A New Approach to Multiple Testing of Grouped Hypotheses[☆]

Yanping Liu, Sanat K. Sarkar, Zhigen Zhao

Department of Statistics, Temple University, Philadelphia, PA, 19122, USA

Abstract

A two-fold loop testing algorithm (TLTA) is proposed for testing grouped hypotheses controlling false discoveries. It is constructed by decomposing a posterior measure of false discoveries across all hypotheses into within- and between-group components, allowing a portion of the overall FDR level to be used to maintain control over within-group false discoveries. Numerical calculations performed under certain model assumption for the hidden states of the within-group hypotheses show its superior performance over its competitors that ignore the group structure, especially when only a few of the groups contain the signals, as expected in many modern applications. We offer data-driven version of the TLTA by estimating the parameters using EM algorithms and provide simulation evidence of its favorable performance relative to these competitors. Real data applications have also produced encouraging results for the TLTA.

Keywords: False Discovery Rate, Grouped Hypotheses, Large-Scale Multiple Testing.

1. Introduction

Statistical queries in modern scientific investigations often involve simultaneous testing of multiple hypotheses appearing in non-overlapping groups. Such group formation occurs naturally in many of these investigations due to the underlying biological or experimental process or can be created to effectively capture certain specific features of the data. Whatever be the reasons for the hypotheses to form groups, ignoring the group structure when constructing multiple testing methods may result in misleading conclusions (Efron (2008)). A considerable amount of research has taken place in the development of multiple testing methods for grouped hypotheses both from frequentist and Bayesian perspectives (Benjamini & Heller (2007), Pacifico et al. (2004)), Subramanian et al. (2005), Heller et al. (2009)), Arbeitman et al. (2002), Calvano et al. (2005), Clements et al. (2011), Clements et al. (2014), Hu et al. (2010), Cai & Sun (2009) and Schildknecht et al. (2016)), mostly in the framework of controlling false discovery rate (FDR), as originally defined by (Benjamini & Hochberg (1995)), or its variants

[☆]The work of Yanping Liu is based on her Ph.D. thesis at Department of Statistics, Temple University. Sanat K. Sarkar is Professor of Department of Statistical Science, Temple University. Zhigen Zhao is Associate Professor of Department of Statistical Science, Temple University. Liu's research was supported in part by NSF Grants DMS-1208735 and DMS-1309273. Sarkar's research was supported by NSF Grants DMS-1208735 and DMS-1309273. Zhao's research was supported by NSF Grants DMS-1208735. E-mail addresses: lyping@temple.edu (Y. Liu), sanat@temple.edu (S.K. Sarkar), zhaozhg@temple.edu (Z. Zhao).

such as local FDR (Efron et al. (2001)) and marginal FDR (Sun & Cai (2007) and Sun & Cai (2009)).

A Bayesian approach can produce powerful method of controlling a rate of false discoveries when testing multiple hypotheses (He et al. (2015), Sun & Cai (2007), Sun & Cai (2009), Efron & Tibshirani (2002), Tang & Zhang (2007), Sarkar et al. (2008), and Newton et al. (2004)). With the posterior probability of a null hypothesis being true, often referred to as the local FDR under a simple mixture model (Efron et al. (2001)), as the key ingredient, such a method relies on controlling the cumulative mean of these posterior probabilities. While these methods are well developed for multiple testing of a single group of hypotheses, extension of the notion of local FDR from single to multiple groups, taking into account false discoveries both within and between groups given that the significance of a hypothesis in a group depends on whether or not the group itself is significant, has not been put forward yet, as far as we know. The novelty of our paper lies in considering such an extension and using that to develop a new method for testing grouped hypotheses integrating both within and between group discoveries.

Let X_{gj} be the observation corresponding to the jth hypothesis in the gth group and θ_{gj} indicate the truth ($\theta_{gj}=0$) or falsity ($\theta_{gj}=1$) of that hypothesis, for $g=1,\ldots,G; j=1,\ldots,m_g$. We express each θ_{gj} as follows: $\theta_{gj}=\theta_g\cdot\theta_{j|g}$, with $\theta_g=0$ (or $=1$) indicating that the gth group, and hence each (or at least one) of its component hypotheses, is non-significant (or significant), and $\theta_{j|g}$ having its interpretation as the hidden state for the jth hypothesis within the gth group depending on the status of that group. In other words, if $\theta_g=0$, then $\theta_{gj}=0$; and if $\theta_g=1$, then $\theta_{gj}=0$ or 1 according to whether $\theta_{j|g}=0$ or 1. This provides an explicit representation of the underlying group structure of the hidden states. It also leads us to the consideration of a similarly defined two-stage multiple decision rule $\delta_{gj}(\boldsymbol{X})=\delta_g(\boldsymbol{X})\cdot\delta_{j|g}(\boldsymbol{X})$, with $\delta_g(\boldsymbol{X})\in\{0,1\}$ and $\delta_{j|g}(\boldsymbol{X})\in\{0,1\}$ being the decision rules for θ_g and $\theta_{j|g}$, respectively, for the problem of deciding between $\theta_{gj}=0$ and $\theta_{gj}=1$ simultaneously for all (g,j). Our method provides such a two-stage rule controlling the FDR aposteriori at a given level α, that is, satisfying the following constraint:

$$E\left[\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}(1-\theta_{gj})\delta_{gj}(\boldsymbol{X})}{\left\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})\right\}\vee 1}\,\middle|\,\boldsymbol{X}\right]\leq\alpha, \tag{1.1}$$

(with $a\vee b=\max(a,b)$) on the expected false discovery proportion conditional on $\boldsymbol{X}=\{X_{gj}\}$.

The above two-stage representations of the hidden states and the decision rules allow us to express the conditional expectation in (1.1), we call it posterior total FDR (across all hypotheses), in terms of posterior FDRs within the truly significant groups ($\theta_g=1$). Given α at which the posterior total FDR is to be controlled, our method screens the hypotheses within each group at the first stage for possible rejections subject to a control over the posterior within-group FDR at a certain level less than or equal to α. At the second stage, it makes the final decision on ultimately rejecting these hypotheses if the groups containing them are identified as significant subject to the desired control over the posterior total FDR. We call this method the "Two-fold Loop Testing Algorithm (TLTA)" for grouped hypotheses. Thus, the TLTA enjoys, unlike the other available methods, the added flexibility in preserving a pre-chosen level of control

over false discoveries within each significant group along with controlling the false discoveries across all hypotheses. Moreover, with this within-group posterior FDR being defined in terms of strengths of evidence towards rejection for the hypotheses in a group conditional on that for the group itself, preserving a control over it presents an effective way of capturing the within-group dependencies caused inherently or structurally by the grouping.

We carried out numerical and simulation studies assessing the performance of the TLTA, both in terms of its oracle and data-driven versions, against its relevant competitors that completely ignore the group structure. These studies were conducted under a model setting that assumes independence between but not within groups and a truncated Bernoulli for the hidden states within each significant group. These studies have revealed superior performance of the TLTA in terms of FDR control and power (measured using false non-discoveries and the expected proportion of correctly rejected false nulls) over its competitors in many practical scenarios. When applied to the data from the Adequate Yearly Progress (AYP) study of California elementary schools in 2013 (http://www.cde.ca.gov/ta/ac/ay/aypdatafiles.asp) comparing the academic performance for socioeconomically advantaged (SEA) versus socioeconomically disadvantaged (SED) students, the TLTA also shows its favorable performance by making more discoveries than its competitors.

The remainder of the paper is organized as follows. We present the TLTA in Section 2. In Section 3, we introduce our model assumption, explicit formulas for the within- and between-group local fdr scores under this model, and steps of estimating the model parameters for derivation of data-driven versions of the TLTA and its competitors. The findings of numerical studies associated with the TLTA and its competitors are presented in Section 4 for their oracle and data-driven versions. The real data application is illustrated in Section 5. Concluding remarks are made in Section 6, and technical details are given in Appendix.

2. Methodolgies

As mentioned above, $\theta_{gj} = 0$ if and only if either $\theta_g = 0$ or $\theta_{j|g} = 0$ when $\theta_g = 1$. With that in mind, we define the following two quantities with the only model assumption made at this point that the hidden states are binary random variables:

$$fdr_g(\boldsymbol{X}) = P(\theta_g = 0|\boldsymbol{X}). \tag{2.1}$$

and

$$fdr_{j|g}(\boldsymbol{X}) = P(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{X}). \tag{2.2}$$

The first is the local fdr score for the gth group, which is measured by its posterior probability of being not significant and represents a key ingredient in discovering significant groups. The second is the posterior probability of $\theta_{j|g} = 0$ given that $\theta_g = 1$, which represents the conditional local fdr score for a hypothesis given that it is in a truly significant group and is a key ingredient in making discoveries within a significant group.

The posterior total FDR, denoted by $\mathrm{PFDR}_T(\boldsymbol{X})$, is given by

$$\begin{aligned} & \frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}(1-\theta_{gj})\delta_{gj}(\boldsymbol{X})}{\left\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})\right\}\vee 1} \\ = \quad & I(\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})>0) - \frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\theta_{gj}\delta_{gj}(\boldsymbol{X})}{\left\{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\delta_{gj}(\boldsymbol{X})\right\}\vee 1}, \end{aligned}$$

and so can be expressed as follows:

$$\begin{aligned} & I(\sum_{g}\sum_{j}\delta_{gj}(\boldsymbol{X})>0) - PFDR_T(\boldsymbol{X}) \\ = \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})P(\theta_g=1|\boldsymbol{X})\sum_j\delta_{j|g}(\boldsymbol{X})P(\theta_{j|g}=1|\theta_g=1,\boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1} \\ = \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})(1-fdr_g(\boldsymbol{X}))\sum_j\delta_{j|g}(\boldsymbol{X})(1-fdr_{j|g})(\boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1}, \\ = \quad & \frac{\sum_g\left\{\delta_g(\boldsymbol{X})(1-fdr_g(\boldsymbol{X}))[I(\sum_j\delta_{j|g}(\boldsymbol{X})>0)-PFDR_{W|g}(\boldsymbol{X})]\sum_j\delta_{j|g}(\boldsymbol{X})\right\}}{\left(\sum_g\delta_g(\boldsymbol{X})\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\right)\vee 1}, \end{aligned}$$

where

$$PFDR_{W|g} = \frac{\sum_j\delta_{j|g}(\boldsymbol{X})fdr_{j|g}(\boldsymbol{X})}{\left\{\sum_j\delta_{j|g}(\boldsymbol{X})\right\}\vee 1}$$

is the posterior FDR within the gth group that is truly significant. In other words,

$$PFDR_T(\boldsymbol{X}) = \frac{\sum_g\delta_g(\boldsymbol{X})\left\{1-(1-fdr_g(\boldsymbol{X}))(1-PFDR_{W|g}(\boldsymbol{X}))\right\}\sum_j\delta_{j|g}(\boldsymbol{X})}{\left(\sum_g\delta_g(\boldsymbol{X})\{\sum_j\delta_{j|g}(\boldsymbol{X})\}\right)\vee 1}. \tag{2.3}$$

This leads us to our proposed method in its oracle form, as stated in the following. For notational convenience, from this point onwards we will often suppress the symbol $\boldsymbol{X}$ in the quantities that obviously depend on the data.

Proposed Method: Two-Fold Loop Testing Algorithm (TLTA)

Step 1. For each g, let $fdr_{(1)|g}\le fdr_{(2)|g}\cdots\le fdr_{(m_g)|g}$ be the ordered $fdr_{j|g}$, with $H_{g(1)},\ldots,H_{g(m_g)}$ being the corresponding hypotheses, and find

$$R_g = \max\left\{k_g : \frac{1}{k_g}\sum_{j=1}^{k_g}fdr_{(j)|g}\le\eta\right\},$$

given $0<\eta\le\alpha<1$. Mark the hypotheses $H_{g(1)},\ldots,H_{g(R_g)}$ for possible rejection and go to the next step.

Step 2. Calculate $\eta_g = \frac{1}{R_g}\sum_{j=1}^{R_g} fdr_{(j)|g}$, and define $fdr_g^* = 1-(1-\eta_g)(1-fdr_g)$, for each g. Order these fdr_g^* values as $fdr_{(1)}^* \le \cdots \le fdr_{(G)}^*$, and find

$$l = \max\left\{k : \frac{\sum_{g=1}^{k} R_{(g)} fdr_{(g)}^*}{\sum_{g=1}^{k} R_{(g)}} \le \alpha\right\},$$

with $R_{(g)}$ being the value of R for the group that corresponds to $fdr_{(g)}^*$. The hypotheses that were marked for possible rejection in the groups $(1),\ldots,(l)$ are ultimately rejected.

Theorem 2.1. *Given any $0 < \eta \le \alpha < 1$, the TLTA for grouped hypotheses controls the PFDR$_T$ at level α.*

Proof. A proof of this theorem is immediate, since PFDR$_T$ of the TLTA is $\frac{\sum_{g=1}^{l} R_{(g)} fdr_{(g)}^*}{\sum_{g=1}^{l} R_{(g)}}$, which is less than or equal to α.

Remark 1. The TLTA is developed with special attention given to that the hypotheses are grouped, with each group having a significance probability of its own. It also takes into account the dependency, which could be naturally present or caused inherently due to grouping, between the significance of a hypothesis and that of the group containing it. More specifically, given data, (i) it measures strength of evidence towards rejection for each hypotheses within a group conditional on that for the group itself by using $1-fdr_{j|g}$, (ii) pulls up the hypotheses with the highest average measure of conditional evidence exceeding $1-\eta$ from each group, and (iii) then sets up a rejection rule for these selected sub-groups of hypotheses by taking into account the measures of evidence towards rejection for the respective groups subject to a control over total false discoveries at the desired level. It provides a new two-fold loop algorithm integrating both within and between group discoveries when testing multiple hypotheses that are grouped.

Let us consider testing a single group of hypotheses, say $H_{1(1)},\ldots,H_{1(m_1)}$, assuming that $G = 1$. Here, since $\eta_1 = \frac{1}{R_1}\sum_{j=1}^{R_1} fdr_{(j)|1}$, where $R_1 = \max\left\{k_1 : \frac{1}{k_1}\sum_{j=1}^{k_1} fdr_{(j)|1} \le \eta\right\}$, and $fdr_1^* = fdr_1 + (1-fdr_1)\frac{1}{R_1}\sum_{j=1}^{R_1} fdr_{(j)|1}$, the TLTA rejects the first R_1 of these hypotheses if

$$R_1 = \max\left\{k : \frac{1}{k}\sum_{j=1}^{k} fdr_{(j)|1} \le \min\left(\eta, [\alpha - fdr_1]/[1-fdr_1]\right)\right\}. \tag{2.4}$$

Our assumption that η should be restricted within the interval $(0,\alpha]$ can be justified from (2.3), since outside this interval η has no effect on R_1. Although our method works for any $\eta \le \alpha$, it works the best when $\eta = \alpha$ with $R_1 = \max\left\{k : \frac{1}{k}\sum_{j=1}^{k} fdr_{(j)|1} \le [\alpha - fdr_1]/[1-fdr_1]\right\}$. This method with $\eta = \alpha$ is slightly different from the SC (Sun & Cai (2007)) method for testing a single group of hypotheses. It actually modifies the SC method by incorporating into the method the strength of significance of the group measured using its local fdr, and thus lets the SC method

to adapt itself according to the group's own significance. The TLTA, of course, reduces to the SC method (irrespective of η) when $m_g = 1$.

Going back to testing multiple groups of hypotheses, although the TLTA allows η to be chosen differently for the different groups, each within the interval $(0, \alpha]$, we consider keeping the η's same. Our reason is that it allows us to use a certain portion of the overall FDR level to maintain control over within-group false discoveries. This may be desirable in some applications where one would like to attach some measure of reliability to decisions made within each group. Of course, the choice of η is subjective and can be made judiciously based on ones prior knowledge or expertise.

Nevertheless, we will be choosing $\eta = \alpha$ in the following sections on simulation studies and real-data application.

3. Model Under Group Structure

3.1. The Model

We consider the following model for $(X_{gj}, \theta_g, \theta_{j|g})$, $g = 1, \ldots, G$; $j = 1, \ldots, m_g$:

$$\theta_1, \ldots, \theta_G \overset{i.i.d.}{\sim} Bernoulli(\pi_1),$$

$$\theta_{1|g}, \ldots, \theta_{m_g|g} | \theta_g = 0 \overset{i.i.d}{\sim} Bernoulli\ (0)\ (i.e.,\ P(\theta_{j|g} = 0 | \theta_g = 0) = 1),$$

$$\theta_{1|g}, \ldots, \theta_{m_g|g} | \theta_g = 1 \sim \frac{\prod_{j=1}^{m_g} \left\{ (1-\pi_{2|1})^{1-\theta_{j|g}} \pi_{2|1}^{\theta_{j|g}} \right\} I(\sum_j \theta_{j|g} > 0)}{1 - (1-\pi_{2|1})^{m_g}}\ [Truncated\ Bernoulli\ (\pi_{2|1})]$$

$$X_{gj} \mid \theta_{gj} \overset{ind}{\sim} (1-\theta_{gj}) f_0(x_{gj}) + \theta_{gj} f_1(x_{gj}),\ for\ some\ given\ densities\ f_0\ and\ f_1\ . \tag{3.1}$$

We call such a model with Truncated **B**ernoulli hidden states within each **S**ignificant **G**roup as **BSG** model.

3.2. Formulas for $fdr_{j|g}$ and fdr_g

Under the BSG model, we can obtain explicit formulas for $fdr_{j|g}$ and fdr_g when all the parameters are known, before using them to calculate the corresponding scores based on the data $\boldsymbol{x} = (x_{11}, \ldots, x_{Gm_G})$.

Let $\boldsymbol{x}_g = (x_{g1}, \ldots, x_{gm_g})$ represent the data vector for the g-th group. Let us introduce the following notations

$$\widetilde{fdr}_{gj} = \frac{(1-\pi_{2|1}) f_0(x_{gj})}{f(x_{gj})} \quad \text{and} \quad \widetilde{fdr}_g = \prod_{k=1}^{m_g} \widetilde{fdr}_{gk},$$

with $f(x) = (1-\pi_{2|1}) f_0(x) + \pi_{2|1} f_1(x)$, to define the local fdr scores for the hypothesis corresponding to x_{gj} within the gth group and for the gth group itself, respectively, under the assumption $\theta_{gj} \overset{i.i.d.}{\sim} Bernoulli(\pi_{2|1})$, which would be appropriate to make if we were to ignore whether or not the group is significant by letting $\sum_j \theta_{j|g} \geq 0$.

Then, as shown in Appendix 7.1, when the significance of a group is taken into account by making the Truncated Bernoulli ($\pi_{2|1}$) assumption for the θ_{gj}'s given that the gth group containing it is truly significant (i.e., $\theta_g = 1$), the local fdr score $\text{fdr}_{j|g}$ within that group is given by

$$fdr_{j|g} = P(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{x}) = \frac{\widetilde{fdr}_{gj} - \widetilde{fdr}_g}{1 - \widetilde{fdr}_g}, \tag{3.2}$$

and the group level local fdr score fdr_g is given by

$$fdr_g = \frac{(1-\pi_1)\widetilde{fdr}_g}{(1-\pi_1)\widetilde{fdr}_g + \pi_1 \cdot \frac{(1-\pi_{2|1})^{m_g}}{1-(1-\pi_{2|1})^{m_g}}(1 - \widetilde{fdr}_g)}. \tag{3.3}$$

The fact that these scores should reduce to zero when the gth group is not significant ($\theta_g = 0$) is immediate from these formulas.

Remark 2. It is interesting to note the following identities:

$$1 - \widetilde{fdr}_{gj} = (1 - \widetilde{fdr}_g)(1 - fdr_{j|g}),$$

and

$$\frac{1-(1-\pi_{2|1})^{m_g}}{(1-\pi_{2|1})^{m_g}} \cdot \frac{\widetilde{fdr}_g}{1 - \widetilde{fdr}_g} = \frac{\pi_1}{1-\pi_1} \cdot \frac{fdr_g}{1 - fdr_g},$$

which explain more directly how within- and between-group local fdr scores obtained by ignoring whether or not a group itself is significant relate to those when the group significance is taken into account using the BSG model.

3.3. Estimation

Here, we present the main steps of estimating the model parameters. These will be used later to derive data-driven versions of the TLTA and its competitors.

We assume that $X_{gj}|\theta_{gj} \sim (1-\theta_{gj})f_0(x_{gj}) + \theta_{gj}f_1(x_{gj})$, with $f_0(x) \equiv \phi(x)$, the density of $N(0,1)$, and

$$f_1(x) = \sum_{l=1}^{L} c_l \frac{1}{\sigma_l} \phi\left(\frac{x - \mu_l}{\sigma_l}\right),$$

after making appropriate transformations to the data. The following algorithm provides the steps to estimate the set of parameters $\boldsymbol{\beta} = (\pi_1, \pi_{2|1}, c_l, \mu_l, \sigma_l)$. It is based on EM algorithm (Dempster et al. (1977); Bilmes (1998)), and its detailed derivation is given in Appendix 7.2.

Definition 1 (Estimation Algorithm for β).

1. *Given the current value $\boldsymbol{\beta}'$, calculate $fdr_g(\boldsymbol{\beta}') = P(\theta_g = 0|\boldsymbol{x}, \boldsymbol{\beta} = \boldsymbol{\beta}')$, $fdr_{j|g}(\boldsymbol{\beta}') = P(\theta_{j|g} = 0|\boldsymbol{x}, \theta_g = 1, \boldsymbol{\beta} = \boldsymbol{\beta}')$ using (3.2) and (3.3), respectively, and*

$$\begin{aligned} & P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\ = \quad & \left(1 - fdr_g(\boldsymbol{\beta}')\right)\left(1 - fdr_{j|g}(\boldsymbol{\beta}')\right) \cdot \frac{c_l \frac{1}{\sigma_l}\phi(\frac{x_{gj} - \mu_l}{\sigma_l})}{\sum_{l=1}^{L} c_l \frac{1}{\sigma_l}\phi(\frac{x_{gj} - \mu_l}{\sigma_l})}, \forall l = 1, 2, \cdots, L; \end{aligned}$$

2. *Update* π_1 *and* $\pi_{2|1}$ *as*

$$\pi_1^{new} = 1 - \frac{\sum_g fdr_g(\boldsymbol{\beta}')}{G},$$

and

$$\pi_{2|1}^{new} = \frac{\sum_g \sum_{j=1}^{m_g} \left(1 - fdr_{j|g}(\boldsymbol{\beta}')\right)}{\sum_g \sum_{j=1}^{m_g} \left(1 - fdr_g(\boldsymbol{\beta}')\right)}.$$

3. *Update* (c_l, μ_l, σ_l) *as*

$$c_l^{new} = \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_{j|g} = 1, \theta_g = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} (1 - fdr_g(\boldsymbol{\beta}'))(1 - fdr_{j|g}(\boldsymbol{\beta}'))},$$

$$\mu_l^{new} = \frac{\sum_g \sum_{j=1}^{m_g} x_{gj} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')},$$

and

$$\sigma_l^{2new} = \frac{\sum_g \sum_{j=1}^{m_g} (x_{gj} - \mu_l)^2 P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}.$$

4. *Repeat steps (1-3) until convergence.*

4. Numerical Studies

We conducted numerical studies to examine how well the TLTA performs for multiple testing in comparison with its relevant competitors in their oracle and data-driven forms under the BSG model. In this section, we present the competing methods and the results of these studies.

4.1. Competing Methods

The most relevant, alternative approaches against which ours should be compared are the ones that ignore the group structure and perform multiple testing by pulling the hypotheses in a single group. With that in mind, we consider the following two methods and present them in their oracle forms under the above model setting.

The SC Method. (Sun & Cai (2007)) Let

$$PLfdr_{gj}(\boldsymbol{x}) \quad = \quad Pr(\theta_{gj} = 0|\boldsymbol{x}) = \frac{P(\theta_{gj} = 0) f_0(x_{gj})}{f(x_{gj})},$$

the pulled local FDR score for each hypothesis. Let $PLfdr_{(1)} \leq \cdots \leq PLfdr_{(N)}$ $(N = \sum_g m_g)$ be the ordered $PLfdr$ values, with $H_{(1)}, \ldots, H_{(N)}$ being the corresponding hypotheses. Find

$$k = max \left\{ i : \frac{1}{i} \sum_{i=1}^{k} PLfdr_{(i)} \leq \alpha \right\},$$

and reject $H_{(i)}$ for all $i = 1, \ldots, k$.

Under the BSG model, we have

$$PLfdr_{gj}(\boldsymbol{x}) = \frac{(1-\pi_1\pi_{2|1})f_0(x_{gj})}{(1-\pi_1\pi_{2|1})f_0(x_{gj})+\pi_1\pi_{2|1}f_1(x_{gj})}.$$

The Adaptive BH Method. (Benjamini & Hochberg (2000)). Let each X_{gj} be transformed to its p-value P_{gj}. Let $P_{(1)} \leq \cdots \leq P_{(N)}$ be the ordered versions of the these p-values when they are pulled into a single group. Compute

$$k = max\left\{i : (1-\pi_1\pi_{2|1})P_{(i)} \leq \frac{i\alpha}{N}\right\}.$$

If such a k exists, then reject the hypotheses associated with $P_{(1)},\ldots,P_{(k)}$; otherwise, do not reject any hypotheses.

Remark 3. It should be noted that Cai & Sun (2009) introduced a multiple-group version of the SC method and Hu et al. (2010) developed an adaptive BH method for grouped hypotheses. However, both papers rely on different model assumptions. In particular, in these two approaches, π_1 has been assumed to be one, implying that there is no sparsity on between-group level which is not appropriate when there are many groups. Secondly, they assumed the independence assumption within each group.

4.2. Oracle Comparison

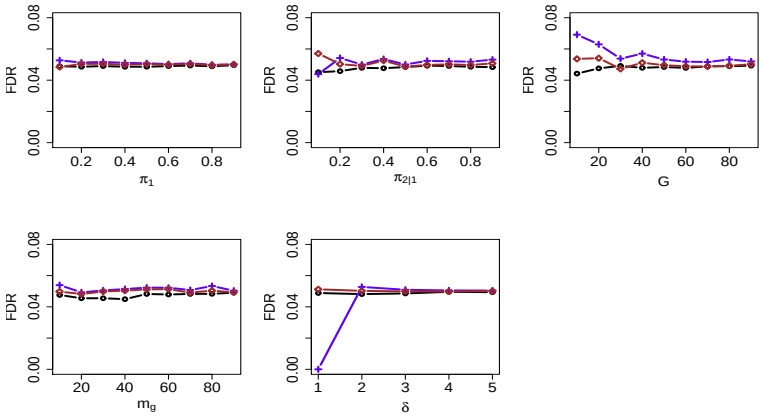
We now present the results of numerical studies conducted under the BSG model to examine the performance of the TLTA relative to its aforementioned competitors in their oracle forms in terms of FDR control and power. Two different definitions of power are used; one is the FNR (the expected proportion of false acceptances among all the accepted hypotheses), and the other is the Average Power (the expected proportion of truly rejected hypothesis).

We set $f_0(x) \equiv \phi(x)$, $f_1(x) = \phi(x-\delta)$, and $\alpha = \eta = 0.05$. There are five unknown quantities, $\pi_1, \pi_{2|1}$, δ, G, and m_g. The simulated values of FDR, FNR and Average Power were calculated based on 500 runs for each of these three methods having chosen some values for these quantities.

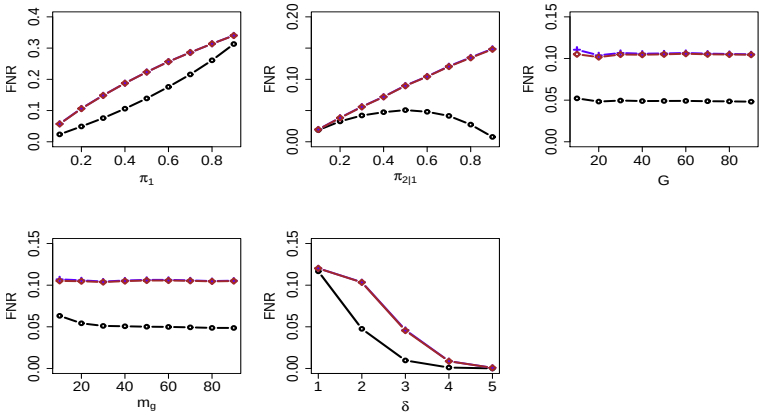
Figure 1 compares the three methods in terms of simulated FDR, FNR and Average Power. The basic values chosen for the unknown quantities are $\pi_1 = 0.2, \pi_{2|1} = 0.6, G = 100, m_g = 100$, and $\delta = 2$. In each figure, we allow the value of one of these quantities to vary, holding the other quantities at the aforementioned values. As seen from the graphs, the TLTA can perform significantly better than the other two methods in all cases.

4.3. Comparison of data-driven methods

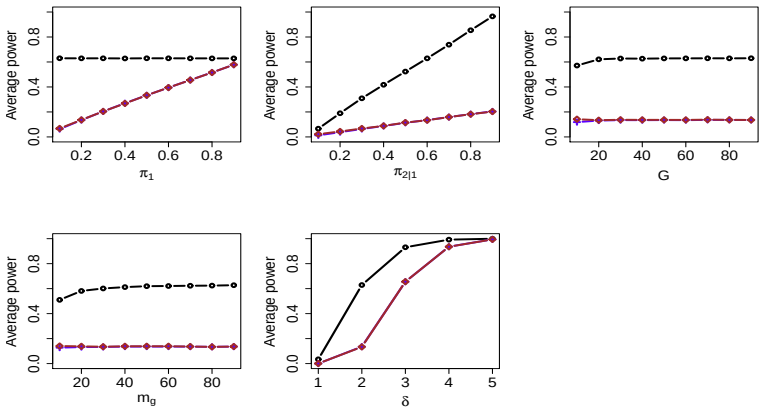
In the oracle versions of the different methods, we assume that all the parameters are known. In this section, they are estimated using the Estimation Algorithm in Section 3. We simulated the FDR, FNR and Average Power, for each of the three methods with different values of the unknown quantities in $\boldsymbol{\beta}$. The components of $\boldsymbol{\beta}$ and the basic values chosen for them are listed in Table 1. The value of $\eta = \alpha$ was set at 0.05. The simulated values were based on 500 runs under each setting for the set of unknown quantities.



(a) Comparison of FDR for Different Procedures



(b) Comparison of FNR for Different Procedures



(c) Comparison of Average Power for Different Procedures

Figure 1: Simulation results for the TLTA(—○—), the SC(—+—), and the Adaptive BH(—◇—) procedures in their oracle forms under the BSG model.

BSG model	
L	Basic values for $\boldsymbol{\beta}$
$L=1$	$\pi_1=0.2$, $\pi_{2\|1}=0.6$, $G=100$, $m_g=100$, $\delta=2$, $\sigma=1$
$L=2$	$c_1=c_2=0.5$, $\delta_2=-2$, $\sigma_2=1$ (in addition to the quantities for $L=1$)

Table 1: The Basic Values for the Unknown Quantities in $\boldsymbol{\beta}$ in the Simulation Studies

The simulation results are displayed in Figures 2 (L=1) and 3 (L=2). In each graph, we allow the value of one of the above quantities to vary while holding the others at the aforementioned values. As seen from these figures: (i) The performance of the data-driven version is very close to that of its oracle version, (ii) the overall FDR of all the procedures are controlled at the desired level α in all the scenarios considered, and (iii) the average power of the TLTA is the highest in all cases (as seen from Figures 2c and 3c).

Our simulations for the data-driven methods were done by starting with $G=3$ and then by increasing G from 10 to 100 in increments of 10. The simulation results showed that the TLTA would have much better performance when there are more groups, for instance, more than 10, although it gets plateaued with G more than 20.

In summary, as demonstrated through our numerical studies, the TLTA seems to outperform its competitors by effectively capturing the group structure. Therefore, we would like to recommend it for application to multiple testing of grouped hypotheses.

5. Real Data Application(AYP Study of California 2013)

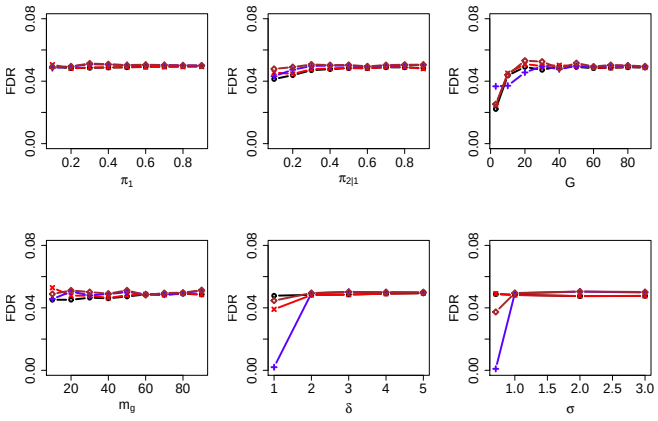
We take up the adequate yearly progress (AYP) study of California elementary schools in 2013 (http://www.cde.ca.gov/ta/ac/ay/aypdatafiles.asp) comparing the academic performance for socioeconomically advantaged (SEA) against socioeconomically disadvantaged (SED) students in the elementary schools. We compare the success rates in Math exams of SEA versus SED students. Although it is generally the case that the average success rate of SEA students is higher than SED students, our focus is in discovering the schools with unusually small or large advantaged-disadvantaged performance differences, and also to identify the school districts with such schools.

Let p_{1i} and p_{2i} be the success rates and n_{1i} and n_{2i} be the numbers of students in the groups of SEA and SED students, respectively, in the ith school, $i=1,\ldots,N$. Similar to Cai & Sun (2009) and Efron (2008), a z-value for school i is computed according to

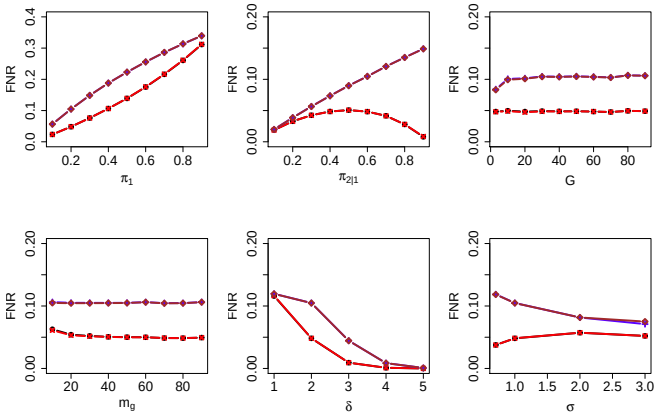
$$z_i=\frac{p_{1i}-p_{2i}-\tau}{\sqrt{p_{1i}(1-p_{1i})/n_{1i}+p_{2i}(1-p_{2i})/n_{2i}}},$$

where τ is the overall difference, median (p_{1i})- median (p_{2i}), which is 18.4% in this AYP study. There are 4118 $(=N)$ elementary schools and 701 qualified school districts (defined as having at least 20 students in each category and $|z|<10$ for each school). We consider these school districts as the groups in our application.

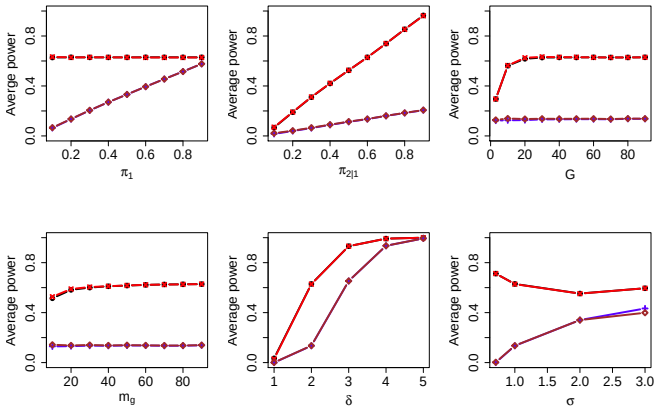
We apply the data-driven versions of the TLTA, the SC and the adaptive BH methods, assuming that $f_0(x)=\phi(x)$ and $f_1(x)$ is a mixture of two normal distributions each with variance



(a) Comparison of FDR for Different Procedures

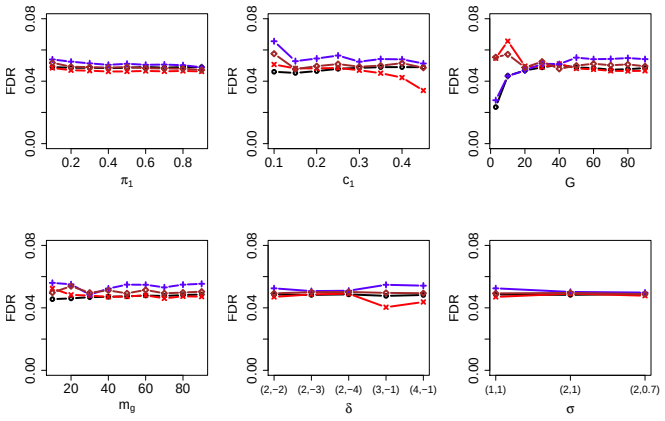


(b) Comparison of FNR for Different Procedures

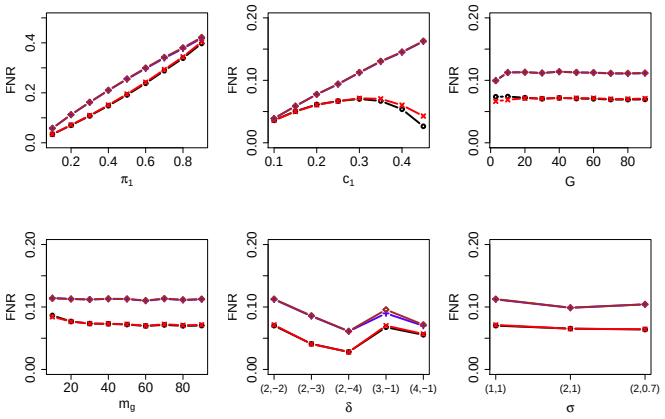


(c) Comparison of Average Power for Different Procedures

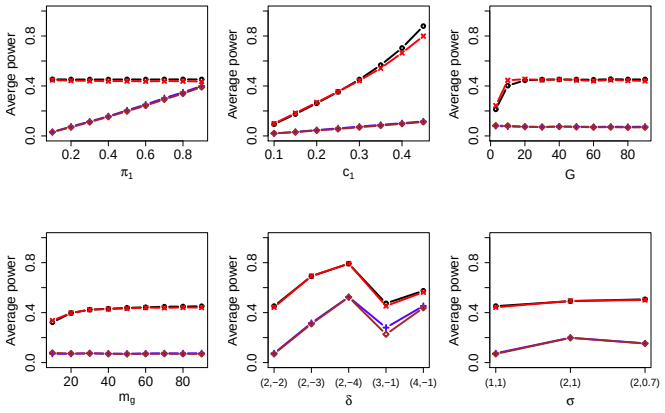
Figure 2: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 1$ under the BSG model.



(a) Comparison of FDR for Different Procedures



(b) Comparison of FNR for Different Procedures



(c) Comparison of Average Power for Different Procedures

Figure 3: Simulation results for our data-driven(—×—), the SC(—+—), and the Adaptive BH(—◇—) procedures with parameters estimated by EM algorithm and our oracle procedure(—○—) with $L = 2$ under the BSG model.

α=0.05 and η=0.05			
Procedures	School Discoveries	Group Decisions	School District Discoveries
TLTA	736	Yes	224
SC	471	No	
Adaptive BH	410	No	
α=0.1 and η=0.1			
TLTA	1085	Yes	284
SC	668	No	
Adaptive BH	629	No	

Table 2: Number of Discoveries Made by the Three Procedures

1. Using the Estimation Algorithm in Section 3, the estimated proportion of group significance $\widehat{\pi_1}$ is seen to be about 0.53, the estimated proportion of within-group significance $\widehat{\pi_{2|1}}$ is about 0.59, and the estimated f_1 is $\widehat{f_1(x)} \sim 0.21N(2.64,1) + 0.79N(-1.88,1)$. We chose two different values, 0.05 and 0.10, for $\eta = \alpha$.

The numbers of discoveries made by the three methods are shown in Table 2. As seen from this table, the TLTA can identify more unusual schools having extremely small or large academic performance difference between SEA and SED students than the other methods. Our discoveries of schools within each district seem statistically more informative than those made by the other methods that, unlike ours, don't attempt to control within-group false discoveries, and so could potentially be of value to district level education policy makers.

6. Concluding Remarks

When testing grouped hypotheses, how overall false discoveries across all hypotheses is intertwined with false discoveries of hypotheses within each group seems fundamental to deeper understanding towards effectively capturing the underlying group structure. This is an important issue that has not been answered in the literature, as far as we know. This article presents a theoretical framework built on this fundamental understanding from a Bayesian viewpoint, and develops a new approach to multiple testing of grouped hypotheses that allows one to maintain some specific control over within-group false discoveries while controlling the overall false discoveries across all hypotheses. This new approach is a two-fold loop algorithm integrating within and between group discoveries.

Having a separate control over within-group false discoveries, we argue, is often an effective way of capturing the underlying group structure when testing grouped hypotheses, particularly when there is high positive dependencies within groups. Moreover, this is often desired in some applications, such as in analyzing the AYP data in Section 4 where discovering schools within a school district controlling a district specific false discovery rate seems practically more useful than discovering these schools through a global discovery process controlling a global false discovery rate. It allows making statistically more reliable district level decisions for policy makers. Of course, the choice of the level η at which within-group false discoveries is to be controlled is

subjective and can be made judiciously based on ones prior knowledge in terms of how stringent that control should be.

There is lot more that can be done following our current research; for instance, (i) extending the TLTA to other types of within-group dependency, such as hidden Markov (Newton et al. (2004)) or time series dependency(Tang & Zhang (2007)), (ii) investigating its optimality, and (iii) developing its frequentist analog providing a multiple-group and improved version of the single-group Benjamini & Hochberg (1995) method.

An R-package, called "GroupTest", which is developed to carry out the numerical calculations associated with the TLTA in this paper is made available at `http://astro.temple.edu/~zhaozhg/software.html`. All simulations involving EM algorithm were run through the high performance computing cluster at Temple University supported by NSF instrumentation grant CNS-09-58854.

7. Appendix

7.1. Proofs of (3.2) and (3.3)

Let $f_{\theta_{k|g}}(x) = (1-\theta_{k|g})f_0(x) + \theta_{k|g}f_1(x)$, $\tilde{\theta}_g = (\theta_{1|g},\ldots,\theta_{m_g|g})$, and $\Omega = \{0,1\}^m \backslash (0,\ldots,0)$. Then, Eqn. (3.2) follows from that fact that

$$fdr_{j|g} = \frac{f(\boldsymbol{x}_g|\theta_{j|g}=0,\theta_g=1)P(\theta_{j|g}=0|\theta_g=1)}{f(\boldsymbol{x}_g|\theta_g=1)},$$

where

$$\begin{aligned}
& f(\boldsymbol{x}_g|\theta_{j|g}=0,\theta_g=1)P(\theta_{j|g}=0|\theta_g=1) \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}} \sum_{\theta_{j|g}=0,\tilde{\theta}_g\in\Omega} \left[\prod_{k=1}^{m_g}\left\{f_{\theta_{k|g}}(x_{gk})\right\}\prod_{k=1}^{m_g}\left\{(1-\pi_{2|1})^{1-\theta_{k|g}}\pi_{2|1}^{\theta_{k|g}}\right\}\right] \\
= & \frac{(1-\pi_{2|1})f_0(x_{gj})}{1-(1-\pi_{2|1})^{m_g}} \sum_{\theta_{j|g}=0,\tilde{\theta}_g\in\Omega} \left[\prod_{k=1,k\neq j}^{m_g}\left\{f_{\theta_{k|g}}(x_{gk})\right\}\prod_{k=1,k\neq j}^{m_g}\left\{(1-\pi_{2|1})^{1-\theta_{k|g}}\pi_{2|1}^{\theta_{k|g}}\right\}\right] \\
= & \frac{(1-\pi_{2|1})f_0(x_{gj})}{1-(1-\pi_{2|1})^{m_g}} \left[\prod_{k=1,k\neq j}^{m_g} f(x_{gk}) - (1-\pi_{2|1})^{m_g-1}\prod_{k=1,k\neq j}^{m_g} f_0(x_{gk})\right] \\
= & \frac{\widetilde{fdr}_{gj}}{1-(1-\pi_{2|1})^{m_g}} \left[1-\prod_{k=1,k\neq j}^{m_g}\widetilde{fdr}_{gk}\right]\prod_{k=1}^{m_g} f(x_{gk}) \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}} \left[\widetilde{fdr}_{gj} - \widetilde{fdr}_{g}\right]\prod_{k=1}^{m_g} f(x_{gk}), \qquad (7.1)
\end{aligned}$$

and

$$
\begin{aligned}
& f(\boldsymbol{x}|\theta_g=1) \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}} \sum_{\tilde{\theta}_g \in \Omega} \left[\prod_{k=1}^{m_g} \left\{ f_{\theta_{k|g}}(x_{gk}) \right\} \prod_{k=1}^{m_g} \left\{ (1-\pi_{2|1})^{1-\theta_{k|g}} \pi_{2|1}^{\theta_{k|g}} \right\} \right] \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}} \sum_{\tilde{\theta}_g \in \Omega} \left[\prod_{k=1}^{m_g} \left\{ f_{\theta_{k|g}}(x_{gk}) \right\} \prod_{k=1}^{m_g} \left\{ (1-\pi_{2|1})^{1-\theta_{k|g}} \pi_{2|1}^{\theta_{k|g}} \right\} \right] \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}} \left[\prod_{k=1}^{m_g} f(x_{gk}) - (1-\pi_{2|1})^{m_g-1} \prod_{k=1}^{m_g} f_0(x_{gk}) \right] \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}\}} \left[1 - \prod_{k=1}^{m_g} \widetilde{fdr}_{gk} \right] \prod_{k=1}^{m_g} f(x_{gk}) \\
= & \frac{1}{1-(1-\pi_{2|1})^{m_g}\}} \left[1 - \widetilde{fdr}_g \right] \prod_{k=1}^{m_g} f(x_{gk}). \qquad (7.2)
\end{aligned}
$$

Eqn. (3.3) follows from the fact that

$$fdr_g = \frac{(1-\pi_1) f(\boldsymbol{x}_g|\theta_g=0)}{(1-\pi_1) f(\boldsymbol{x}_g|\theta_g=0) + \pi_1 f(\boldsymbol{x}_g|\theta_g=1)},$$

where

$$f(\boldsymbol{x}_g|\theta_g=0) = \prod_{k=1}^{m_g} f_0(x_{gk}) = \frac{\widetilde{fdr}_g}{(1-\pi_{2|1})^{m_g}} \prod_{k=1}^{m_g} f(x_{gk}),$$

and $f(\boldsymbol{x}_g|\theta_g=1)$ equals what is given in (7.2).

7.2. Detailed Derivation of Estimation Algorithm for $\boldsymbol{\beta}$

For better presentation of the results, let us define $\pi_1^1 = \pi_1, \pi_1^0 = 1-\pi_1, \pi_{2|1}^1 = \pi_{2|1}$ and $\pi_{2|1}^0 = 1-\pi_{2|1}$. Consider $(\boldsymbol{x}, \boldsymbol{\theta})$ as the complete data. Then the complete log-likelihood function can be written as:

$$
\begin{aligned}
& l(\boldsymbol{x}, \boldsymbol{\theta}) \\
& = \sum_g \sum_{k=0}^{1} I(\theta_g = k)(log\pi_1^k + logf(x_{gj}|\theta_g = k)) \\
& = \sum_g \left\{ I(\theta_g=0) \left[log\pi_1^0 + \sum_{j=1}^{m_g} logf(x_{gj}|\theta_g=0) \right] + I(\theta_g=1) \left[log\pi_1^1 + logf(\boldsymbol{x}_g|\theta_g=1) \right] \right\} \\
& = \sum_g \sum_{k=0}^{1} I(\theta_g=k) log\pi_1^k + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} \left[I(\theta_g=1, \theta_{j|g}=1, m_{j|g}=l) log(\pi_{2|1} c_l) + I(\theta_g=1, \theta_{j|g}=0) log(\pi_{2|1}^0) \right] \\
& \quad + \sum_g \left[I(\theta_g=0) \sum_{j=1}^{m_g} logf_0(x_{gj}) + I(\theta_g=1) \sum_{j=1}^{m_g} I(\theta_{j|g}=0) logf_0(x_{gj}) \right] \\
& \quad + \sum_g I(\theta_g=1) \sum_{j=1}^{m_g} \sum_{l=1}^{L} I(\theta_{j|g}=1, m_{j|g}=l) logf_l(x_{gj}|\theta_{j|g}=1),
\end{aligned}
$$

where $m_{j|g} = l$ implies that $x_{j|g}$ is generated from $N(\mu_l, \sigma_l^2)$.

The expected value of the complete-data log-likelihood $l(\boldsymbol{x}, \boldsymbol{\theta})$ with respect to the unknown θ_g and $\theta_{j|g}$, given the observed data $\boldsymbol{x}$ and the current value $\boldsymbol{\beta}'$ of the parameter, is:

$$
\begin{aligned}
&Q(\boldsymbol{\beta}, \boldsymbol{\beta}') = E\left[l(\boldsymbol{x}, \boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{\beta}'\right] \\
&= \sum_g \sum_{k=0}^{1} log\pi_1^k P(\theta_g = k|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{k=0}^{1} \log \pi_{2|1}^k P(\theta_j = 1, \theta_{j|g} = k|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log c_l P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_{j=1}^{m_g} log f_0(x_{gj}) fdr_g(\boldsymbol{\beta}') + \sum_g \sum_{j=1}^{m_g} log f_0(x_{gj})(1 - fdr_g(\boldsymbol{\beta}')) \\
&\quad + \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log f_l(x_{gj}) P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}').
\end{aligned}
$$

Note that

$$
\begin{aligned}
& P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}') \\
= \quad & P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x}, \beta') \frac{c_l \frac{1}{\sigma_l} \phi(\frac{x_{gj} - \mu_l}{\sigma_l})}{\sum_{l=1}^{L} c_l \frac{1}{\sigma_l} \phi(\frac{x_{gj} - \mu_l}{\sigma_l})},
\end{aligned}
$$

where

$$
P(\theta_g = 1, \theta_{j|g} = 1|\boldsymbol{x}, \beta') = (1 - fdr_{j|g}(\boldsymbol{\beta}'))(1 - fdr_g(\boldsymbol{\beta}')).
$$

We want to maximize the Q function, which can be realized by maximizing each of these parts to get the estimates of $(\pi_1, \pi_{2|1}, c_l)$ and (μ_l, σ_l^2), since these parts are not related. To maximize the first part with the restriction that $\pi_1^0 + \pi_1^1 = 1$, using the Lagrange multipliers, we can find the maximizers for π_1^1 as

$$
\pi_1^{new} = 1 - \frac{\sum_g fdr_g(\boldsymbol{\beta}')}{G},
$$

Similarly, we can find the maximizer for $\pi_{2|1}$ and c_l as

$$
\begin{aligned}
\pi_{2|1}^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} (1 - fdr_g(\boldsymbol{\beta}'))(1 - fdr_{j|g}(\boldsymbol{\beta}'))}{\sum_g \sum_{j=1}^{m_g} (1 - fdr_g(\boldsymbol{\beta}'))} \\
c_l^{new} &= \frac{\sum_g \sum_{j=1}^{m_g} P(\theta_{j|g} = 1, \theta_g = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} (1 - fdr_g(\boldsymbol{\beta}'))(1 - fdr_{j|g}(\boldsymbol{\beta}'))}.
\end{aligned}
$$

For the last part of Q function, we know that $f_0(x) \sim N(0,1)$ and $f_l(x) \sim N(\mu_l, \sigma_l^2)$ with probability c_l . Therefore, for each l, we need to find the MLEs for μ_l and σ_l^2 by maximizing

the following log-likelihood function:

$$\sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} log f_l(x_{gj}) P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')$$
$$= \sum_g \sum_{j=1}^{m_g} \sum_{l=1}^{L} \left[-\frac{1}{2} log\sigma_l^2 - \frac{1}{2\sigma_l^2}(x_{gj} - \mu_l)^2 \right] P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}').$$

Taking derivatives with respect to μ_l and σ_1^2 and equating them to zero, we can get:

$$\mu_l^{new} = \frac{\sum_g \sum_{j=1}^{m_g} x_{gj} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}$$
$$\sigma_l^{2new} = \frac{\sum_g \sum_{j=1}^{m_g} (x_{gj} - \mu_l)^2 P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}{\sum_g \sum_{j=1}^{m_g} P(\theta_g = 1, \theta_{j|g} = 1, m_{j|g} = l|\boldsymbol{x}, \boldsymbol{\beta}')}.$$

Acknowledgements

The authors thank a reviewer for valuable suggestions that led to improved presentation of the paper.

References

Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R., & White, K. (2002). Gene expression during the life cycle of Drosophila melanogaster. *Science*, *297*, 2270–2275.

Benjamini, Y., & Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, *102*, 1272–1281.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57(1)*, 289–300.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60–83.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, *4*, 126.

Cai, T., & Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, *104*, 1467–1481.

Calvano, S., Xiao, W., Richards, D., Felciano, R., Baker, H., Cho, R., Chen, R., Brownstein, B., Cobb, J., Tschoeke, S., Miller-Graziano, C., Moldawer, L., Mindrinos, M., Davis, R., Tompkins, R., & Lowry, S. (2005). A network-based analysis of systemic inflammation in humans. *Nature*, *437*, 1032–1037.

Clements, N., Sarkar, S. K., & Guo, W. (2011). Astronomical transient detection using grouped p-values and controlling the false discovery rate. *Statistical Challengers in Modern Astronomy*, *V*, 383–396.

Clements, N., Sarkar, S. K., Zhao, Z., & Kim, D. (2014). Applying multiple testing procedures to detect changes in East African vegetation. *Annals of Applied Statistics*, *8*, 286–308.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, *39(1)*, 1–38.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statisitcal Science*, *23*, 1–22.

Efron, B., Tibshirani, B., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*, 1151–1160.

Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, *23*, 70–86.

He, L., Sarkar, S. K., & Zhao, Z. (2015). Capturing the severity of type II errors in high-dimensional multiple testing. *Journal of Multivariate Analysis*, *142*, 106–116.

Heller, R., Manduchi, E., Grant, G., & Ewens, W. (2009). A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics*, *25*, 1019–1205.

Hu, J., Zhao, H., & Zhou, H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, *105*, 1215–1227.

Newton, M., Noueiry, A., Sarkar, D., & Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, *5(2)*, 155–176.

Pacifico, M., Genovese, C., Verdinelli, I., & Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, *99*, 1002–1014.

Sarkar, S., Zhou, T., & Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica*, *18*, 925–946.

Schildknecht, K., Tabelow, K., & Dickhaus, T. (2016). More specific signal detection in functional magnetic resonance imaging by false discovery rate control for hierarchically structured systems of hypotheses. *PLoS One*, *11(2)*, e0149016.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 15545–15550.

Sun, W., & Cai, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, *102(479)*, 901–912.

Sun, W., & Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society. Series B*, *71(2)*, 393–424.

Tang, W., & Zhang, C. (2007). Empirical Bayes methods for controlling the false discovery rate with dependent data. *Lecture Notes–Monograph Series*, *54*, 151–160.