Symmetric directional false discovery rate control

Sarah E. Holte, Eva K. Lee, Yajun Mei

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form 2 June 2016
Accepted 12 August 2016
Available online 24 August 2016

MSC:
62G10
62C25
62P10

Keywords:
Column permutation
Directional FDR
False discovery rate
Multiple testing
Symmetric decision
Three-decisions

ABSTRACT

This research is motivated from the analysis of a real gene expression data that aims to identify a subset of “interesting” or “significant” genes for further studies. When we blindly applied the standard false discovery rate (FDR) methods, our biology collaborators were suspicious or confused, as the selected list of significant genes was highly unbalanced: there were ten times more under-expressed genes than the over-expressed genes. Their concerns led us to realize that the observed two-sample $t$-statistics were highly skewed and asymmetric, and thus the standard FDR methods might be inappropriate. To tackle this case, we propose a symmetric directional FDR control method that categorizes the genes into “over-expressed” and “under-expressed” genes, pairs “over-expressed” and “under-expressed” genes, defines the $p$-values for gene pairs via column permutations, and then applies the standard FDR method to select “significant” gene pairs instead of “significant” individual genes. We compare our proposed symmetric directional FDR method with the standard FDR method by applying them to simulated data and several well-known real data sets.

© 2016 Elsevier B.V. All rights reserved.

* Corresponding author.
E-mail addresses: sholte@fredhutch.org (S.E. Holte), evakylee@isye.gatech.edu (E.K. Lee), ymei@isye.gatech.edu (Y. Mei).
This research is motivated from the analysis of a real gene expression data. As in the typical comparative genomics studies with high-throughput technologies, the data set we faced is from measuring the expression levels of \( m = 54,675 \) genes on \( n = 16 \) microarrays for two groups: \( n_1 = 8 \) healthy subjects and \( n_2 = 8 \) cancer subjects. The goal is to identify genes that are significantly differentially expressed between two groups with a potential of offering biomarker candidates.

Initially we thought this was a standard multiple hypothesis testing problem that often arises in modern biomedical applications such as genomic, proteomic, and metabolomic, and thus we blindly applied the standard false discovery rate (FDR) control method of Benjamini and Hochberg [2]: we calculated a two-sample \( t \)-statistic \( t_i \) for each gene \( i \), permuted column data (randomly label cancer/normal subjects) to simulate the null distribution of the \( t \)-statistics, computed the corresponding two-sided \( p \)-values \( p_i = \Pr_0(|T| > |t_i|) \)'s for each gene \( i \), and then used the standard Benjamini–Hochberg FDR method to select significant genes. However, when we reported the list of significant genes to our biology collaborators, they were suspicious, and felt the results did not make biology sense. We thought that this might be due to the simplicity of the Benjamini–Hochberg FDR method, and thus we re-analyzed data by applying more advanced FDR methods such as the robust FDR method of Benjamini and Yekutieli [3], the \( q \)-value of Storey [15,16], and the empirical Bayes estimate of the null distribution of Efron [4]. Unfortunately, our biology collaborators were still unsatisfactory to the results. After lengthy discussions, we realized that in our list of significant genes, we have selected ten times more negatively expressed genes than the positively expressed genes, but our biology collaborators preferred the list of significant genes to be balanced, since symmetry is common in many biology systems. More importantly, our biology collaborators did not use any specific biology knowledge to purposely choose negatively or positively expressed genes in the experiments.

It is natural to ask what happened to the data set we analyzed? Fig. 1 plots the histogram and QQ-norm plot of the observed \( t \)-statistics \( t_i \)'s in our data set and both plots clearly suggest that the observed \( t_i \)'s are highly skewed to negative and any normal distribution \( N(\mu_0, \sigma^2) \) will likely be a poor approximation to the null distribution of \( t_i \)'s. In other words, it is not clear how to estimate the null distribution \( \Pr_0 \) of \( t_i \)'s for our data set. It is important to emphasize the role of the null distribution \( \Pr_0 \) of \( t_i \)'s when genes are insignificantly differentiated expressed, since otherwise the corresponding \( p \)-values can be useless and thus the standard FDR methods are inappropriate. As mentioned in Efron [4], there are several methods to derive the null distribution of \( t_i \)'s in the literature. The first one is the theoretical \( t \)-distribution under the assumption that the data \( x_i \)'s are independent normally distributed, and this is often referred as the theoretical null distribution. The second method is data permutation methods by randomly labeling normal and cancer subjects and using the re-calculated \( t_i^* \) to simulate the null distribution. As pointed out in Efron [4], data permutation methods essentially approximate the null distribution of \( t_i \)'s as \( N(0, \sigma^2) \) after some suitable transformations, and do not help if the observed \( t_i \)'s is not symmetric at 0. This view motivated Efron [4] to propose the third method that approximates the null distribution based on empirical Bayes: it is assumed that the null distribution is \( N(\mu_0, \sigma^2) \) after transformations, where the null mean \( \mu_0 \) is estimated from the observed \( t_i \)'s that are likely from the null, say those between the first and third quartiles.

Unfortunately all these three existing approaches of estimating the null distribution of \( t_i \)'s do not work in the case when the observed \( t_i \)'s are asymmetric and highly skewed. One possible remedy is to extend the empirical Bayes method of Efron [4] by considering a mixture of normal or other distributions that can take into account the skewed or asymmetric properties of the observed \( t_i \)'s. See, for instance, Zhao et al. [19] and Beana et al. [1], which applied the mixture distribution to address the skewness that is due to the non-null or significant genes. When the null distribution is skewed, one may still be able to use the mixture model to fit both null and non-null distribution of \( t_i \)'s, but it is unclear how to classify the components of the mixture model between the null and non-null distribution. Moreover, such approach essentially assumes that the extreme behavior of the \( t_i \)'s can be predicted based on the non-extreme values of \( t_i \)'s, which is questionable or at least debatable.

In this article, we propose a novel FDR method that can circumvent the difficulty of estimating the null distribution of \( t_i \)'s when they are highly skewed. Motivated by the rationale and remarks of our biologist collaborators, we note that the ultimate goal in FDR is not necessarily on estimating
Fig. 1. Histogram and q-q plot for the $t$-statistics in a data set, which range from $-10.15$ to $7.64$. Upper Panel: the histogram, with the two fitted pdfs: one is for the theoretical null $N(0, 1)$ distribution, and the other is for the null distribution $N(0.5713, 1.5623^2)$ estimated from the test statistics in $[-3, 3]$. Bottom Panel: the QQ normal plot with a theoretical line. The figures show that the observed $t$-statistics are highly skewed to the left, and the normal approximations are poor fit.

thenulldistributionofthe$t_i$’s, but on choosing the significant genes where the numbers of selected over-expressed and under-expressed genes are desired to be comparable. Thus we propose to put an additional assumption on the significant genes: the number of over-expressed genes (i.e., those significant genes with $t_i > 0$) is the same as the number of under-expressed genes (i.e., those significant genes with $t_i < 0$) among the chosen significant genes. As compared to standard FDR methods, our method shifts the symmetric assumption from the null-distribution of $t_i$’s to our decision making on the selected significant genes.

Note that the decision of selecting significant under-expressed and over-expressed genes is also called the directional FDR in the literature, see, for example, Lewis and Thayer [10] and Sarkar and Zhou [13]. However, the existing directional FDRs assume that the null distribution of the $t_i$’s is known (normal or $t$-distribution), so that one can compute the one-sided $p$-values, and the emphasis is often to control the Type III errors that mis-classify under-expressed genes as over-expressed or vice-versa. There are a couple of differences between our setting and the existing directional FDR. From the general decision theoretic formulation viewpoint, instead of 0–1 indicator-type loss in the
directional FDRs, our formulation is inspired by the three decisions problem in Lehmann [8,9] and put different losses on different kinds of mis-classifications. From the optimization constraint viewpoint, we assume that our decision of chosen significant genes are symmetric: the number of over-expressed genes is the same as the number of under-expressed genes, whereas the directional FDR does not have such symmetric constraints. In our practical experiences of using the FDR methods to select significant genes, the numbers of selected over-expressed and under-expressed genes are often in the same ball park range in most real data sets when the observed $t_i$’s are symmetric, and thus the assumption of symmetric decision is not that bad. In fact, when the standard FDR methods selected very different numbers of over-expressed and under-expressed genes, one needs to be very cautious and should double check whether the corresponding observed $t_i$’s are skewed or not, as the standard FDR methods might be used inappropriately.

We acknowledge that the assumption of symmetric decision might sound too restrictive in some applications, where one might want the ratio between negatively and positively expressed genes to be $k_1 : k_2$ for selected significant genes. Our proposed directional FDR method can be extended to those cases by grouping genes as blocks of $k_1$ negatively expressed genes and $k_2$ positively expressed genes, estimating the $p$-values for each block, and then selecting significant blocks, although further theoretical investigation on the blocks is needed. We should also mention that while this article focuses on the two-sample $t$-statistics $t_i$’s as the local test statistic to generate $p$-value for each gene, the idea can also be applied to other local test statistics in a more complicated design of high-throughput biology experiments. As illustrated in Leek [7], and Love, Huber, and Anders [11], in many latest high-throughput experiments, one often needs to build a more advanced model than the independent two-sample model and needs to generate $p$-values for all genes from some local test statistics that can remove the batch effects and/or take into account other explanatory factors. Our proposed directional FDR method can be adapted to the case when the observed local test statistics are skewed and the theoretical null distribution is a poor fit.

The remainder of this paper is as follows. In Section 2, we present our motivating example in more details and illustrate that the standard FDR method produces a very unbalanced decision. We formulate the symmetric decision problem as a compound three-decisions problem in Section 3, and the corresponding optimal solution motivates us to propose a symmetric directional FDR method in Section 4 that is computationally simple. To illustrate our main ideas, we will focus on the gene expression problem in the independent two-sample model and demonstrate how to modify the standard Benjamini–Hochberg FDR method in [2] to make symmetric decisions, but our ideas can be easily applied to other applications, other FDR methods and other models. Then we compare the performance of our proposed symmetric directional FDR method with the standard FDR methods in simulated data in Section 5 and several well-known real data sets in Section 6. Section 7 includes a discussion of our method and further research. The proof of Theorem 3.1 is presented in the Appendix.

2. The motivating example

Suppose that we observe an $m \times n$ data matrix $X = (x_{ij})$, where $m$ denotes the number of genes, $n$ denotes the number of subjects and the entry $x_{ij}$ is the expression level for gene $i$ on subject $j$. We assume that the first $n_1$ columns of $X$ represent normal subjects and the last $n_2$ columns represent cancer subjects, where $n_1 + n_2 = n$. For each gene $i$, the $i$th row of the data matrix $X$ represents its data, and we can compute Welch’s $t$-statistic:

$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sqrt{s_{i1}^2/n_1 + s_{i2}^2/n_2}},$$

(1)

where $\bar{x}_{i1} = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{ij}$ and $\bar{x}_{i2} = \frac{1}{n_2} \sum_{j=n_1+1}^{n_1+n_2} x_{ij}$ are the averages of $x_{ij}$ for normal and cancer groups and $s_{i1}^2 = \sum_{j=1}^{n_1} (x_{ij} - \bar{x}_{i1})^2/(n_1 - 1)$ and $s_{i2}^2 = \sum_{j=n_1+1}^{n_1+n_2} (x_{ij} - \bar{x}_{i2})^2/(n_2 - 1)$ are the corresponding sample variances.

In the FDR research, it is important to investigate the (null and alternative) distribution of the observed $t_i$’s in (1). While the theoretical null distribution of $t_i$’s is a $t$-distribution that is symmetric
about 0, in practice the distribution of $t_i$'s is often different from the theoretical null $t$-distribution, since the data $x_{ij}$'s are not necessarily normally distributed, see Efron [4].

Fig. 1 illustrates the histogram and qqnorm plot for $t$-statistics in (1) from the real data set mentioned in the Introduction when $m = 54675$, $n = 16$ and $n_1 = n_2 = 8$. To illustrate the asymmetric properties of $t$-statistics, we not only plot the pdf for the $N(0, 1)$ distribution in the histogram, but also plot another normal distribution that approximates the null distribution of the $t_i$'s from the observed value. In this approximation, we use the observed $t$-statistics $t_i$'s in $[-3, 3]$ to estimate the pdf of null distribution, since most of the $t$-statistics whose values near 0 should come from the null hypothesis. For this data set, those $t$-statistics $t_i$'s in $[-3, 3]$ have sample mean 0.5713 and sample variance 1.5623. Thus one may approximate the null distribution of the $t$-statistics as $N(0.5713, 1.5623^2)$. As we can see from Fig. 1, it is unfortunate that the $t$-statistics are highly skewed to the left, and the normal approximations are poor fit. Not surprisingly, if we blindly apply the standard FDR methods, we would have selected ten times more negatively expressed genes than the positively expressed genes.

To further understand the null distribution of data, we follow Efron [4] to assume that most genes are in the null, and regard the $i$th row (or genes) as null whenever the observed $t$-statistics $t_i$'s are between 25% and 75% percentiles of all $t_i$'s. This leads to $m_1 = m/2$ null genes. Then we analyze the four-dimensional vectors $(\bar{x}_1, s_1^2, \bar{x}_2, s_2^2)$'s that are used to compute the $t$-statistics $t_i$'s in (1) for the $m_1$ null genes/rows: we do the K-means clustering analysis on these $m_1 = m/2$ four-dimensional vectors $(\bar{x}_1, s_1^2, \bar{x}_2, s_2^2)$'s, and find the center and proportion of each cluster. To highlight our main ideas, we considered only $K = 3$ clusters. By using the “kmeans ()” function from the free statistical software R (version 2.15.2), the centers of these three clusters were found to be $(2.73, 0.16, 2.61, 0.07), (5.17, 0.81, 4.96, 0.62)$, and $(8.17, 0.99, 8.03, 0.81)$ with the proportions of the three clusters being 73.6%, 17.8%, 8.6%, respectively. Note that $\bar{x}_1 > \bar{x}_2$ for all three group centers, and the $s_1^2$ and $s_2^2$ are also different for three groups. Thus there appeared to have a systematic bias on the means between normal and cancer groups, although such bias seemed to be not that significant given the corresponding standard deviations.

3. Problem formulation and oracle rule

Inspired by the three decisions problem in Lehmann [8,9] and the compound binary-decision framework for multiple-testing problems in Sun and Cai [18], we present a Bayesian formulation of symmetric decision in the context of FDR control. For the purpose of convenience, one will transform the $t$-statistic $t_i$ to a $z$-value by $Z_i = \Phi^{-1}(F_i(t_i))$, where $F_i$ is the null cumulative distribution function (cdf) of the $t$-statistic $t_i$ and $\Phi$ is cdf of the standard normal variable. Thus, we now observe $m$ derived random variables, $Z_1, \ldots, Z_m$.

We assume that $Z_i$ has a distribution function $F_{\theta_i}$ with the unknown parameters $\theta_i$ being one of three possible values, $\{-1, 0, 1\}$. Here $\theta_i = 0$ implies that the $i$th gene is “null”, and $\theta_i = -1$ means that the $i$th gene is under-expressed, and $\theta_i = 1$ implies that the $i$th gene is over-expressed. Denote by $f_{\theta_i}$ the corresponding density function. In practice, the null density $f_0$ is often assumed to be known at least theoretically, say $f_0 = \text{pdf of } N(0, 1)$, but $f_{-1}$ or $f_1$ is unlikely to be known to the statistician a priori. The goal is to classify each $\theta_i \in \{-1, 0, 1\}$ based on the observed values of $Z_i$’s.

To simplify our notation, denote by $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ the unknown parameters, and let $\delta = (\delta_1, \ldots, \delta_m)$ be a general decision rule on $\tilde{\theta}$, where $\delta_i = \delta_i(\tilde{Z}) = \delta_i(z_1, \ldots, z_m) \in \{-1, 0, 1\}$ be the decision rule for the $i$th parameter $\theta_i$ when we observe $Z_1 = z_1, \ldots, Z_m = z_m$. Note that each $\delta_i$ depends on not only $z_i$, but also possibly other $z_k$’s, $k \neq i$, since we want to estimate $\theta_i$’s simultaneously.

The Bayesian problem can be formulated as follows. The unknown parameters $\theta_1, \ldots, \theta_m \in \{-1, 0, 1\}$ are assumed to be independent variables with a prior probability mass function

$$P(\theta_i = -1) = P(\theta_i = 1) = p \quad \text{and} \quad P(\theta_i = 0) = 1 - 2p$$

for some $0 < p < 1/2$. In other words, the underlying true model is assumed to be symmetric in terms of over-expressed and under-expressed genes. Moreover, when the true parameter vector is
\( \hat{\theta} = (\theta_1, \ldots, \theta_m) \), the loss function of a decision rule \( \hat{\delta} = (\delta_1, \ldots, \delta_m) \) is of the form
\[
L(\hat{\theta}, \hat{\delta}) = \sum_{i=1}^{m} L(\theta_i, \delta_i)
\]
where the individual loss \( L(\theta_i, \delta_i) \) is given below:

<table>
<thead>
<tr>
<th>( \delta_i )</th>
<th>( \delta_i = -1 )</th>
<th>( \delta_i = 0 )</th>
<th>( \delta_i = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_i = -1 )</td>
<td>0</td>
<td>( \lambda )</td>
<td>( \sigma )</td>
</tr>
<tr>
<td>( \theta_i = 0 )</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \theta_i = 1 )</td>
<td>( \sigma )</td>
<td>( \lambda )</td>
<td>0</td>
</tr>
</tbody>
</table>

Specifically, if the true value of \( \theta_i \) is 0 but we make a wrong decision of \( \delta_i = 1 \) or \(-1 \), the loss is 1. On the other hand, when the true value of \( \theta_i \) is 1 (or \(-1 \), the loss is \( \lambda > 0 \) if we make a wrong decision of 0, and the loss becomes \( \sigma \) if the wrong decision is \(-1 \) (or 1).

Mathematically, we want to find a decision rule \( \hat{\delta} \in \{-1, 0, 1\}^m \) that minimizes the Bayesian risk
\[
R(\hat{\delta}) = \sum_{a_1, \ldots, a_m \in \{-1,0,1\}} E_{\hat{\theta}}(L(\hat{\theta}, \hat{\delta})) P(\hat{\theta} = (a_1, \ldots, a_m)).
\] (2)

Note that similar formulations have been considered for the directional FDR, see Lewis and Thayer [10] and Sarkar and Zhou [13], but our loss functions here are different, and our specific loss functions are similar to those in Lehmann [8,9]. Also it is interesting to note that in the extreme case of \( \sigma = 0 \), we can simply combine the decision \( \delta_i = -1 \) and \( \delta_i = 1 \), and thus the problem becomes the classical binary hypothesis problem in which one is interested in selecting significant genes, see Sun and Cai [18]. Another special case is when \( \sigma = 2\lambda \), which turns out to be of most interest to us since its Bayes decision rule has a symmetric structure. The following theorem characterizes the corresponding Bayes decision rule, which is called oracle rule, as it needs an oracle that tells us the values of \( p, \lambda \) and the density functions \( f_{-1}, f_0, f_1 \).

**Theorem 3.1 (Oracle Rule).** When \( \sigma = 2\lambda \), the oracle rule that minimizes the Bayesian risk (2) is given by:
\[
\hat{\delta}_i = \begin{cases} 
1, & \text{if } \frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \geq \frac{1 - 2p}{\lambda p}; \\
-1, & \text{if } \frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \leq -\frac{1 - 2p}{\lambda p}; \\
0, & \text{otherwise.}
\end{cases}
\]
for \( i = 1, \ldots, m \).

It is interesting to note that this oracle rule has a symmetric decision structure based on the values of the variables
\[
g(z_i) = \frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)}, \tag{3}
\]
as the oracle rule declares a \( \hat{\delta}_i = \pm 1 \) if and only if \( |g(z_i)| \geq (1 - 2p)/\lambda p \). However, \( g(z) \) in (3) is not necessarily a symmetric function of \( z \) even when \( f_0(z) = f_0(-z) \) as \( f_{-1} \) and \( f_1 \) can be very different. In other words, the oracle rule makes a symmetric decision based on \( |g(z_i)| \)'s in (3), but it does not necessarily make a symmetric decision based on \( |z_i| \)'s themselves.

From the oracle rule, it motivates us to rank the observed values \( z_i \)'s in an increasing order from the “most negative” to the “most positive” based on the values of \( g(z_i) \) in (3). It turns out that in many cases, \( g(z_i) \) in (3) is a monotonic function of \( z_i \). For example, this is true if \( f_{-1}, f_0, f_1 \) are the pdfs of normal distributions with different means and same variance \( \sigma^2 \). Likewise, this also holds when \( f_0 \) is the pdf of \( N(0, \sigma^2) \) and \( f_{-1} \) and \( f_1 \) are the pdfs of the mixture distributions of \( N(\mu, \sigma^2) \) for
\( \mu < 0 \) and \( \mu > 0 \), respectively, i.e., \( f_1(z) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) d\Pi_1(\mu_1) \) and \( f_{-1}(z) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) d\Pi_2(\mu_2) \). Thus, the order of \( g(Z) \) is often simply based on the order of the actual values of \( Z \) themselves. This key property motivates our proposed symmetric directional FDR method in the practical situations when the parameters \( \lambda, p \) or the pdfs \( f_0 \) and \( f_{\pm1} \) are unknown.

It is also useful to point out that for a given observed \( Z_i = z_i \), the local Bayes false discovery rate (see Efron [4]) can be defined as the posterior probability of nullity

\[
\text{lbfdr}(z_i) = \mathbb{P}(\text{null}|Z_i = z_i) = \frac{(1 - 2p)f_0(z_i)}{pf_{-1}(z_i) + pf_1(z_i) + (1 - 2p)f_0(z_i)}.
\]

However, unlike the binary decision problems, the \( \text{lbfdr}(z_i) \) does not directly involve in the oracle rule in Theorem 3.1, and this reflects another challenge of the three-decisions problem.

In our problem, we want to ensure the equal numbers of chosen under-expressed and over-expressed genes, and thus we actually have the following “symmetric decision” constraint on the decision rule \( \delta = (\delta_1, \ldots, \delta_m) \):

\[
\sum_{i=1}^m I(\delta_i = -1) = \sum_{i=1}^m I(\delta_i = 1).
\] (4)

Note that the symmetric constraint (4) is also valid even if the underlying true model for \( \theta_i \)'s is asymmetric. That is, we want to find the best policy among those symmetric policies. This is in parallel in the offline estimation problem where one is interested in restricting to unbiased estimators, and wanting to find the best unbiased estimators.

For the oracle rule in Theorem 3.1, let \( \hat{k}' = \sum_{i=1}^m I(\hat{\delta}_i = -1) \) and \( \hat{k} = \sum_{i=1}^m I(\hat{\delta}_i = -1) \). Clearly, if \( \hat{k}' = \hat{k} \), then the oracle rule in Theorem 3.1 is also the optimal rule under the constraint (4). Unfortunately, if \( \hat{k}' \neq \hat{k} \), it is non-trivial to minimize the Bayesian risk in (2) subject to the symmetric decision constraint in (4), as the corresponding optimal rule no longer has a simple form. Fortunately, we can still characterize some of its theoretical properties as follows.

**Theorem 3.2.** Let \( \delta^* \) be the optimal rule that minimizes the Bayesian risk (2) subject to the constraint (4) when \( \sigma = 2\lambda \). Assume \( \delta_{i_0}^* (\bar{z}) = -1 \) and \( \delta_{i_1}^* (\bar{z}) = 1 \) for some given index \( i_0 \) and \( i_1 \). Then for any \( k \) with \( \delta_k^* (\bar{z}) = 0 \) we have

\[
f_0(z_k) \left[ g(z_k) + \frac{1-2p}{\lambda p} \right] \geq f_0(z_{i_0}) \left[ g(z_{i_0}) + \frac{1-2p}{\lambda p} \right]
\]

and

\[
f_0(z_k) \left[ g(z_k) - \frac{1-2p}{\lambda p} \right] \leq f_0(z_{i_1}) \left[ g(z_{i_1}) - \frac{1-2p}{\lambda p} \right],
\]

where the function \( g(\cdot) \) is defined in (3).

The proof of this theorem follows directly from Theorem 3.1 by constructing a new rule that switches the decisions \( \delta_{i_0}^* \) and \( \delta_k^* \) (or \( \delta_{i_1}^* \) and \( \delta_k^* \)) and then using the fact that the Bayes risk of this new rule should be larger than that of \( \delta^* \). The detailed arguments are straightforward and thus omitted.

Other properties of the optimal rule \( \delta^* \) can also be obtained by comparing it with other decision rules. For instance, it is impossible for \( \delta^* \) to have \( \delta_{k_1}^* = \delta_{k_2}^* = 0 \) but \( g(z_{k_1}) \leq -(1 - 2p)/(\lambda p) \) and \( g(z_{k_2}) \geq (1 - 2p)/(\lambda p) \). This is because otherwise setting \( \delta_{k_1}^* = -1 \) and \( \delta_{k_2}^* = 1 \) will lead a new decision rule that also satisfies the constraint (4) but would have a smaller Bayes risk. Moreover, it is not difficult to see from the proofs of Theorem 3.2 that loosely speaking, the optimal rule \( \delta^* \) assigns \( \delta_i^* (\bar{z}) = -1 \) for those \( i \)'s that have small values of \( f_0(z_i)[g(z_i) + (1 - 2p)/(\lambda p)] \) and assigns \( \delta_i^* (\bar{z}) = 1 \) for those \( i \)'s that have large values of \( f_0(z_i)[g(z_i) - (1 - 2p)/(\lambda p)] \).

In the special case when \( g(z) \) in (3) is an increasing function of \( z \) while \( f_0(z) \) is an increasing function when \( z \leq 0 \) but a decreasing function when \( z \geq 0 \), the optimal rule \( \delta^* \) assigns \( \delta_i^* (\bar{z}) = -1 \)
or 1 corresponds to those i that have smaller and larger values of z_i’s, respectively, when z satisfies \(- (1 - 2p)/(\lambda p) \leq g(z) \leq (1 - 2p)/(\lambda p)\). Unfortunately, this result no longer holds when \(g(z) \leq - (1 - 2p)/(\lambda p)\) or \(g(z) \geq (1 - 2p)/(\lambda p)\), which can happen, say, in the scenario that we have chosen all \(k\) most negative genes, but there are \(k > \hat{k}\) genes that satisfies \(g(z_i) \geq (1 - 2p)/(\lambda p)\) in the oracle rule. In such scenario, maximizing \(\hat{f}_0(z_i)[g(z_i) - (1 - 2p)/(\lambda p)]\) could be different from maximizing \(z_i\).

Nevertheless, it is not bad to approximate \(\delta^*\) by a rule in the following subclass \(C\) of the decision rule satisfying (4): choose \(k\) “most positive” observations and \(k\) “most negative” observations when \(k(1 \leq k \leq m/2)\) significant pairs are chosen. In the next section, we will propose an empirical method on how to choose such \(k\).

4. Our proposed symmetric FDR method

In this section we focus on the subclass \(C\) of the decision rules that chooses \(k\) “most positive” observations and \(k\) “most negative” observations. Since all decision rules in the subclass \(C\) satisfy the constraint (4), it remains to decide how to select such \(k\) in order to minimize the Bayesian risk (2). Clearly, the optimal choice of \(k\) depend on the parameters of \(\lambda, p\) and \(f_{-1}, f_1\), which may be unknown in practice. In order to choose such \(k\) in practice, we explore the nested structure and will develop a data-driven method based on the “p-values” for a pair of positive and negative genes.

A high-level description of our proposed symmetric directional FDR algorithm is as follows. For the \(m\) by \(n\) data matrix \(X = (x_{ij})\), we first calculate the two-sample \(t\) statistic \(t_i\) for each row (or gene) as in (1) for \(i = 1, \ldots, m\). Then we rank these \(m\)-statistics in an increasing order from negative to positive: \(t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(m)}\), and pair them as \((t_{(1)}, t_{(m)}), (t_{(2)}, t_{(m-1)}), \ldots, (t_{(i)}, t_{(m+1-i)}), \ldots\) for \(i = 1, 2, \ldots, \lfloor m/2 \rfloor\). We then assign the \(p\)-values to each pair under the null hypothesis when there are no significant pairs, and select “significant pairs” via some standard FDR methods.

It remains to define the \(p\)-value for each two-dimensional vector \((t_{(i)}, t_{(m+1-i)})\). To do so, we propose to choose a suitable non-negative function \(g\), say \(g(t) = |t|\) or \(g(t) = t^2\), and then define the “score” of a pair \((t_{(i)}, t_{(m+1-i)})\) as \(S_i = g(t_{(i)}) + g(t_{(m+1-i)})\). In general, the function \(g(\cdot)\) can be any non-negative function satisfying (i) \(g(t) = g(-t)\) and (ii) \(g(t)\) is increasing function of \(t \geq 0\), as long as it will make sure that more extreme pairs of \((t_{(i)}, t_{(m+1-i)})\) will have larger scores. In our numerical studies below, we found that \(g(t) = |t|\) or \(t^2\) actually yield the same results on selected significantly genes despite the slight difference on the actual numerical \(p\)-values.

Next, we propose to derive the null distribution of the scores \(S_i\’s\) by using the column permutation method. To be more specific, we permute among \(n = n_1 + n_2\) columns to randomly label \(n_1\) and \(n_2\) columns as normal and cancer subjects, and then re-compute the pairs \((t_{(i)}, t_{(m+1-i)})\)’s and their corresponding scores \(S_i\’s\). Thus, for each of \(B\) (e.g., \(B = 200\)) permutations, we get a set of null score statistics, \(S_1^b, \ldots, S_{\lfloor m/2 \rfloor}^b\) for \(b = 1, \ldots, B\). The \(p\) value of the observed ith pair, \((t_i, t_{(m+1-i)})\), is then defined as

\[
p_i = \frac{B}{[m/2]B} \sum_{b=1}^{B} \# [k : S_k^b \geq S_i, k = 1, \ldots, [m/2]] = \frac{1}{[m/2]B} \sum_{b=1}^{B} \sum_{k=1}^{\lfloor m/2 \rfloor} 1 \{ S_k^b \geq S_i \}.
\]

With the defined \(p\)-values for the \([m/2]\) pairs, we can apply any standard FDR (or family-wise error rate control) method to select significant \(p\)-values, thereby yielding the selected significant pairs. For instance, let us consider the Benjamini–Hochberg (step-up) FDR procedure, and our proposed symmetric algorithm is as follows. For the \([m/2]\) ordered pairs \((t_{(i)}, t_{(m+1-i)})\’s\), we obtain the \([m/2]\) ordered \(p\)-values: \(p_1 \leq p_2 \leq \cdots \leq p_{[m/2]}\). Set a FDR control level \(q\) (say \(q = 1\%\)), find \(k^*\) = the largest \(k\) such that \(p_k \leq \frac{k}{[m/2]} \cdot q\), and then declare significant genes as all those genes corresponding to \(t_{(i)}\) (under-expressed) or \(t_{(m+1-i)}\) (over-expressed) for \(i = 1, \ldots, k^*\). Hence, our proposed algorithm will select \(k^*\) under-expressed genes and \(k^*\) over-expressed genes.
Table 1
Simulation Results. Monte Carlo mean with standard deviations in parentheses based on 1000 repetitions. When the observed t-statistics are asymmetric and skewed, the results show that the standard FDR methods are inappropriate due to a large number of false under-expressed genes.

<table>
<thead>
<tr>
<th></th>
<th># of true over-expressed</th>
<th># of false over-expressed</th>
<th># of true under-expressed</th>
<th># of false under-expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>The standard FDR method</td>
<td>100 (0.0)</td>
<td>0.02 (0.16)</td>
<td>100 (0.0)</td>
<td>57.3 (10.8)</td>
</tr>
<tr>
<td>Our proposed FDR method</td>
<td>100 (0.0)</td>
<td>2.1 (1.1)</td>
<td>100 (0.0)</td>
<td>2.1 (1.1)</td>
</tr>
</tbody>
</table>

5. Numerical studies

In this section we run numerical simulations to illustrate the usefulness of our proposed symmetric directional FDR method when the truly significant genes are symmetric in the sense that the number of over-expressed genes is the same as the number of under-expressed genes.

The summary statistics of the real data in Section 2 motivate us to generate simulated $m \times n$ data matrix $X = (X_{ij})$ under the null hypothesis as follows. Assume $m = 10,000$ and $n = 50$, and the first 25 subjects are the normal (control) group and the last 25 subjects consist of the cancer (treatment) group. For the first $0.736m = 7360$ rows_genes (i.e., for $1 \leq i \leq 7360$), we assume $X_{ij}$ are iid $N(2.73, 0.16)$ for the normal group but iid $N(2.61, 0.07)$ for the cancer group. This corresponds to the center of the first cluster in the previous discussion that accounts for 73.6% of all data. Likewise, for the next $0.178m = 1780$ rows (i.e., for $7361 \leq i \leq 7360 + 1780 = 9140$), we assume $X_{ij}$ are iid $N(5.17, 0.81)$ if $1 \leq j \leq 25$ and iid $N(4.96, 0.62)$ if $26 \leq j \leq 50$. For the last $0.086m = 860$ rows (i.e., for $9141 \leq i \leq 10000$), we assume $X_{ij}$ are iid $N(8.17, 0.99)$ if $1 \leq j \leq 25$ and iid $N(8.03, 0.81)$ if $26 \leq j \leq 50$. Note that our simulation setting mimics the real data in Section 2 in the sense that under the null hypothesis, the distribution of the two-sample t-statistic $t_i$'s in (1) is asymmetric and skewed.

Given the $m \times n$ data matrix under the null hypothesis, we are now ready to generate significant genes/rows under the alternative hypothesis. Among $m = 10000$ genes/rows, we randomly select two subsets of genes/rows and each subset contains $k^+ = 100$ genes/rows. For these two subsets, we change the values in the cancer groups (e.g., $x_{ij}$ for $26 \leq j \leq 50$) to reflect the significant genes. One subset is treated as “over-expressed” genes/rows in the sense that $x_{ij}$’s is replaced by $x_{ij} + 3$ for all $26 \leq j \leq 50$, and the other subset is thought of as “under-expressed” genes/rows in the sense that $x_{ij}$’s is replaced by $x_{ij} - 3$ for all $26 \leq j \leq 50$. This leads to a desired $m \times n$ data matrix $X = (X_{ij})$. This process is repeated 1000 times in our simulations.

For each of 1000 simulated $m \times n$ data matrices $X = (X_{ij})$, two FDR methods are applied to select significant genes under the FDR = 1%; one is the standard Benjamini–Hochberg FDR method with column permutations to generate $p$-values for each individual gene, and the other is our proposed symmetric directional FDR method with the Benjamini–Hochberg FDR method applied to the pair $p$-values via column permutations. For each of simulated $m \times n$ data matrices $X = (X_{ij})$ and for each of these two FDR methods, we record the numbers of true over-expressed and true under-expressed genes as well as the numbers of false over-expressed and false under-expressed genes. The overall results are summarized in Table 1.

It is clear from Table 1 that when the null distribution of the $t$-statistics is skewed to the left, the standard FDR method did a good job to select significant over-expressed genes, but did a very poor job to select significant under-expressed genes. This demonstrates that the standard FDR methods are inappropriate if the observed $t$-statistics are asymmetric and skewed. Meanwhile, our proposed symmetric directional FDR method borrows the evidence from over-expressed genes to make a much better decision on under-expressed genes, which greatly improve the overall decision.

6. Application

To further illustrate the usefulness of our proposed symmetric directional FDR method, we apply it to several real data sets in the literature, see Storey and Tibshirani [17] and Efron [4]. For the purpose of comparison, we also report the results of the standard Benjamini–Hochberg FDR method. To make...
a fair comparison, all $p$-values for individual genes or paired genes were computed via the column permutation approaches.

**Example 1.** The Prostate Cancer Data Set from Singh et al. [14]. It consists of 6033 genes for 102 men, among which, 52 prostate cancer patients and 50 healthy controls. With FDR = 10%, the standard Benjamini–Hochberg FDR method chose 64 significant genes: 31 over-expressed and 33 under-expressed. Meanwhile, our proposed symmetric directional FDR method selected 35 for each category, and included all 64 significant genes chosen by the standard FDR method. In this example, the empirical null distribution of $t$-statistics $t_i$’s in (1) seems to be symmetric, and there is little difference between the standard FDR method and our proposed symmetric directional FDR method.

**Example 2.** The Leukemia data set from Golub et al. [5]. There are $m = 7219$ genes on $n = 72$ subjects, with 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) arrays. With FDR = 1%, the standard method selects 1624 “significant” leukemia genes: 716 over-expressed genes and 908 under-expressed genes. Our proposed symmetric directional FDR method selects 1612 “significant” leukemia genes with 806 for over-expressed and 806 for under-expressed genes. In particular, all 716 over-expressed genes from the standard FDR method are also chosen by our symmetric directional FDR method, whereas all 806 under-expressed genes from our symmetric directional FDR method are also chosen by the standard FDR method.

In this case, our proposed symmetric directional FDR method and the standard FDR method are complementary to each other, and can also be combined to make decisions: while there are $908 - 716 = 192$ more significant under-expressed genes than over-expressed genes by the standard FDR method, our symmetric directional FDR method suggests that as many as $908 - 806 = 102$ under-expressed genes are likely insignificant, as the standard FDR methods might be misleading here due to the asymmetric properties of the observed $t_i$’s values.

**Example 3.** The Breast Cancer Microarray Data from Hedenfalk et al. [6]. It consists of 3,226 genes on 7 BRCA1 and 8 BRCA2 arrays. The measurements exceeding 20 have been deleted and log transformation with base 2 has been made to the data.

With FDR = 5%, the standard FDR method selects 94 significant genes: 26 under-expressed and 68 over-expressed. On the one hand, our proposed symmetric directional FDR method selects 110 significant genes: 55 over-expressed and 55 under-expressed. Note that all 26 under-expressed genes from the standard FDR method are also chosen by our symmetric directional FDR method, while all 55 over-expressed genes from our symmetric directional FDR method are selected by the standard FDR method. In other words, compared to the standard FDR method, our symmetric directional FDR method adds more under-expressed genes at the expense of some over-expressed genes.

7. Discussion

In this paper, we proposed a symmetric directional FDR method that pairs the over-expressed and under-expressed genes, and defines the $p$-values for the pairs instead of the individual genes. The standard FDR method is then applied to select significant pairs, thereby yielding equal numbers of over-expressed and under-expressed genes.

Our proposed method is designed for the scenario when it is desirable to have a balance between the over-expressed genes and under-expressed genes among the selected significant genes. A challenge that still remains is how to get the information on the balance or rather imbalance of the over-expressed genes and under-expressed genes in a given data analysis. The short answer is that it depends on the specific application and data set. From the biology viewpoint, it depends on whether the researchers use any specific domain knowledge to purposely choose negatively or positively expressed genes in their experiments. Or whether the biological difference between two groups affects gene pathways in such a way that there are significantly more negatively or positively expressed genes. From the statistical viewpoint, there are two main factors. The first one is whether the gene set is large or not. As mentioned in Rahmatallah, Emmert-Streib, and Glazko [12], small
gene sets have higher probability of large imbalance between the over-expressed genes and under-expressed genes than large gene sets. Note that the gene set in our motivating example is much larger than the public data sets available in the FDR literature, and our biologist collaborators feel the balance makes more sense in such a large gene set. The other factor is whether the observed t-statistics are highly skewed or not. In such case, if the standard FDR method yields highly unbalanced results that are suspicious or questionable, our proposed symmetric directional FDR method might be a reasonable alternative statistical tool.

There are several interesting problems that deserve further research. First, it will be interesting to see how efficient the proposed symmetric directional FDR method is as compared to the oracle rule in Theorem 3.1. Can we characterize the optimal symmetric policy in the context of Theorem 3.2? Intuitively, another possible symmetric directional FDR approach is to choose two different cut-off values, \(-c_1 < 0 < c_2\), and declare a gene is significantly under-expressed if \(t_i < -c_1\) and significantly over-expressed if \(t_i > c_2\). Mathematically, the choice of two cut-off values \(-c_1\) and \(c_2\) should allow us to make a symmetric decision, though it is unclear how to choose these two values simultaneously, especially when the null distribution of \(t_i\)'s is asymmetric or skewed. It is also interesting to compare this alternative method with our proposed symmetric directional FDR method, which is computationally simpler. Moreover, the idea of selecting significant paired genes can easily be extended to selecting significant gene blocks where each block consists of \(k_1\) negatively expressed genes and \(k_2\) positively expressed genes, but further theoretical investigation is needed. Finally, we hope that this paper opens new research directions on FDR or multiple testing problems by extending large-scale binary decision problems to large-scale three or multiple decision problems.

Acknowledgments

The authors thank Professors John McDonald and Yuhong Fan for sharing with us their expertise in cancer and cell biology. S. E. Holte’s research was supported in part by University of Washington Center for AIDS Research NIH P30 AI027757. E. K. Lee’s research was partially supported by NIH 1R01 DA030996-01, NSF IIP-0832390 and IIP-1361532. Y. Mei was supported in part by NSF grants DMS-0954704 and CMMI-1362876.

Appendix. Proof of Theorem 3.1

By the prior distribution of \(\theta_i\)'s, the Bayesian risk in (2) can be re-written as:

\[
R(\hat{\delta}) = \sum_{a_1 \in \{-1,0,1\}} \ldots \sum_{a_m \in \{-1,0,1\}} \int_{z_1} \ldots \int_{z_m} \left[ L(\theta_i = a_i, \delta_i(\tilde{z})) f_{a_i}(z_i) P(\theta_i = a_i) \right] dz_1 \ldots dz_m
\]

\[
= \int_{z_1} \ldots \int_{z_m} \left[ L(\theta_i = -1, \delta_i(\tilde{z})) f_{-1}(z_i) p + L(\theta_i = 0, \delta_i(\tilde{z})) f_0(z_i)(1 - 2p) \right]
\]

\[
+ L(\theta_i = 1, \delta_i(\tilde{z})) f_1(z_i)p \right] dz_1 \ldots dz_m.
\]

Thus \(\hat{\delta}(\tilde{z}) = (\hat{\delta}_1(\tilde{z}), \ldots, \hat{\delta}_m(\tilde{z}))\) is Bayes rule if and only if, for every observed vector \((Z_1, \ldots, Z_m) = (z_1, \ldots, z_m)\), it assigns a decision \(\hat{\delta}(\tilde{z}) = d\) which minimizes \((d \in \{-1, 0, 1\})\)

\[
h_i(d) = L(\theta_i = -1, d) f_{-1}(z_i)p + L(\theta_i = 0, d) f_0(z_i)(1 - 2p) + L(\theta_i = 1, d) f_1(z_i)p.
\]

Next, by the definition of individual loss function \(L(\theta_i, d_i)\), we have

\[
h_i(d) = \begin{cases} 
(1 - 2p)f_0(z_i) + \sigma pf_1(z_i), & \text{if } d = -1; \\
\lambda pf_{-1}(z_i) + \lambda pf_1(z_i), & \text{if } d = 0; \\
\sigma pf_{-1}(z_i) + (1 - 2p)f_0(z_i), & \text{if } d = 1.
\end{cases}
\]
Hence, the Bayes rule makes a decision \( \hat{\delta}_i(\bar{z}) = -1 \) if and only if \( h_i(d = -1) \leq h_i(d = 0) \) and \( h_i(d = 0) \leq h_i(d = -1) \), or equivalently, if and only if
\[
(1 - 2p)f_0(z_i) + \sigma pf_1(z_i) \leq \lambda pf_{-1}(z_i) + \lambda pf_1(z_i) \quad \text{and} \quad \sigma p f_1(z_i) \leq \sigma pf_{-1}(z_i).
\]
When \( \sigma = 2\lambda > 0 \) and \( 0 < p < 1/2 \), the above inequalities become
\[
\frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \leq -\frac{1 - 2p}{\lambda p} \quad \text{and} \quad f_1(z_i) \leq f_{-1}(z_i).
\]
When \( 0 < p < 1/2 \), the second inequality is implied by the first inequality, and thus the Bayes rule makes a decision \( \hat{\delta}_i(\bar{z}) = -1 \) if and only if
\[
\frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \leq -\frac{1 - 2p}{\lambda p}.
\]
Likewise, the Bayes rule makes a decision \( \hat{\delta}_i(\bar{z}) = 1 \) if and only if
\[
\frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \geq \frac{1 - 2p}{\lambda p},
\]
and makes a decision \( \hat{\delta}_i(\bar{z}) = 0 \) if and only if
\[
-\frac{1 - 2p}{\lambda p} \leq \frac{f_1(z_i) - f_{-1}(z_i)}{f_0(z_i)} \leq \frac{1 - 2p}{\lambda p},
\]
completing the proof of Theorem 3.1.

References