Marquette University
e-Publications@Marquette

Mathematics, Statistics and Computer Science Faculty Research and Publications

Mathematics, Statistics and Computer Science, Department of (-2019)

8-1-2018

Empirical Bayesian Approach to Testing Multiple Hypotheses with Separate Priors for Left and Right Alternatives

Naveen K. Bansal
Marquette University, naveen.bansal@marquette.edu

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

Steven J. Schrodi
University of Wisconsin-Madison

Published version. *Statistical Applications in Genetics and Molecular Biology*, Vol. 17, No. 4 (August 2018). DOI. © 2018 by Walter de Gruyter GmbH. Ised with permission

Naveen K. Bansal[1] / Mehdi Maadooliat[1,2] / Steven J. Schrodi[2]

Empirical Bayesian approach to testing multiple hypotheses with separate priors for left and right alternatives

[1] Department of Mathematics, Statistics, and Computer Sciences, Marquette University, Milwaukee, WI 53201-1881, USA, E-mail: mehdi@mscs.mu.edu. http://orcid.org/0000-0002-5408-2676.
[2] Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA, E-mail: mehdi@mscs.mu.edu. http://orcid.org/0000-0002-5408-2676.

Abstract:
We consider a multiple hypotheses problem with directional alternatives in a decision theoretic framework. We obtain an empirical Bayes rule subject to a constraint on mixed directional false discovery rate ($mdFDR \leq \alpha$) under the semiparametric setting where the distribution of the test statistic is parametric, but the prior distribution is nonparametric. We proposed separate priors for the left tail and right tail alternatives as it may be required for many applications. The proposed Bayes rule is compared through simulation against rules proposed by Benjamini and Yekutieli and Efron. We illustrate the proposed methodology for two sets of data from biological experiments: HIV-transfected cell-line mRNA expression data, and a quantitative trait genome-wide SNP data set. We have developed a user-friendly web-based shiny App for the proposed method which is available through URL https://npseb.shinyapps.io/npseb/. The HIV and SNP data can be directly accessed, and the results presented in this paper can be executed.

Keywords: directional alternatives, EM algorithm, false discovery rates, HIV, SNP
DOI: 10.1515/sagmb-2018-0002

1 Introduction

Many genetics and molecular/cellular biology studies require multiple hypothesis testing with directional alternatives, where, for example, the left directional and the right directional are associated with gene down-regulation and up-regulation, respectively. A traditional way to solve this problem is to use a two-sided test procedure with threshold determined based on a controlled error rate and the directional decision made based on the sign of the test statistics. This procedure is generally optimal under various directional false discovery rates (DFDRs) if the left and right directional alternatives are equally likely (Shaffer 2002). However, this assumption may not be satisfied for many biological applications. For example, a gene mutation (either experimental or natural) may result in more genes under-expressed than over-expression due to suppression of one more key regulatory genes. In addition, the pathways of under-expressed and over-expressed genes may be different. In such a case, considering separate priors for the left and right tails alternatives will be reasonable.

The problem can be described in terms of parameters θ_i, $i = 1, 2, \ldots, m$, as the directional hypotheses

$$H_i^{(0)} : \theta_i = 0 \text{ vs. } H_i^{(-1)} : \theta_i < 0 \text{ or } H_i^{(1)} : \theta_i > 0,\ i = 1,\ 2,\ \ldots m. \tag{1}$$

Suppose, $\theta_i s$, $i = 1, 2, \ldots, m$, are assumed to be generated from

$$\pi(\theta_i) = p_- \pi_-(\theta_i) + p_0 I(\theta_i = 0) + p_+ \pi_+(\theta_i), \tag{2}$$

where $p_- + p_0 + p_+ = 1$, and $\pi_-(\theta)$ and $\pi_+(\theta)$ are densities with supports $(-\infty, 0)$ and $(0, \infty)$, respectively. Note that prior (2) can be interpreted as $p_- = \Pr(H_i^{(-1)})$, $p_+ = \Pr(H_i^{(1)})$, $p_0 = \Pr(H_i^{(0)})$, $\pi_-(\theta_i) = \pi(\theta_i | H_i^{(-1)})$, and $\pi_+(\theta_i) = \pi(\theta_i | H_i^{(1)})$.

Mehdi Maadooliat is the corresponding author.
©2018 Walter de Gruyter GmbH, Berlin/Boston.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

The idea behind prior (2) is that it allows, under the non-null case, some components as negatively affected and some as positively affected, and it also allows separate priors for the positive and negatively affected components with non-symmetry controlled by p_- and p_+. This phenomenon occurs in many genetic studies where, under the influence of an external factor, some genes may be up-regulated and some may be down-regulated, and down-regulation may be more prevalent than up-regulation or *vice versa*. Consider, for example, the HIV data described in van't Wout et al. (2003). In this study, Van't Wout and colleagues measured, through a cDNA array, the mRNA expression changes in established cell lines during infection with HIV *in vitro*, compared to mRNA preparation from the cell lines under control conditions. The tranfection of HIV virus mediates differential expression of several host genes, some of which are up-regulated and others down-regulated. The regulatory impact of the HIV transfection is dependent upon disparate molecular mechanisms–host genes overexpressed following transfection may result from the activation of immunological defense pathways (e.g. upregulation of interferon-encoding genes in dendritic cells), whereas some genes involved in steroid and fatty acid biosynthesis may exhibit suppressed expression levels. Additionally, expression of cell surface proteins are known to be modified by HIV infection. Hence, a separate priors are required for up-regulated and down-regulated genes with a higher number of genes down-regulated than up-regulated. In that case, one would have $p_- > p_+$. We will analyze this data under the prior (2) with p_-, $\pi_-(\theta)$, p_+, and $\pi_+(\theta)$ estimated from the data itself. Note that frequentist's approach to this problem assumes equal weight to up-regulated and down-regulated genes as the direction of the selected genes are determined by the signs of the test statistics; see Benjamini and Yekutieli (2005). The proposed methodology gives different weights. We will demostrate that this approach produces more powerful rules.

Another example, we consider, is concerning genome-wide SNP data. To study the genetic architecture of body fat distribution, Shungin et al. (2015) performed a large-scale genome-wide association study in an effort to identify SNPs with genotypes significantly correlated with waist-to-hip ratio. Following standard QC procedures used in genome-wide association studies, the statistical test employed by the authors was a linear regression model, adjusted for confounding covariates. The genotypes are positively or negatively correlated depending upon whether the minor allele genotype increases or decreases the waist-to-hip phenotype. As rare alleles carry higher probabilities of conferring dysfunction in both the regulation and function of genes compared to alleles segregating at higher frequencies in the population, it is unreasonable to assume that the non-null SNPs will be symmetrically distributed in their effects on phenotypes. We will analyze this data using prior (2) with parameters of the prior estimated from the data itself.

Both of these examples demand that the left-tailed and right-tailed $\theta_i s$ should not be symmetrically distributed, and separate priors must be assigned to left and right tails.

Variety of procedures on Bayesian decision theoretic based multiple testing have been proposed in the literature over the past decade (see, e.g. Do, Muller & Tang, 2005; Muller, Parmigiani & Rice, 2006; Bogdan, Ghosh & Tokdar, 2008), but there is no systematic method to incorporate the unequal prior weights for the alternative tails. In this work, by considering free priors for left-tailed and right-tailed $\theta_i s$, we can extract the non-symmetric distribution for the non-null $\theta_i s$. As a consequence, we claim that the proposed Bayes solution yields better power of true discoveries.

In Section 2, we discuss briefly the Bayesian decision theoretic methodology. The details can be found in Bansal and Miescke (2013) and Bansal, Hamedani, and Maadooliat (2016). In Section 3, we develop empirical Bayesian methodology and discuss how to estimate the parameters (p_-, p_0, p_+), σ^2, and the nonparametric densities. A simulation study comparing the proposed rules against rules proposed by Benjamini and Yekutieli (2005) and Efron (2007a, 2007b), and Bansal, Hamedani, and Maadooliat (2016) are presented in Section 4. Analyses of HIV and genome-wide SNP, and miroRNA data sets are presented in Section 5. We end with some concluding remarks in Section 6.

2 Bayesian decision theoretic methodology

Let $\boldsymbol{d} = (d_1, d_2, \ldots, d_m)$ with $d_i \in \{-1, 0, 1\}$ denote a selection rule, where $d_i = -1$ means that $H_i^{(-1)}$ is selected, $d_i = 0$ means $H_i^{(0)}$ is selected, and $d_i = 1$ means $H_i^{(+1)}$ is selected. Let $\boldsymbol{u} = (u_1, u_2, \ldots, u_m)$ be the vector of true hypotheses defined similarly, i.e. $u_i = -1$ if $\theta_i < 0$, $u_i = 0$ if $\theta_i = 0$, and $u_i = +1$ if $\theta_i > 0$. Consider the loss

$$L(\boldsymbol{\theta}, d) = \sum_{i=1}^{m} L_i(\theta_i, d_i), \tag{3}$$

where L_i is an individual loss for the ith component, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. If the loss is a "0-1" loss, then (3) will be

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

$$L_0(\boldsymbol{\theta}, d) = \sum_{i=1}^{m} \sum_{j=-1}^{1} I(d_i = j) I(u_i \neq j),$$

where function $I(\cdot)$ denotes the indicator function. Note that the "0-1" loss yields the number of false discoveries and the number of false non-discoveries. Thus, minimizing the expected loss would yield the minimum number of expected false discoveries and expected false non-discoveries. If an objective is to also control false discoveries, then appropriate control can be imposed as we illustrate below.

Let the observed data $\boldsymbol{X} \sim P(\boldsymbol{x}; \boldsymbol{\theta}, \eta)$, where η denotes the set of all nuisance parameters including nonparametric components. The Bayes rule can be obtained by minimizing the posterior expected loss. Define,

$$v_i^{(-1)} = P(H_i^{(-1)} \mid \boldsymbol{x}),\ v_i^{(0)} = P(H_i^{(0)} \mid \boldsymbol{x}),\ \text{and}\ v_i^{(1)} = P(H_i^{(+1)} \mid \boldsymbol{x}).$$

Since the posterior expected loss for the "0-1" loss is $E[L_0(\theta, d) \mid x] = \sum \sum_{i,j} I(d_i = j)(1 - v_i^{(j)})$, it is easy to see that the Bayes rule, for all i, selects according to maximum of $\{v_i^{(j)},\ j = -1,\ 0,\ 1\}$.

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ with $x_i \sim f(x \mid \theta_i)$, $i = 1, 2, \ldots, m$ independently distributed, then

$$v_i^{(-1)} = \frac{p_- f_-(x_i)}{f(x_i)},\ v_i^{(1)} = \frac{p_+ f_+(x_i)}{f(x_i)},\ v_i^{(0)} = \frac{p_0 f_0(x_i)}{f(x_i)}, \tag{4}$$

where,

$$f_-(x) = \int_{-\infty}^{0} f(x \mid \theta) \pi_-(\theta) d\theta,\ f_+(x) = \int_{0}^{\infty} f(x \mid \theta) \pi_+(\theta) d\theta,\ f_0(x) = f(x \mid 0) \tag{5}$$

and $f(x_i) = p_- f_-(x_i) + p_+ f_+(x_i) + p_0 f_0(x_i)$. Since, maximizing $v_i^{(0)}$, $v_i^{(-1)}$, $v_i^{(1)}$ is equivalent to maximizing $p_0 f_0(x)$, $p_- f_-(x)$, and $p_+ f_+(x)$, the Bayes rule can be described as follows:

Reject $H_i^{(0)}$ *if*

$$T_-(x_i) = \int_{-\infty}^{0} \frac{f(x_i \mid \theta)}{f(x_i \mid 0)} \pi_-(\theta) d\theta > \frac{p_0}{p_-},\ T_+(x_i) = \int_{0}^{\infty} \frac{f(x_i \mid \theta)}{f(x_i \mid 0)} \pi_+(\theta) d\theta > \frac{p_0}{p_+}. \tag{6}$$

When $H_i^{(0)}$ *is rejected, select* $H_i^{(-1)}$ *or* $H_i^{(1)}$ *according to the maximum of* $p_- f_-(x)$ *and* $p_+ f_+(x)$. *Otherwise, select* $H_i^{(0)}$.

If $f(x \mid \theta)$ holds the MLR property, which we assume throughout the paper, then it is easy to see that $T_-(x)$ is a monotonically decreasing function, and $T_+(x)$ is a monotonically increasing function. Then (6) can be written as

$$x_i < T_-^{-1}\left(\frac{p_0}{p_-}\right) \text{ and } x_i > T_+^{-1}\left(\frac{p_0}{p_+}\right). \tag{7}$$

As an example, when $f(x \mid \theta)$ is normal density, π_+ and π_- are standard half-normal distribution, and when $(p_-, p_0, p_+) = (0.15, 0.80, 0.05)$, then (7) yields $x_i < -1.31$ and $x_i > 1.94$. Note, here, an asymmetric structure of the Bayes rule.

The above unconstrained Bayes rule does not necessarily control a false discovery rate. However, in practice, it is important that a decision rule controls a false discovery rate so that, whenever left or right discoveries are made, one would have confidence in those discoveries. This can be achieved by minimizing the posterior expected loss subject to a constraint on a false discovery rate. Since, we are interested in directional discoveries, we use mixed directional false discovery rate (mdFDR), which is defined as (see, Benjamini and Yekutieli 2005)

$$\text{mdFDR} = E\left[\frac{\sum_{i=1}^{m} \{I(d_i = -1) I(u_i = 0,\ 1) + I(d_i = +1) I(u_i = -1,\ 0)\}}{(|D_-| + |D_+|) \vee 1}\right], \tag{8}$$

where $D_{\mp}$ is the set of indices of selected $H_i^{(\mp 1)}s$, $|\cdot|$ denotes the cardinality of the set, and the expectation is with respect to both $\boldsymbol{X}$ and $\boldsymbol{\theta}$. The posterior version of mdFDR is

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

$$\text{mdPFDR} = 1 - \frac{\sum_{i=1}^{m}\{I(d_i = -1)v_i^{(-1)} + I(d_i = +1)v_i^{(1)}\}}{(|D_-| + |D_+|) \vee 1}. \tag{9}$$

Thus, the problem now is to find the Bayes rule that minimizes $E[L_0(\theta, d) \mid x] = m - \sum_i \sum_j I(d_i = j)v_i^{(j)}$ subject to the constraint mdPFDR$\leq\alpha$. We call this rule as a constrained Bayes rule. As described in Bansal and Miescke (2013), this can be achieved as follows:

Let $D_B^{(-1)} = \{i : v_i^{(-1)} > v_i^{(1)}, v_i^{(0)}\}$, and $D_B^{(1)} = \{i : \nu_i^{(1)} > \nu^{(-1)}, \nu^{(0)}\}$ be the sets of indices of the selected $H_i^{(-1)}s$ and $H_i^{(1)}s$ under the unconstrained Bayes rule, respectively. Define $\xi_i = v_i^{(-1)}$ for $i \in D_B^{(-1)}$, and $\xi_i = \nu_i^{(1)}$ for $i \in D_B^{(1)}$, and then rank all ξ_i, $i \in D_B^{(-1)} \cup D_B^{(1)}$ from the lowest to the highest. Let the ranked values be denoted by $\xi_{[1]} \leq \xi_{[2]} \leq \ldots \leq \xi_{[\hat{k}]}$, where $\hat{k} = |D_B^{(-1)} \cup D_B^{(1)}|$. Denote

$$\hat{i}_0 = \max\left\{j \leq \hat{k} : \frac{1}{j}\sum_{i=1}^{j} \xi_{[\hat{k}-i+1]} \geq 1 - \alpha\right\}. \tag{10}$$

Let D_ξ denotes the set of indices corresponding to $\xi_{[\hat{k}]} \geq \xi_{[\hat{k}-1]} \geq \cdots \geq \xi_{[\hat{k}-\hat{i}_0+1]}$. Now select $H_i^{(-1)}$ for $i \in D_B^{(-1)} \cap D_\xi$, and $H_i^{(1)}$ for $i \in D_B^{(1)} \cap D_\xi$.

3 Empirical Bayes rule with nonparametric alternatives

We first discuss the estimation of the nonparametric components. When $f(x \mid \theta)$ belongs to a location family, the estimation of π_- and π_+ can be obtained through deconvolution methods as described in Lee et al. (2010, 2013), and van Es, Gugushvili, and Spreij (2008). However, our Bayesian methodology described in Section 2 does not require the direct estimation of π_- and π_+. All we need are the estimations of $v_i^{(-1)}$, $v_i^{(0)}$, and $v_i^{(1)}$ as defined in (4), which involve $p_-, p_0, p_+, f_-(x), f_0(x)$, and $f_+(x)$. Note that the observations x_i, $i = 1, 2, \ldots, m$ are marginally distributed with the mixture density $f(x) = p_-f_-(x) + p_+f_+(x) + p_0f_0(x)$. This mixture density cannot be treated as a regular mixture of three distributions since $f_-(x), f_+(x)$, and $f_0(x)$ have specific characteristics. Before we discuss an estimation approach, we want to make a note that $f_-(x), f_+(x)$, and $f_0(x)$ are uniquely represented in the mixture density of $f(x)$. Due to MLR property of $f(x \mid \theta)$, as discussed in Section 2, $f_-(x)/f_0(x)$ (say $h_-(x)$) is a decreasing function of x, and $f_+(x)/f_0(x)$ (say $h_+(x)$) is an increasing function of x. This implies that $f_-(x) = f_0(x)h_-(x)$ and $f_+(x) = f_0(x)h_+(x)$ are distinct and different from $f_0(x)$. Thus, $f_-(x), f_+(x)$, and $f_0(x)$ in the mixture density of $f(x)$ are uniquely characterized.

We use a discretized approach of Lee et al. (2013) by discretizing $\pi_+(\theta)$ (and similarly $\pi_-(\theta)$) at the equally spaced points $\{h, 2h, 3h, \ldots, Bh\}$, where the bandwidth $h > 0$ and the number of bins B are appropriately determined. Thus, we consider

$$\pi(\theta) \approx p_0 I(\theta = 0) + p_- \sum_{k=1}^{B} \omega_{-1,k} I(\theta = -kh) + p_+ \sum_{k=1}^{B} \omega_{1,k} I(\theta = kh),$$

where $\sum \omega_{-1,k} = \sum \omega_{1,k} = 1$. Note that we could have taken different bandwidths h and different number of bins B for $\pi_-(\theta)$ and $\pi_+(\theta)$, but that would make computation difficult when applying AIC criterion. Since, the AIC picks the highest possible B based on the observed data, the magnitudes of the weights $\omega_{-1,k}$ and $\omega_{1,k}$ would still reflect the nature of $\pi_-(\theta)$ and $\pi_+(\theta)$, respectively.

To describe the empirical Bayesian methodology, for simplicity, we will assume that the observed $X_i \mid \theta_i \sim N(\theta_i, \sigma^2)$; however, the methodology described below can be applied to any location-scale family. In addition, we also assume that X_i, $i = 1, 2, \ldots, m$ are independently distributed. Thus, the observed X_i, $i = 1, 2, \ldots, m$ are i.i.d. with marginal density approximated by

$$f(x) \approx p_0 \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) + p_- \sum_{k=1}^{B} \omega_{-1,k} \frac{1}{\sigma}\phi\left(\frac{x + kh}{\sigma}\right) + p_+ \sum_{k=1}^{B} \omega_{1,k} \frac{1}{\sigma}\phi\left(\frac{x - kh}{\sigma}\right).$$

We now use the EM algorithm to estimate (p_0, p_-, p_+), σ^2, and weights $\omega_{j,k}$, $k = 1, 2, \ldots, B$, $j = -1, 1$.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

Let $(\hat{p}_0^{(r-1)}, \hat{p}_-^{(r-1)}, \hat{p}_+^{(r-1)})$, $\hat{\sigma}^{(r-1)}$, and $\hat{\omega}_{j,k}^{(r-1)}$, $j = -1, 1$, $k = 1, 2, \ldots, B$ be the estimates from the rth iteration ($r \geq 1$). Denote

$$\hat{p}_{0i}^{(r)} = \frac{\hat{p}_0^{(r-1)}\phi\left(\frac{x_i}{\hat{\sigma}^{(r-1)}}\right)}{\hat{Q}^{(r-1)}(x_i)}, \ \hat{p}_{(-)ki}^{(r)} = \frac{\hat{p}_-^{(r-1)}\hat{\omega}_{-1,k}^{(r-1)}\phi\left(\frac{x_i+kh}{\hat{\sigma}^{(r-1)}}\right)}{\hat{Q}^{(r-1)}(x_i)},$$
$$\hat{p}_{(+)ki}^{(r)} = \frac{\hat{p}_+^{(r-1)}\hat{\omega}_{1,k}^{(r-1)}\phi\left(\frac{x_i-kh}{\hat{\sigma}^{(r-1)}}\right)}{\hat{Q}^{(r-1)}(x_i)},$$

where,

$$\hat{Q}^{(r-1)}(x_i) = \hat{p}_0^{(r-1)}\phi\left(\frac{x_i}{\hat{\sigma}^{(r-1)}}\right)$$
$$+ \sum_{k=1}^{B}\left[\hat{p}_-^{(r-1)}\hat{\omega}_{-1,k}^{(r-1)}\phi\left(\frac{x_i+kh}{\hat{\sigma}^{(r-1)}}\right) + \hat{p}_+^{(r-1)}\hat{\omega}_{1,k}^{(r-1)}\phi\left(\frac{x_i-kh}{\hat{\sigma}^{(r-1)}}\right)\right].$$

Following the standard E-step and M-step, the EM algorithm yields the following estimates

$$\hat{p}_0^{(r)} = \frac{1}{m}\sum_{i=1}^{m} p_{0i}^{(r)}, \ \hat{p}_-^{(r)} = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{B}\hat{p}_{(-)ki}^{(r)}, \ \hat{p}_+^{(r)} = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{B}\hat{p}_{(+)ki}^{(r)},$$
$$\hat{\omega}_{-1,k}^{(r)} = \frac{\sum_{i=1}^{m}\hat{p}_{(-)ki}^{(r)}}{\sum_{i=1}^{m}\sum_{\tilde{k}=1}^{B}\hat{p}_{(-)\tilde{k}i}^{(r)}}, \ \hat{\omega}_{1,k}^{(r)} = \frac{\sum_{i=1}^{m}\hat{p}_{(+)ki}^{(r)}}{\sum_{i=1}^{m}\sum_{\tilde{k}=1}^{B}\hat{p}_{(+)\tilde{k}i}^{(r)}}$$

and

$$\hat{\sigma}^{2(r)} = \frac{1}{m}\sum_{i=1}^{m}\left[\hat{p}_{0i}^{(r)}x_i^2 + \sum_{k=1}^{B}\hat{p}_{(-)ki}^{(r)}(x_i+kh)^2 + \sum_{k=1}^{B}\hat{p}_{(+)ki}^{(r)}(x_i-kh)^2\right].$$

A higher value of B (and a smaller h) would amount to overfitting. As discussed in Lee et al. (2013), this can be resolved by maximizing the likelihood function subject to a penalty function, where a higher penalty is assigned for overfitting. We use the AIC criterion,

$$\mathrm{AIC}(B) = 2\sum_{i=1}^{m}\log\hat{f}(x_i) - 2(2B+2),$$

by maximizing AIC(B) with respect to B. Note that $(2B+2)$ is the number of parameters, and $\hat{f}(x) = \hat{p}_-\hat{f}_-(x) + \hat{p}_0\hat{f}_0(x) + \hat{p}_+\hat{f}_+(x)$, where $\hat{f}_0(x) = \frac{1}{\hat{\sigma}}\phi\left(\frac{x}{\hat{\sigma}}\right)$, and the estimates $\hat{f}_+(x)$ and $\hat{f}_-(x)$ are given by

$$\hat{f}_+(x) = \sum_{k=1}^{B}\hat{\omega}_{-1,k}\frac{1}{\hat{\sigma}}\phi\left(\frac{x-kh}{\hat{\sigma}}\right), \text{ and } \hat{f}_-(x) = \sum_{k=1}^{B}\hat{\omega}_{1,k}\frac{1}{\hat{\sigma}}\phi\left(\frac{x+kh}{\hat{\sigma}}\right) \tag{11}$$

Note that the value h depends on B, which can be chosen in such a way that the left end distribution $N(-Bh, \hat{\sigma}^2)$ covers the lowest observed value $x_{[1]}$, and the right end distribution $N(Bh, \hat{\sigma}^2)$ covers the largest observed value $x_{[m]}$. For example, h can be the smallest value such that $x_{[m]} < Bh + 3\hat{\sigma}$, and $x_{[1]} > -Bh - 3\hat{\sigma}$.

Once, we have estimates of (p_-, p_0, p_+), π_-, π_+, and σ, the empirical Bayes rule can be obtained by the Bayesian methodology developed in Section 2 using these estimates. For example, it is easy to see from (6) that

$$\hat{T}_-(x) = \sum_{k=1}^{B}\hat{\omega}_{-1,k}\phi\left(\frac{x+kh}{\hat{\sigma}}\right)/\phi\left(\frac{x}{\hat{\sigma}}\right), \text{ and } \hat{T}_+(x) = \sum_{k=1}^{B}\hat{\omega}_{1,k}\phi\left(\frac{x-kh}{\hat{\sigma}}\right)/\phi\left(\frac{x}{\hat{\sigma}}\right),$$

and

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

$$\hat{v}_i^{(-1)} = \frac{\hat{p}_-\hat{T}_-(x_i)}{\hat{p}_0 + \hat{p}_-\hat{T}_-(x_i) + \hat{p}_+\hat{T}_+(x_i)},\ \hat{v}_i^{(1)} = \frac{\hat{p}_+\hat{T}_+(x_i)}{\hat{p}_0 + \hat{p}_-\hat{T}_-(x_i) + \hat{p}_+\hat{T}_+(x_i)},$$

$$\text{and } \hat{v}_i^{(0)} = 1 - \hat{v}_i^{(-1)} - \hat{v}_i^{(1)}$$

As shown in Lee et al. (2013), the estimated parameters and estimated densities will be consistent as $m \to \infty$. Since, in applications dealing with high dimensional multiple hypothesis problems, m is likely to be very large (in thousands), the proposed empirical Bayes procedure is likely to be closed to the Bayes optimal.

Remark 2:

Note that the above methodology is dictated by the weights $\hat{\omega}_{-1,k}$ and $\hat{\omega}_{1,k}$. If suppose the weights are negligible on the left and right extremes, then there will be few instances (x values) when $\hat{T}_-(x)$ and $\hat{T}_+(x)$ will cross the threshold values. On the other hand, if the weights are high on the left and right extremes, there will be many instances when $\hat{T}_-(x)$ and $\hat{T}_+(x)$ will cross the threshold values. The weights can also be used to determine the pattern of the discovered components if there is any visible pattern in the weights. For example, a bimodal pattern in $\{\omega_{-1,k}, k = 1, 2, \ldots, B\}$ or in $\{\omega_{1,k}, k = 1, 2, \ldots, B\}$ would be an indication two sets of clusters of the discovered components.

4 Monte Carlo simulations

In this section, we compare the performance of the proposed nonparametric empirical Bayes procedure (NPB) with the directional procedure (BY) proposed by Benjamini and Yekutieli (2005), a local false discovery rate (LFDR) rule proposed by Efron (2007a,b), and a skew normal Bayes rule (SNB) proposed by Bansal, Hamedani, and Maadooliat (2016). Note that BY and LFDR are nonparametric procedures against the choice of alternative priors. SNB procedure was developed under the skew normal alternative prior; however, it may be interesting to see its performance when the true alternative prior is not skew normal. In order to consider the robustness against different types of alternative priors, we simulate the test statistics $z_i \sim N(\theta_i, 1)$, $i = 1, 2, \ldots, m$ with $[mp_0]$ of $\theta_i = 0$ and the remaining $m - [mp_0]$ of θ_is generated from set of distributions as described below.

In the setups below, we use $\gamma_- = p_-/(1-p_0)$ and $\gamma_+ = p_+/(1-p_0)$ which are the expected proportions of rejected nulls belonging to the left and right tails, respectively, and denote $\pi_\gamma = \gamma_-\pi_-(\theta) + \gamma_+\pi_+(\theta)$ as the prior under the alternative.

- Setup 1:

$$\text{MSN}.j : \text{mixture of } 2j \text{ skew normals } \gamma_- \sum_{k=1}^{j} \text{SN}(-k\xi,\ \sigma^2,\ \lambda) + \gamma_+ \sum_{k=1}^{i} \text{SN}(k\xi,\ \sigma^2,\ \lambda)$$

  For the definition and concept of skew normal distribution, see Azzalini (1985). We present the results for this setup in the online Supplementary Materials

- Setup 2:

$$\begin{cases} \text{MN : mixture of normals} & \gamma_- \cdot N(-\mu,\ \sigma_1^2) + \gamma_+ \cdot N(\mu,\ \sigma_1^2), \\ \text{SDE : shifted double exponential} & \gamma_- \cdot \lambda e^{\lambda(y+1)} I(-\infty,\ -1) + \gamma_+ \cdot \lambda e^{-\lambda(y-1)} I(1,\ \infty), \\ \text{SN : skew normal} & \text{SN}(0,\ \sigma_2^2,\ \eta). \end{cases}$$

  Throughout this simulation study, we assume $\mu = 2$, $\lambda = \sigma_1 = 1$, $\sigma_2 = 2.5$ and consider different combinations of p_0 and (γ_-, γ_+) from the following ranges: $p_0 = 0.8, 0.9, 0.95$, and $(\gamma_-, \gamma_+) = (0.15, 0.85), (0.3, 0.7), (0.5, 0.5), (0.7, 0.3), (0.85, 0.15)$. For the skew normal alternatives, the shape parameter η is chosen in a way as to match the associated weights (γ_-, γ_+). In other word, the shape parameter η satisfies the following relationships:

$$\int_{-\infty}^{0} f_{\text{SN}}(x|\eta)dx = \gamma_- \text{ and } \int_{0}^{\infty} f_{\text{SN}}(x|\eta)dx = \gamma_+,$$

  where $f_{\text{SN}}(x\,|\,\eta)$ is the probability density function of the $\text{SN}(0,\ \sigma_2^2,\ \eta)$. For the above setting of (γ_-, γ_+), the values of η are $\eta = 1.96, 0.73, 0, -0.73, -1.96$, respectively. Figure 1 illustrates the distributions of the generating models for $p_0 = 0.8$ and $(\gamma_-, \gamma_+) = (0.15, 0.85)$.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

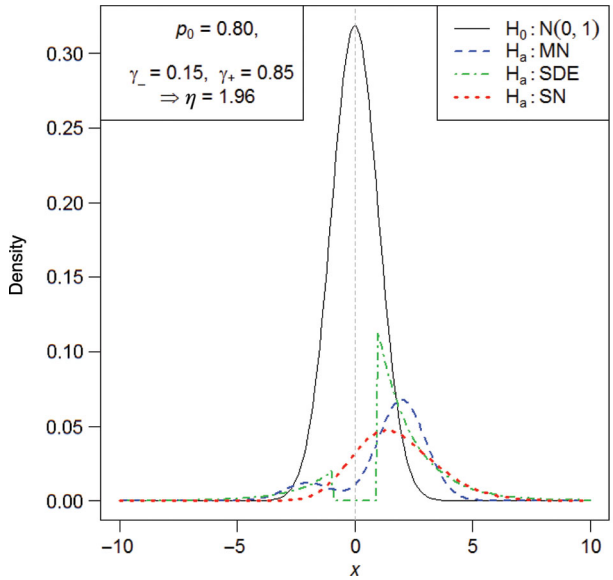


Figure 1: Graph of the null distribution and three choices for alternatives, mixture of normal (MN), shifted double exponential (SDE), and skew normal (SN) for $p_0 = 0.8$, and $(\gamma_-, \gamma_+) = (0.15, 0.85)$.

From each of the above combinations, we simulate 100 datasets with $m = 10{,}000$. Empirical estimates of the parameters were obtained first, and AIC criterion was used to determine the optimal B. The empirical estimates were then plugged in for the parameters before applying the algorithm discussed in Section 2 to obtain the empirical Bayes rule.

Benjamini and Yekutieli's procedure (BY) is performed using p-values based on z_i and the the direction of the alternative determine by sign of z_i as described in their paper. LFDR procedure is based on the local false discovery rate with null hypothesis rejected if LFDR(z_i) < 0.20 as suggested by Efron, and the direction of the alternative determined by the sign of z_i. For SNB procedure, parameters of skew normal alternative are estimated first before applying the procedure proposed in Bansal, Hamedani, and Maadooliat (2016). For comparison, we use the following two measurements:

a. Correct Discovery: The mean number of correct discoveries observed in the simulations. To avoid potential confusion with averaging over number of the hypotheses, m, we refer to this measurement as the expected number of correct discoveries.

b. False Discovery Rate: The expected ratio of false discoveries to the total number of discoveries.

To compare the procedures in left (right) directions, we further consider the expected number of correct left (right) discoveries and the expected rate of false left (right) discoveries. Figure 2 represents the aforementioned measurements for different values of $p_0 = (0.8, 0.9, 0.95)$, $\gamma_- = (0.15, 0.3, 0.5, 0.7, 0.85)$ and the generating model SDE. Expected number of right discoveries and the expected rate of false right discoveries are not presented since they can be interpreted from the results of the left discoveries. Similar Figures associated to other two generating models (MN and SN) can be found in the Supplementary Materials.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

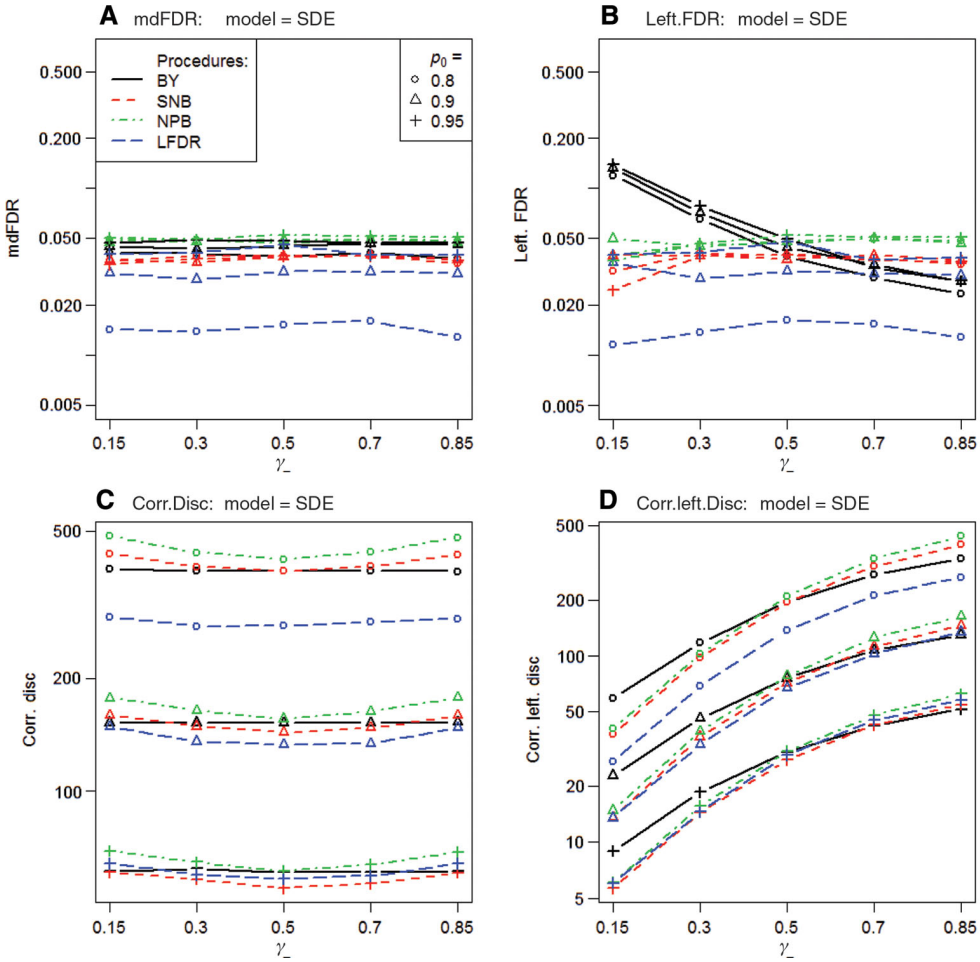


Figure 2: Comparison of the discoveries for the proposed nonparametric procedure (NPB) and the BY, SNB and LFDR procedures: (A) mixed directional false discovery rates (mdFDR), (B) false left discovery rate, (C) expected number of correct discoveries, (D) expected number of correct left discoveries.

Figure 2 shows that, overall, the results of the proposed Empirical Bayes rule NPB is superior to BY, LFDR, and SNB procedures for the shifted double exponential alternative prior as the parameters (γ_-, γ_+) departs from (0.5, 0.5). There are higher number of correct discoveries by NPB with false discovery rate close to pre-assigned rate of 0.05. Note that the mdFDR and expected number of correct discoveries for BY are flat with varying (γ_-, γ_+). The mdFDR for LFDR procedure is lower than 0.05 for smaller p_0 values, however, it approached 0.05 as p_0 approaches 1. The figures for left discoveries show a gain made in expected number of correct discoveries when γ_- is greater than 0.5 while keeping the false left discovery rates comparable with the BY procedure. Note also that the false left discovery rates are much higher under BY rule for smaller γ_-.

We see the same pattern for MN and SN priors (see the supplementary material). The advantage of the proposed NPB procedure is more prominent as (γ_-, γ_+) gets further away from (0.5, 0.5). In addition to this, we also tried mixture of skew normal priors with number of mixtures of 2, 4, and 6. For details, see the Supplementary Materials. This clearly shows an optimal robustness of the proposed NPB procedure in the sense that NPB allows a high number of correct discoveries subject to the constraint of mdFDR ≤ 0.05 irrespective of the alternative prior as if it possesses a skewness.

For further illustration, we compare the performance of the NPB, LFDR, SNB, and BY procedures in Table 1 based on 100 simulations for models MN, SDE, SN, $\gamma_- = 0.3, 0.5, 0.85$, and $p_0 = 0.8, 0.9, 0.95$. Some notable results of Table 1 are enumerated below:

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Table 1: Comparison of BY, SNB, NPB and LFDR rules for model = (MN, SDE, SN), $\gamma_- = 0.3, 0.5, 0.85$ and $p_0 = 0.8, 0.9, 0.95$ based on expected number of correct discoveries and false discovery rates for 100 simulations.

Model	γ_-	p_0	Correct discovery				False discovery rate			
			BY	SNB	NPB	LFDR	BY	SNB	NPB	LFDR
MN	0.30	0.80	655.0 (3.2)	671.4 (2.9)	689.6 (3.2)	555.8 (4.7)	0.0406 (0.0007)	0.0420 (0.0007)	0.0474 (0.0007)	0.0183 (0.0007)
		0.90	282.3 (2.1)	278.7 (1.9)	291.1 (2.1)	258.8 (2.6)	0.0454 (0.0012)	0.0398 (0.0011)	0.0494 (0.0011)	0.0295 (0.0014)
		0.95	121.3 (1.3)	116.0 (1.2)	123.2 (1.2)	115.1 (1.4)	0.0463 (0.0022)	0.0352 (0.0016)	0.0476 (0.0019)	0.0365 (0.0022)
	0.50	0.80	656.1 (2.8)	662.0 (2.5)	679.9 (2.9)	551.5 (4.3)	0.0392 (0.0008)	0.0409 (0.0008)	0.0462 (0.0009)	0.0184 (0.0007)
		0.90	279.9 (2.0)	272.4 (1.8)	283.8 (2.0)	253.5 (2.6)	0.0442 (0.0011)	0.0394 (0.0011)	0.0474 (0.0012)	0.0291 (0.0012)
		0.95	122.3 (1.3)	116.0 (1.1)	122.5 (1.2)	116.4 (1.5)	0.0464 (0.0019)	0.0373 (0.0017)	0.0474 (0.0018)	0.0387 (0.0017)
	0.85	0.80	651.7 (2.9)	694.9 (2.9)	713.7 (3.0)	580.9 (4.4)	0.0402 (0.0007)	0.0419 (0.0006)	0.0476 (0.0007)	0.0191 (0.0007)
		0.90	282.6 (2.5)	291.6 (2.2)	304.5 (2.5)	269.5 (2.8)	0.0434 (0.0012)	0.0390 (0.0010)	0.0467 (0.0012)	0.0274 (0.0012)
		0.95	120.7 (1.3)	121.2 (1.2)	128.7 (1.3)	119.9 (1.5)	0.0473 (0.0018)	0.0374 (0.0016)	0.0472 (0.0018)	0.0360 (0.0021)
SN	0.30	0.80	537.6 (3.1)	586.9 (2.9)	579.0 (3.2)	465.8 (3.4)	0.0387 (0.0008)	0.0478 (0.0008)	0.0451 (0.0008)	0.0165 (0.0007)
		0.90	227.6 (1.8)	242.9 (1.7)	241.5 (1.8)	214.7 (2.1)	0.0456 (0.0013)	0.0494 (0.0013)	0.0487 (0.0012)	0.0303 (0.0014)
		0.95	98.8 (1.1)	102.5 (1.1)	103.8 (1.1)	96.6 (1.3)	0.0494 (0.0022)	0.0502 (0.0021)	0.0524 (0.0021)	0.0411 (0.0023)
	0.50	0.80	530.9 (2.7)	556.9 (2.5)	549.4 (2.8)	438.2 (4.0)	0.0402 (0.0008)	0.0492 (0.0009)	0.0463 (0.0009)	0.0186 (0.0007)
		0.90	227.6 (2.3)	231.4 (2.0)	231.2 (2.2)	203.8 (2.5)	0.0438 (0.0013)	0.0474 (0.0013)	0.0470 (0.0013)	0.0283 (0.0015)
		0.95	99.8 (1.3)	99.5 (1.2)	99.9 (1.3)	94.5 (1.3)	0.0464 (0.0021)	0.0467 (0.0018)	0.0481 (0.0020)	0.0387 (0.0023)
	0.85	0.80	536.1 (2.7)	638.1 (2.5)	627.7 (2.7)	496.7 (3.3)	0.0405 (0.0009)	0.0509 (0.0008)	0.0476 (0.0008)	0.0173 (0.0007)
		0.90	231.0 (1.9)	267.3 (1.8)	266.7 (1.9)	234.3 (1.9)	0.0465 (0.0015)	0.0483 (0.0011)	0.0479 (0.0012)	0.0268 (0.0012)
		0.95	99.0 (1.2)	112.6 (1.2)	113.3 (1.2)	103.9 (1.2)	0.0496 (0.0022)	0.0472 (0.0020)	0.0483 (0.0021)	0.0327 (0.0021)
SDE	0.30	0.80	390.3 (2.8)	400.4 (2.4)	437.1 (2.8)	277.2 (3.6)	0.0406 (0.0010)	0.0390 (0.0008)	0.0487 (0.0009)	0.0138 (0.0008)
		0.90	152.7 (1.6)	149.2 (1.5)	164.3 (1.6)	135.7 (2.1)	0.0431 (0.0016)	0.0357 (0.0014)	0.0479 (0.0017)	0.0284 (0.0017)
		0.95	61.9 (1.0)	57.9 (1.1)	64.4 (1.0)	59.5 (1.2)	0.0488 (0.0026)	0.0377 (0.0022)	0.0494 (0.0025)	0.0410 (0.0026)
	0.50	0.80	390.4 (2.5)	389.7 (2.3)	418.8 (2.5)	277.6 (3.6)	0.0390 (0.0010)	0.0393 (0.0010)	0.0466 (0.0011)	0.0151 (0.0008)
		0.90	152.1 (1.8)	143.6 (1.5)	156.4 (1.7)	132.9 (2.1)	0.0451 (0.0017)	0.0388 (0.0017)	0.0491 (0.0017)	0.0316 (0.0018)
		0.95	60.4 (1.0)	54.9 (1.1)	60.9 (1.0)	58.0 (1.2)	0.0484 (0.0030)	0.0390 (0.0028)	0.0524 (0.0031)	0.0461 (0.0033)
	0.85	0.80	389.0 (2.5)	431.1 (2.5)	479.9 (2.9)	290.2 (3.2)	0.0378 (0.0010)	0.0353 (0.0008)	0.0481 (0.0010)	0.0128 (0.0008)
		0.90	153.1 (1.9)	159.4 (1.7)	178.1 (1.8)	147.9 (2.0)	0.0458 (0.0019)	0.0373 (0.0015)	0.0472 (0.0015)	0.0308 (0.0017)
		0.95	60.9 (0.9)	60.2 (1.0)	68.6 (1.0)	63.8 (1.2)	0.0472 (0.0029)	0.0368 (0.0021)	0.0507 (0.0021)	0.0398 (0.0024)

Values within parentheses are standard errors.

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

1. While mdFDR is almost always controlled by $q = 0.05$ level by all procedures, the expected number of correct discoveries by the NPB procedure is generally higher than other procedures, especially when there is high skewness in the alternative priors.

2. For larger values of p_0, expected number of discoveries are smaller by all methods, but the proposed NPB method is still better than others.

A similar table for mixture of skew normal alternatives with different number of mixtures can be found in Supplementary Materials.

5 Real data

We applied the proposed method to (i) a gene expression HIV study by van't Wout et al. (2003); and (ii) a genome-wide SNP association study on the waist-to-hip ratio (WHR) for European-ancestry females by Shungin et al. (2015). Van't Wout and colleagues measured, through a cDNA array, the mRNA expression changes in established cell lines during transfection with HIV *in vitro*, compared to mRNA preparation from the cell lined under control conditions. The genome-wide SNP association study examined the correlation between WHR measurements and SNP genotypes in a large number of samples.

Note that our methodology is described in terms of test statistics $X_i \sim N(\theta_i, \sigma^2)$, $i = 1, 2, \ldots, m$. However, any test statistics can be converted to normal variate at least under the null. Let T_i be the test statistic for testing ith hypothesis with $F(T_i)$ as its distribution function under the null. Let $X_i = \Phi^{-1}(F(T_i))$, where Φ is the distribution function of the $N(0, 1)$. Note that $X_i \sim N(0, 1)$ under the null. However, following Efron (2007a,b), we assume that $X_i \sim N(0, \sigma^2)$ under the null, where σ^2 may be <1 or >1 due to possible correlation among the components. Thus, we assume that $X_i \sim N(\theta_i, \sigma^2)$, $i = 1, 2, \ldots, m$, where $\theta_i s$ are independently distributed as (2).

Remarks 3:

a. In both datasets, we observe that the mean of the null-distribution needs to be shifted; otherwise, there are unusually high number of discoveries. Note that this phenomenon has also been reported by Efron (2007a,b). Thus, we assume that the null distribution is $N(\mu_0, \sigma^2)$, where $\mu_0 \neq 0$. In order to apply the approach of Section 4, where we assume that $\mu_0 = 0$, all we have to do is to transform $x \to x - \hat{\mu}_0$, where $\hat{\mu}_0$ is an estimate of μ_0. We obtain the estimate $\hat{\mu}_0$ using EM algorithm steps as discussed in Section 4. In each iterative step, we transform $x \to x - \hat{\mu}_0^{(r)}$, where

$$\hat{\mu}_0^{(r)} = \frac{\sum_{i=1}^m p_{0i}^{(r)} x_i}{\sum_{i=1}^m p_{0i}^{(r)}}, \; r \geq 1, \; \hat{\mu}_0^{(0)} = \bar{x}.$$

b. It is natural to expect a slight decay in the estimate of σ (the standard deviation under the null), as the number of bases (of the mixture components) B under the non-null distribution increases. However, for both data sets we considered the estimates of σ decreases dramatically as B increases. This may be due to the correlation among the test statistics as also hypothesized by Efron (2007a,b). The information criterions (i.e. AIC or BIC) picked the most complex model (with very large B and small σ), which lead to underestimation of the null and overestimation of the alternative distributions. To resolve this, we first estimated d μ_0 and σ of the null distribution from the simplest model (i.e. $B = 2$) and then estimated the alternative distributions (ω_k weights and B) using the criterion to pick the optimal model.

5.1 The HIV data

The data consists of mRNA expression levels from eight cDNA microarrays – four from cell lines transfected by HIV retrovirus and four from uninfected cell lines, each measuring expression levels of 7680 transcripts. For each gene, we obtain a two-sample t-statistic, comparing the infected versus the uninfected subjects, which is then transformed to a z-value, where $z_i = \Phi^{-1}\{F_6(t_i)\}$. Here $F_6(\cdot)$ denotes the cumulative distribution function (cdf) of t-distribution with 6 degrees of freedom, and Φ denotes the cdf of standard normal distribution. The histogram of the z-values is given in the left panel of Figure 3 with a nonparametric fit, which is picked by the AIC as proposed in Section 4.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

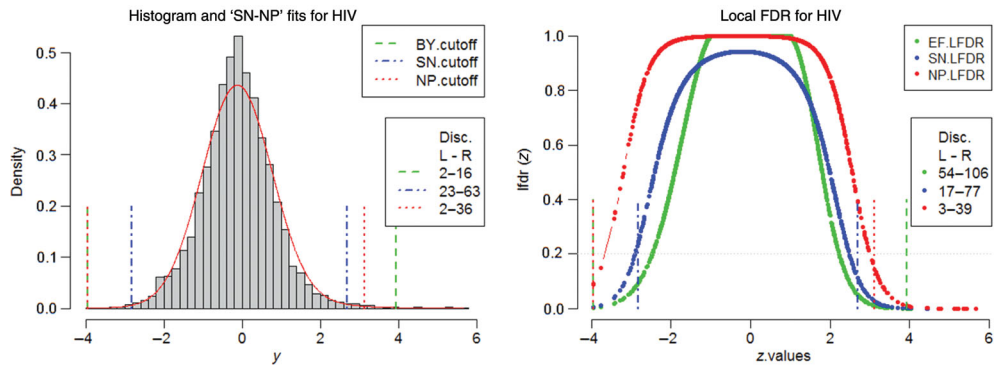


Figure 3: HIV-data: left panel – histogram with cut-off points by BY, SNB and the proposed Bayesian nonparametric method (NPB); right panel – local fdrs based on Efron's (EF.LFDR) and the new nonparametric approach (NP.LFDR) with cut-off points of BY, SNB and NPB.

The BY procedure resulted in cutoffs $(-3.94, 3.94)$, which resulted in 18 total discoveries with 2 genes declared as under-expressed and 16 as over-expressed. For the proposed Bayes rule, we used AIC that picks the model with $B = 4$ bases, where the parameter estimates from the EM algorithm are: $\hat{\mu}_0 = -0.127$, $\hat{\sigma} = 0.9$, $\hat{p}_0 = 0.984$, $\hat{p}_- = 0.003$, and $\hat{p}_+ = 0.014$. Next, we plugged in these estimates and use procedure discussed in Section 3 to obtain the Bayes rule. We ended up with cutoffs $(-3.94, 3.12)$ with total 38 discoveries (under-expressed genes: 2 and over-expressed genes: 36).

The HIV data was also studied by Efron (2007a,b) using a local false discovery approach. In analogy to Efron's local false discovery rate, we introduce

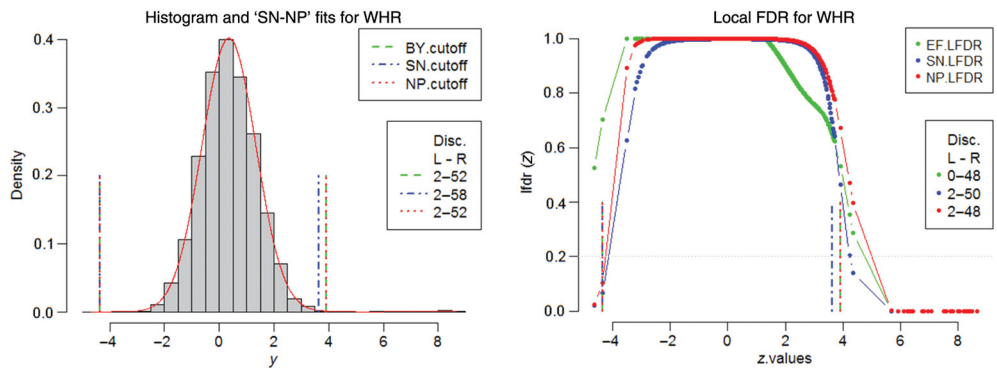
$$\mathrm{LFDR}(z) = \frac{p_0 f_0(z)}{p_- f_-(x_i) + p_+ f_+(x_i) + p_0 f_0(x_i)}, \tag{12}$$

where $f_0(z)$ is the estimated null density, and estimates of $f_-(x_i)$ and $f_+(x_i)$ are give by (11). The cutoff threshold $\mathrm{LFDR} \leq 0.2$ is used to discover infected genes (Efron 2007a, 2007b) with $H_i^{(-)}$ and $H_i^{(+)}$ discoveries made based on the sign of z_i.

Efron's LFDR approach resulted in 160 discoveries (left: 54 and right:106), while the LFDR approach based on (12) resulted in 42 discoveries (left: 3 and right: 39). It is interesting to note that there is not a large difference, in terms of the total number of discoveries, between the proposed approach (NPB) and the LFDR approach based on (12). Perhaps the Efron's LFDR approach over predicts the number of discoveries as it was also pointed out by Guttardo et al. (2006).

5.2 The WHR data

To study the genetic architecture of body fat distribution, Shungin et al. (2015) performed a large-scale genome-wide association study in an effort to identify SNPs with genotypes significantly correlated with waist-to-hip ratio. Following standard QC procedures used in genome-wide association studies, the statistical test employed by the authors was a linear regression model, adjusted for confounding covariates. A major finding of this and related studies is the chromosome 16 locus. Hence, we focused on genetic markers from chromosome 16 residing between 40 Mbp and 70 Mbp. The histogram of the test statistics is given in the left panel of Figure 4 with a nonparametric fit, which is picked by the AIC as proposed in Section 4.



Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

Figure 4: WHR-data: left panel – histogram with cut-off points by BY, SNB and the proposed Bayesian nonparametric method (NPB); right panel – local fdrs based on Efron's (EF.LFDR) and the new nonparametric approach (NP.LFDR) with cut-off points of BY, SNB and NPB.

The BY procedure resulted in 54 total discoveries where 2 SNPs negatively correlated, and 52 SNPs positively correlated with waist-to-hip ratios. For the proposed Bayes rule, we again used AIC and selected the model with $B = 4$ basis, where the parameter estimates from the EM algorithm are: $\hat{\mu}_0 = 0.353$, $\hat{\sigma} = 0.988$, $\hat{p}_0 = 0.996$, $\hat{p}_- = 0.0$, and $\hat{p}_+ = 0.004$. The Bayes procedure yielded total 54 discoveries, similar to the BY procedure.

Efron's LFDR approach resulted in 48 discoveries (left: 0 and right: 48). The LFDR approach based on (12) also resulted in 48 right discoveries, plus 2 left discoveries.

To assess the differential performance of the three methods, we explored the biological plausibility of SNPs that were identified in our method compared to Efron LFDR approach using the extreme BMI GWAS dataset. Our method discovered two left-tail (rs12446228 and rs1477196) and four right-tail (rs4783819, rs7190492, rs1861869 and rs11075986) SNPs, which were absent from the results of Efron LFDR method. All six of these SNPs reside within the *FTO* gene on chromosome 16q. *FTO* is a well-established obesity-associated gene (Peters, Ausmeier & Rüther, 1999; Fischer et al., 2009; Meyre et al. 2009; 2009). Further, all six of these SNPs are significantly correlated with SNP rs1421085 –also a SNP in *FTO*; the alleles segregating at which exhibit strong functional effects on adipocyte biology mediated through enhancer activity (Claussnitzer et al. 2015). Hence, we can reasonably conclude that our method identifies true positive results not discovered through Efron LFDR method.

6 Concluding Remarks

Generally, the hypotheses problems are stated as either one-tailed or two-tailed tests. However, many practical situations would dictate that one tail is more likely than the other tail under the alternative hypotheses. In such cases, one-tailed or two-tailed tests will not be appropriate. The proposed Empirical Bayesian methodology develops procedure for testing directional hypothesis with a mixture of null and nonparametric alternative with separate priors on the left and right tails. We showed in this paper that the use of such nonparametric mixture model yields a better power theoretically and empirically. The proposed Bayes rule is based on an optimality criterion, and it also control mixed directional false discovery rate (mdFDR). Although, we have mainly focused on the "0-1" loss, but any other loss function can be considered without much difficulties. In addition, as a by-product of the proposed nonparametric fit, a local false discovery rate approach is proposed, which can be considered as an alternative to Efron (2007a,b) nonparametric LFDR approach for directional hypotheses.

It is worthy to note that, we observed some level of inconsistency in the number of discoveries between the FDR controlled by the BY procedure (e.g. less discovery in HIV data and more discovery in WHR data) and the local fdr controlled via Efron's LFDR procedure. However, the proposed method yields consistent results between the FDR controlled via the new procedure (NPB) and the local fdr procedure based on (12).

A web application that can be used by the research community to reproduce the results in this paper or to use the proposed methodology for any other related applications is available at https://npseb.shinyapps.io/npseb/.

Supplementary material

The results of simulation (setup 2) is available with this paper at the web-based supplementary files.

Acknowledgements

The authors would like to thank two reviewers for the thoughtful comments that helped us to improve the paper.

References

Azzalini, A. (1985): "A class of distributions which includes the normal ones," Scand. J. Stat., 12, 171–178.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM

Bansal, N. K. and K. J. Miescke (2013): "A Bayesian decision theoretic approach to directional multiple hypotheses problems," J. Multivariate Anal., 120, 205–215.
Bansal, N. K., G. G. Hamedani and M. Maadooliat (2016): "Testing multiple hypotheses with skewed alternatives," Biometrics, 72, 494–502.
Benjamini, Y. and D. Yekutieli (2005): "False discovery rate controlling confidence intervals for selected parameters," J. Am. Stat. Assoc., 100, 71–80.
Bogdan, M., J. K. Ghosh and S. T. Tokdar (2008): "A comparison of the Benjamini–Hochberg procedure with some Bayesian rules for multiple testing," Math. Stat. Collect., 1, 211–230.
Claussnitzer, M., S. N. Dankel, K. H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I. S. Sousa, J. L. Beaudry, V. Puviindran, N. A. Abdennur, J. Liu, P.-A. Svensson, Y.-H. Hsu, D. J. Drucker, G. Mellgren, C.-C. Hui, H. Hauner and M. Kellis (2015): "FTO obesity variant circuitry and adipocyte browning in humans," N. Engl. J. Med., 373, 895–907.
Do, K., P. Muller and F. Tang (2005): "A Bayesian mixture model for differential gene expression," J. R. Stat. Soc. C-Appl., 54, 627–664.
Efron, B. (2007a): "Size, power and false discovery rates," Ann. Stat., 35, 1531–1377.
Efron, B. (2007b): "Correlation and large-scale simultaneous significance testing," J. Am. Stat. Assoc., 102, 93–103.
Fischer, J., L. Koch, C. Emmerling, J. Vierkotten, T. Peters, J. C. Brüning and U. Rüther (2009): "Inactivation of the FTO gene protects from obesity," Nature, 458, 894–898.
Guttardo, R., A. E. Raftery, K. Y. Yeung and R. E. Bumgarner (2006): "Bayesian robust inference for differential gene expression in microarrays with multiple samples," Biometrics, 62, 10–18.
Lee, M., H. Shen, C. Burch and J. S. Marron (2010): "Direct deconvolution density estimation of a mixture distribution motivated by mutation effects distribution," J. Nonparametr. Stat., 22, 1–22.
Lee, M., P. Hall, H. Shen, J. S. Marron, J. Tolle and C. Burch (2013): "Deconvolution estimation of mixture distributions with boundaries," Electron. J. Stat., 7, 323–341.
Locke, A. E., B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, et al. (2015): "Genetic studies of body mass index yield new insights for obesity biology," Nature, 518, 197–206.
Meyre, D., J. Delplanque, J. C. Chèvre, C. Lecoeur, S. Lobbens, S. Gallina, E. Durand, V. Vatin, F. Degraeve, C. Proença, S. Gaget, A. Körner, P. Kovacs, W. Kiess, J. Tichet, M. Marre, A.-L. Hartikainen, F. Horber, N. Potoczna, S. Hercberg, C. Levy-Marchal, F. Pattou, B. Heude, M. Tauber, M. I. McCarthy, A. I. F. Blakemore, A. Montpetit, C. Polychronakos, J. Weill, L. J. M. Coin, J. Asher, P. Elliott, M.-R. Järvelin, S. Visvikis-Siest, B. Balkau, R. Sladek, D. Balding, A. Walley, C. Dina and P. Froguel (2009): "Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations," Nat. Genet., 41, 157–159.
Muller, P., G. Parmigiani and K. Rice (2006): "FDR and Bayesian multiple comparisons rules," Proc. of the 8th Valencia World Meeting on Bayesian Statistics, Oxford University Press.
Peters, T., K. Ausmeier and U. Rüther (1999): "Cloning of Fatso (FTO), a novel gene deleted by the Fused toes (Ft) mouse mutation," Mamm. Genome, 10, 983–986.
Shaffer, J. P. (2002): "Multiplicity, directional (Type III) errors, and the null hypothesis," Psychol. Methods, 7, 356–369.
Shungin, D., T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Magi, et al. (2015): "New genetic loci link adipose and insulin biology to body fat distribution," Nature, 518, 187–196.
van Es, B., S. Gugushvili and P. Spreij (2008): "Deconvolution for an atomic distribution," Electron. J. Stat., 2, 265–297.
van't Wout, A. B., G. K., Lehrman, S. A. Mikheeva, G. C. O'Keeffe, M. G. Katze, R. E. Bumgarner, G. K. Geiss and J. I. Mullins (2003): "Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-cell lines," J. Virol., 77, 1392–1402.

Supplemental Material: The online version of this article offers supplementary material (DOI: https://doi.org/10.1515/sagmb-2018-0002).

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Brought to you by | North Carolina State University (NCSU) Libraries
Authenticated
Download Date | 11/7/18 2:00 PM