article. The proportion of NDE genes was set to be $\pi = 0.8$ and $\pi = 0.5$ to represent the situations of low proportion of DE genes and the medium to high proportion of DE genes. To take into account the variability of the statistics, the results were summarized based on 50 simulated datasets.

### Table 1. True positives and FDRs for 2 × 2 unbalanced design with $\pi = 0.8$ under normal noise

<table>
<thead>
<tr>
<th>Method $\alpha$</th>
<th>$T_P$</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
<th>0.0001</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{.f}$</td>
<td>$T_P$</td>
<td>440</td>
<td>362</td>
<td>348</td>
<td>319</td>
<td>303</td>
<td>248</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>1.40</td>
<td>0.71</td>
<td>0.58</td>
<td>0.99</td>
<td>1.64</td>
<td>3.96</td>
</tr>
<tr>
<td>$F_{.p}$</td>
<td>$T_P$</td>
<td>96.28</td>
<td>17.25</td>
<td>8.13</td>
<td>0.90</td>
<td>0.35</td>
<td>0.03</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>1.69</td>
<td>0.72</td>
<td>0.46</td>
<td>0.16</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>$F_{.D}$</td>
<td>$T_P$</td>
<td>0.1747</td>
<td>0.0381</td>
<td>0.0187</td>
<td>0.0022</td>
<td>0.0009</td>
<td>6e-5</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>0.0025</td>
<td>0.0015</td>
<td>0.0010</td>
<td>0.0004</td>
<td>0.0002</td>
<td>6e-5</td>
</tr>
<tr>
<td>$\hat{F_{.D}}$</td>
<td>$T_P$</td>
<td>0.1820</td>
<td>0.0442</td>
<td>0.0230</td>
<td>0.0050</td>
<td>0.0026</td>
<td>0.0007</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>0.0006</td>
<td>9e-5</td>
<td>4e-5</td>
<td>2e-5</td>
<td>1e-5</td>
<td>1e-5</td>
</tr>
<tr>
<td>$F_{1. f_1}$</td>
<td>$T_P$</td>
<td>483</td>
<td>403</td>
<td>387</td>
<td>366</td>
<td>361</td>
<td>355</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>1.52</td>
<td>0.95</td>
<td>0.80</td>
<td>0.78</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>$F_{2. f_2}$</td>
<td>$T_P$</td>
<td>145.45</td>
<td>56.45</td>
<td>40.30</td>
<td>21.00</td>
<td>17.25</td>
<td>11.43</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>1.86</td>
<td>1.08</td>
<td>0.92</td>
<td>0.82</td>
<td>0.80</td>
<td>0.57</td>
</tr>
<tr>
<td>$F_{2. f_2}$</td>
<td>$T_P$</td>
<td>0.2404</td>
<td>0.1119</td>
<td>0.0831</td>
<td>0.0457</td>
<td>0.0381</td>
<td>0.0258</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>0.0024</td>
<td>0.0019</td>
<td>0.0018</td>
<td>0.0017</td>
<td>0.0017</td>
<td>0.0013</td>
</tr>
<tr>
<td>$\hat{F_{2. f_2}}$</td>
<td>$T_P$</td>
<td>0.1656</td>
<td>0.0397</td>
<td>0.0207</td>
<td>0.0044</td>
<td>0.0022</td>
<td>0.0005</td>
</tr>
<tr>
<td>(std)</td>
<td></td>
<td>0.0005</td>
<td>9e-5</td>
<td>4e-5</td>
<td>2e-5</td>
<td>4e-6</td>
<td>1e-6</td>
</tr>
</tbody>
</table>

When maintaining the same estimated false positive numbers $\text{FP}$, the subpartition method identifies more significant genes $T_P$. For instance with a medium proportion of DE genes $\pi = 0.5$, when $\text{FP}$ is controlled at 0.2, in average the subpartition method identifies 161 more genes than the traditional permutation method. In terms of FDRs, the false positive number $\text{FP}$ of the subpartition method is much closer to the estimated false positive number $\text{FP}$ compared with the traditional permutation method. For instance, when $\text{FP}$ is controlled at 10, with $\pi = 0.5$, the average false positive number $\text{FP}$ for the permutation method is 6.80, whereas the average $\text{FP}$ for the subpartition method is 9.92. Consequently the estimated FDR of the subpartition method is more accurate than that of the permutation method. For instance, when $\text{FP}$ is controlled at 10, with $\pi = 0.5$, the average realized and estimated false discovery rates FDR and $\hat{\text{FDR}}$ for the traditional permutation method are 0.0040 and 0.0059, whereas the average FDR and $\hat{\text{FDR}}$ for the subpartition method are 0.0059 and 0.0059. Comparing with the paired-difference method, the subpartition method offers a more valid procedure. It is shown in Tables 1 and 2 that the paired-difference method produces a largely inflated false positive number $\text{FP}$ and severely underestimated FDR rates. Although the paired-difference method constantly identifies more significant genes than its two other competitors, it contains much more false positives than it aims to control. For instance, with $\pi = 0.5$, and the $\text{FP}$ is aimed to be controlled at 10 in average the paired-difference method produces 39.32 false positives, whereas the permutation method and subpartition method produce 6.80 and 9.92 false positives. Consequently, the corresponding average $\text{FDR}$ of the paired-difference method is 0.0055 which is significantly underestimated compared with the true FDRs.
FDR = 0.0218. In contrast, the proposed subpartition method always produces FP numbers close to the estimated FP and provides a much more accurate estimation of the FDR.

In order to investigate the performance of the proposed methods under unbalanced factorial designs with more than two groups by factor, we simulated a microarray of 2000 genes with three levels of the treatment (J = 3) and three levels of the clinical covariate (K = 3). An unbalanced design of N1 = 4, N2 = 5, N3 = 6, N12 = 4, N13 = 4, N23 = 4, and N3 = 5 was considered. The gene-specific mean was simulated as $\mathbf{\mu} = \mathbf{0}$. For those NDE genes, the treatment effect was set to be $\alpha_j = \beta_{jk} = 0$, whereas for those DE genes, the treatment effect was set to be $\alpha_1 \sim U(-4,2)$, $\alpha_2 \sim U(0,2)$, and $\alpha_3 \sim U(4,6)$. The noise distribution was simulated as symmetric normal distribution $\mathbf{\epsilon}_{ijk} \sim N(0, 2)$. The proportion of NDE genes was set to be $\pi = 0.8$.

The results were summarized in Table 3 based on 50 simulated datasets. It is shown from the simulation results that the subpartition method maintains its validity and power in the unbalanced multifactorial settings with more than two groups per factor. Compared with the traditional method, the subpartition method still detects higher numbers of true positives and obtains more accurate estimates for the number of false positives. It is also interesting to note that as the number of levels increases the total number of replications increases as well. Consequently the gap between the traditional permutation method and the proposed subpartition method is lessened as the bias pertained to the traditional permutation method has been alleviated by the increase of sample size.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha$</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
<th>0.0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>100</td>
<td>20</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>0.0524</td>
<td>0.0115</td>
<td>0.0059</td>
<td>0.0013</td>
<td>0.0007</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>(std)</td>
<td>8e-5</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. True positives and false discovery rates for 3 × 3 unbalanced design with even and odd number of replicates when $\pi =$ 0.8 under normal noise

<table>
<thead>
<tr>
<th>Method</th>
<th>$\alpha$</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
<th>0.0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>100</td>
<td>20</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>FDR</td>
<td>0.0524</td>
<td>0.0115</td>
<td>0.0059</td>
<td>0.0013</td>
<td>0.0007</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>(std)</td>
<td>8e-5</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td>6e-6</td>
<td></td>
</tr>
</tbody>
</table>

4 RESULTS ON BIOLOGICAL DATA

Here we present a statistical analysis on a microarray expression data of Drosophila melanogaster (Jin et al., 2001) that has multiple classes of treatments attributing to the expression level. In the data, there were 24 cDNA microarrays, 6 for each combination of two genotypes (Oregon R and Samarkand) and two genders. In total, the gene expression values of the 3931 clone were obtained on the array representing a third of the genome. Focused on individual gene’s expression level measured at the first week, the objective of the analysis was to identify genes whose expression levels are significantly affected by the genotypes and the genders. In the original analysis of Jin et al. (2001), F distribution was employed to assess the statistical significance of the F-statistics obtained from ANOVA tests. Similar to Jin’s approach we modelled the expression data using a multifactorial model and used the F-statistic to summarize the differential gene expression across genotypes and across genders. Without normality assumption, owing to the small number of replications, the exact $F$ distribution may not provide a more accurate assessment of the statistical significance. Therefore we were motivated to reanalyze this dataset using the non-parametric methods proposed in this article. The three competing methods were applied to the datasets. For each method, 10 permutations were employed to generate $F$ distribution. Using the empirical distributions of the null statistics to approximate the null distributions, the empirical $p$-values were obtained for each individual genes. We used a series of thresholds for the $p$-value, ranging from 0.05, 0.01, 0.005, 0.001, 0.0005 to 0.00025. The corresponding estimated number of false positives are 196.55, 39.31, 19.65, 3.93, 1.96 and 0.98. The number TP of significant genes differentially expressed across the two genders and across the two genotypes were provided in Tables 4 and 5.
adult fruit flies are greatly influenced by different genders. Thus this dataset is one of the examples that the microarray contains a large proportion of genes belonging to the alternative situation. Therefore, the traditional permutation method will be a convoluted procedure due to the heavier tail of the null statistic. Contrarily the proposed subpartition method is well suited for this data set and it identifies a greater number of significant genes. For example, when the estimated false positives \( FP \) is controlled at about 1, the subpartition method identifies about 97 genes compared to the 50 genes identified by the traditional permutation method. It has been revealed in the simulations that the paired-difference method offers a very liberal number of significant genes due to the underestimated false positive numbers. Therefore, the seemingly large number of significant genes identified by the paired-difference method is less reliable than the list produced by the subpartition method. We also applied the three methods to detect the differential gene expression across the two genotypes. Compared with the gender effect, less numbers of differentially expressed genes were identified, which also agrees with the findings of Jin et al. (2001) that the transcriptional profiles of adult fruit flies are affected most strongly by sex and less so by genotypes. Nevertheless, the subpartition method is consistently more powerful than the traditional method. For example, when the estimated false positive \( FP \) is controlled at 1.96 the subpartition method identifies about five genes in contrast to the three genes identified by the standard permutation method. With regard to the paired-difference method, the number of significant genes is again liberally large. In conclusion, our analysis confirms the result by Jin et al. (2001) that genders and genotypes have played important roles in determining the transcriptional profiles of adult flies. Furthermore, our proposed subpartition method is more powerful in identifying significant genes owing to the better approximated null distribution.

5 DISCUSSION

This paper presents a set of non-parametric permutation-based tests to detect differential gene expression for multifactorial microarray experiments. It is shown through the simulation results that the traditional permutation method provides a conservative procedure when the array contains a mixture distribution of DE and NDE genes. This owes to the fact that the DE genes induce a larger variation and a heavier tail to the permutation distribution. With regard to the paired-difference method, it is observed that its null statistic has an undue correlation between the numerator and denominator. Thus the distribution of the null statistic is rather different from the null distribution of the summary statistic. Consequently, although the paired-difference method is seemingly more powerful than the traditional permutation method, it identifies a lot more false positives and offers an unsatisfactory control of the FDR.

The aforementioned drawbacks of the traditional permutation method and the paired-difference method have been pointed out by Pan (2003), Zhao and Pan (2003) and Xie (2005) for the two sample comparison problem. We confirmed and extended the findings to the multifactorial setting. In contrast to the traditional permutation method and the paired-difference method, we propose a subpartition method to construct a null statistic which not only neglects the treatment effects for both DE and NDE genes, but also is corrected for the undue correlation problem. Furthermore, by explicitly considering the covariance structure of the mean vector under subpartition, the method works for arbitrary unbalanced factorial designs with even or odd cell sizes. From the simulations and the real biological data analysis, it is shown that the subpartition method offers a much better approximation to the null distribution compared with its two competitors. It is more powerful to identify significant genes than the traditional permutation method and it offers a more satisfactory control of the FDR than the paired-difference method.

It is worthy to point out that the validity of the proposed subpartition method requires a number of basic assumptions. First of all, each cell should have at least four replications, so that the variances for each subpartition could be calculated. Another restriction of the proposed subpartition method, which applies to the paired-difference method as well, is that it requires the assumption of symmetric noise distribution. Third, the model specifies that the \( \varepsilon_{ijk} \) are independent and identically distributed according to a common distribution \( F \). Nevertheless in practice, there might exist correlated random noise in the data. Theoretically it has been shown by Pollard and van der Lann (2004) and Pollard et al. (2005) that the permutation method cannot preserve the correct covariance structure unless the design is balanced or the covariance structure is same across different populations. Thus if the multivariate distributions vary across the different treatment groups and the covariate groups, we should be cautious that the subpartition method may not be invariant under the permutations. To demonstrate this limitation of the subpartition permutation method in the presence of correlated random noise, we conducted the following simulations. A microarray of 2000 genes was simulated with three levels of the treatment \( (J = 3) \) and four levels of the clinical covariate \( (K = 4) \). An unbalanced design of \( n_{11} = 4, n_{12} = 6, n_{13} = 8, n_{14} = 4, n_{21} = 8, n_{22} = 6, n_{23} = 6, n_{24} = 4, n_{31} = 7, n_{32} = 6, n_{33} = 6, n_{34} = 7 \) was considered. The gene specific mean was simulated as \( \theta_i \sim U(0, 1) \). The covariate effects were simulated as \( \beta_1 \sim N(0, 1), \beta_2 \sim N(0, 2), \beta_3 \sim N(0, 3), \beta_4 \sim N(0, 4) \). For those NDE genes, the treatment effect was set to be \( \alpha_1 = \alpha_2 = \alpha_3 = 0 \), whereas the for those DE genes, the treatment effect was set to be \( \alpha_1 \sim U(0, 1), \alpha_2 \sim U(2, 1), \alpha_3 \sim U(4, 5) \). To simulate unequal correlations across different covariate groups, we
Permutation methods for factorial gene expression

Table 6. True positives and false discovery rates for 3 × 4 unbalanced design with even and odd number of replicates when π = 0.8 under correlated noise

<table>
<thead>
<tr>
<th>Method</th>
<th>α</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
<th>0.0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>FP</td>
<td>100</td>
<td>20</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

F_1, f

<table>
<thead>
<tr>
<th>(std)</th>
<th>3.08</th>
<th>2.38</th>
<th>2.53</th>
<th>2.14</th>
<th>2.26</th>
<th>3.76</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP</td>
<td>496.65</td>
<td>241.80</td>
<td>178.43</td>
<td>86.90</td>
<td>64.60</td>
<td>35.75</td>
</tr>
<tr>
<td>(std)</td>
<td>3.78</td>
<td>2.70</td>
<td>2.73</td>
<td>1.76</td>
<td>1.59</td>
<td>1.51</td>
</tr>
<tr>
<td>FDR</td>
<td>0.5103</td>
<td>0.3508</td>
<td>0.2931</td>
<td>0.1823</td>
<td>0.1485</td>
<td>0.0980</td>
</tr>
<tr>
<td>(std)</td>
<td>0.0019</td>
<td>0.0025</td>
<td>0.0031</td>
<td>0.0029</td>
<td>0.0030</td>
<td>0.0033</td>
</tr>
<tr>
<td>FDR</td>
<td>0.1029</td>
<td>0.0291</td>
<td>0.0165</td>
<td>0.0042</td>
<td>0.0023</td>
<td>0.0006</td>
</tr>
<tr>
<td>(std)</td>
<td>0.0004</td>
<td>0.0001</td>
<td>9e-5</td>
<td>2e-5</td>
<td>8e-6</td>
<td></td>
</tr>
</tbody>
</table>

F_2, f_2

<table>
<thead>
<tr>
<th>TP</th>
<th>809</th>
<th>528</th>
<th>452</th>
<th>328</th>
<th>290</th>
<th>224</th>
</tr>
</thead>
<tbody>
<tr>
<td>(std)</td>
<td>2.91</td>
<td>2.41</td>
<td>2.40</td>
<td>2.48</td>
<td>2.84</td>
<td>4.38</td>
</tr>
<tr>
<td>FP</td>
<td>533.60</td>
<td>220.40</td>
<td>148.28</td>
<td>56.18</td>
<td>37.95</td>
<td>17.03</td>
</tr>
<tr>
<td>(std)</td>
<td>3.67</td>
<td>2.67</td>
<td>2.35</td>
<td>1.57</td>
<td>1.43</td>
<td>1.16</td>
</tr>
<tr>
<td>FDR</td>
<td>0.5273</td>
<td>0.3335</td>
<td>0.2620</td>
<td>0.1364</td>
<td>0.1037</td>
<td>0.0587</td>
</tr>
<tr>
<td>(std)</td>
<td>0.0017</td>
<td>0.0026</td>
<td>0.0030</td>
<td>0.0031</td>
<td>0.0032</td>
<td>0.0031</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0989</td>
<td>0.0303</td>
<td>0.0177</td>
<td>0.0049</td>
<td>0.0028</td>
<td>0.0007</td>
</tr>
<tr>
<td>(std)</td>
<td>0.0003</td>
<td>0.0001</td>
<td>0.0001</td>
<td>4e-5</td>
<td>3e-5</td>
<td>2e-5</td>
</tr>
</tbody>
</table>

first simulated the variables \(z_{1,\ldots,N} \sim N(0, 1), z_{2,\ldots,N} \sim N(0, 2), \) and \(z_{3,\ldots,N} \sim N(0, 2).\) Then the noise distribution was simulated as the sum of independent symmetric normal error plus the correlated errors across the covariate groups \(e_{j+k} \sim N(0, 1) + z_{j+k}, \) if \(s \leq \lceil n_j/2 \rceil, k = 1, 2, e_{j+k} \sim N(0, 1) + z_{j+k}, \) if \(s > \lceil n_j/2 \rceil, k = 1, 2, e_{j+k} \sim N(0, 1) + z_{j+k}, \) if \(s \leq \lceil n_j/2 \rceil, k = 3, 4, e_{j+k} \sim N(0, 1) + z_{j+k}, \) if \(s > \lceil n_j/2 \rceil, k = 3, 4.\) The proportion of NDE genes was set to be \(\pi = 0.8.\) The results are summarized in Table 6 based on 50 simulated datasets. The simulation result clearly demonstrates that in the presence of correlations which are different across covariate groups, both the traditional permutation method and the subpartition permutation method no longer maintain the good control of the FDRs.

Another assumption to validate our approach requires that all the test statistics are independent and identically distributed, which applies to many other existing FDR approaches as well (Efron et al., 2001, Tusher et al., 2001 and Storey and Tibshirani, 2003). As the mRNAs for all the genes on an array are extracted from the same sample and are subject to the same experimental condition and normalization process, it is reasonable to assume that the marginal distributions for the majority of the genes are of the same type. Furthermore for most microarray studies, genes either act independently or have weak dependence whose effect on the FDR becomes negligible as the number of genes goes to infinity (Storey and Tibshirani, 2003). Under this condition, we can employ only a small number of permutations for each gene and pool all the permutations from the genes together to form the empirical null distribution. Collapsing the distributions of the test statistics not only reduces the computational cost but also overcomes the discrete nature of the permutation distribution of a test statistic based on few observations (Tusher et al., 2001, and Reiner et al., 2003, etc.). Nevertheless in practice there may exist violations such that strong dependency structure does occur among the test statistics. For example in cancer studies, the co-regulations of genes based on genomic locations and regulation network can lead to strong correlations among the genes (Reiner et al., 2003). To account for the dependency structure, we can relax the assumption and modify our procedures. Instead of pooling all the test statistics together, we employ a large number of permutations. For each permutation, we randomly shuffle the arrays within the same combination of treatment and covariate group and all the genes on the same array are shuffled simultaneously. Upon each permutation, the vector of the subpartition null statistics \(F_j\) for all the genes is computed. The data are repeatedly permuted with the empirical dependency structure of the data being preserved. By this means the joint multivariate null distribution of the test statistics can be generated from this resampling scheme. Denote the null statistic for gene \(i\) generated from the \(B\) permutations by \(F_{ij}, i = 1, \ldots, B.\) Also denote the \(F\)-statistics from the gene \(i\) by \(F_i, i = 1, \ldots, n.\) The corresponding empirical \(p\)-value for the statistic \(F_i\) is defined as \(p\)-value \((F_i) = \frac{1}{B} \sum_{j=1}^{B} I(F_{ij} \geq F_i).\) Given a fixed cut-off value \(\alpha\) for \(p\)-value, we can obtain the realized FDR and its estimates. It is noted that using the multivariate null distribution as the reference distribution, given the \(p\)-value threshold \(\alpha\) for each marginal distribution, the estimated number of false positives still equals \(n\alpha.\) This is because the expectation of sums of Bernoulli false discovery outcomes do not depend on the correlation structure of the joint distribution (Hu et al., 2005). Therefore we are able to control the estimated FDRs under the dependency structure.

ACKNOWLEDGEMENTS

This research was supported by Canadian NSERC Grant.

Conflict of Interest: none declared.

REFERENCES


