Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks

T. Tony Cai[1] and Wenguang Sun[2]

Abstract

In large-scale multiple testing problems, data are often collected from heterogeneous sources and hypotheses form into groups that exhibit different characteristics. Conventional approaches, including the pooled and separate analyses, fail to efficiently utilize the external grouping information. We develop a compound decision theoretic framework for testing grouped hypotheses and introduce an oracle procedure that minimizes the false non-discovery rate subject to a constraint on the false discovery rate. It is shown that both the pooled and separate analyses can be uniformly improved by the oracle procedure. We then propose a data-driven procedure that is shown to be asymptotically optimal. Simulation studies show that our procedures enjoy superior performance and yield the most accurate results in comparison with both the pooled and separate procedures. A real data example with grouped hypotheses is studied in detail using different methods. Both theoretical and numerical results demonstrate that exploiting external information of the sample can greatly improve the efficiency of a multiple testing procedure. The results also provide insights on how the grouping information is incorporated for optimal simultaneous inference.

Keywords: Compound decision problem; Conditional local false discovery rate; Exchangeability; False discovery rate; Grouped hypotheses; Large-scale multiple testing.

[1] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
The research of Tony Cai was supported in part by NSF Grant DMS-0604954.
[2] Department of Statistics, North Carolina State University, Raleigh, NC 27695

1 Introduction

Conventional multiple testing procedures, such as the false discovery rate analyses (Benjamini and Hochberg 1995; Efron et al. 2001; Storey 2002; Genovese and Wasserman 2002; van der laan et al. 2004), implicitly assume that data are collected from repeated or identical experimental conditions, and hence hypotheses are exchangeable. However, in many applications, data are known to be collected from heterogeneous sources and hypotheses intrinsically form into different groups. The goal of this article is to study optimal multiple testing procedures for grouped hypotheses in a compound decision theoretical framework.

The following examples motivate our study. The adequate yearly progress (AYP) study of California high schools (Rogosa 2003) aimed to compare academic performances of social-economically advantaged (SEA) versus social-economically disadvantaged (SED) students. Standard tests in mathematics were administered to 7867 schools and a z-value for comparing SEA and SED students was obtained for each school. The estimated null densities of the z-values for small, medium and large schools are plotted on the left panel of Figure 1. It is interesting to see that the null density of the large group is much wider than those of the other two groups. The differences in the null distributions have significant effects on the outcomes of multiple testing procedures. See more detailed analysis of this example in Section 6. Another example is the brain imaging study analyzed in Schwartzman et al. (2005). In this study, 6 dyslexic children and 6 normal children received diffusion tensor imaging brain scans on the same 15443 brain locations (voxels). A z-value (converted from a two-sample t-statistic) for comparing dyslexic versus normal children was obtained for each voxel. The right panel in Figure 1 plots the estimated null densities of the z-values for the front and back halves of the brain. We can see that the null cases from two groups center on different means, and the density of the back half is narrower. There are many other examples where the hypotheses are naturally grouped. For instance, in analysis of geographical survey data, individual locations are aggregated into several large clusters; and in meta-analysis of large biomedical studies, the data are collected from different clinical centers. An important common feature of these examples is that data are collected from heterogeneous sources and the hypotheses

being considered are grouped and no longer exchangeable. We shall see that incorporating the grouping information is important for optimal simultaneous inference with samples collected from different groups.

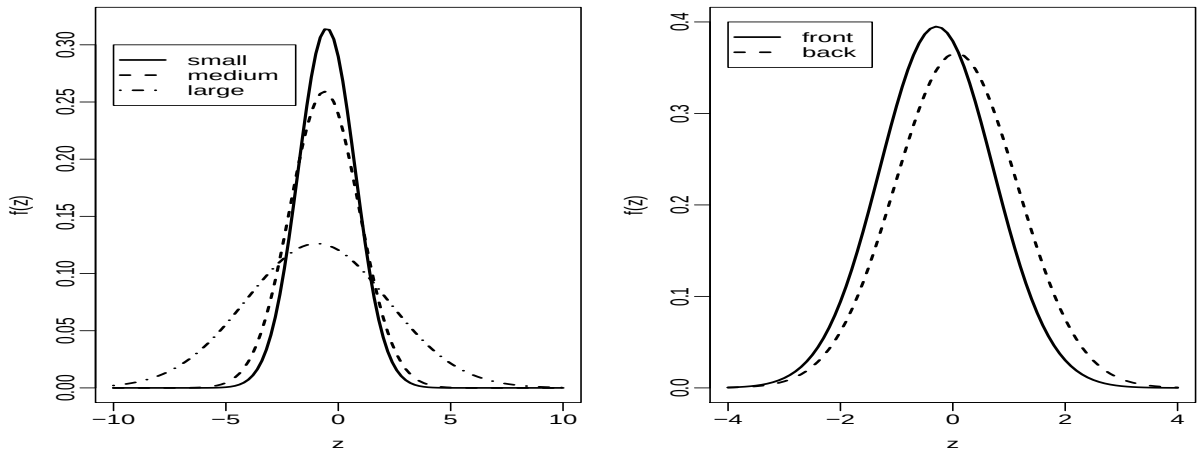

Figure 1: Empirical null densities of the AYP study and the brain imaging study. The null densities of the school data in the left panel are specified in Table 3 in Section 6. The null density of the large group is much wider than those of the other two densities. In the right panel, the null densities of the front and back halves of the brain are $N(0.06, 1.09^2)$ and $N(-0.29, 1.01^2)$, respectively, which are centered at different means.

The analysis of above examples involves simultaneous testing of thousands of hypotheses. In large-scale multiple testing, the false discovery rate (FDR, Benjamini and Hochberg 1995) is often used to combine the type I errors from individual tests and serves as a target for control. The outcomes of a multiple testing procedure can be summarized as in Table 1.

Table 1: Classification of tested hypothesis

	Claimed non-significant	Claimed significant	Total
Null	N_{00}	N_{10}	m_0
Non-null	N_{01}	N_{11}	$m - m_0$
Total	S	R	m

The FDR, defined as $\text{FDR} = E(N_{10}/R)P(R > 0)$, is the expected proportion of false positives among all rejections. The marginal FDR (mFDR), defined as $\text{mFDR} = E(N_{10})/E(R)$, is an asymptotically equivalent measure to the FDR in the sense that $\text{mFDR} = \text{FDR} + O(m^{-\frac{1}{2}})$

under mild conditions (Genovese and Wasserman 2002), where m is the total number of tests. A dual quantity of the FDR is the false non-discovery rate (FNR, Genovese and Wasserman 2002), which is defined as $\mathrm{FNR} = \mathrm{E}(N_{01}/S|S > 0)\mathrm{Pr}(S > 0)$, the expected proportion of false negatives among all non-rejections. An FDR procedure is said to be *valid* if it controls the FDR at a prespecified level α and *optimal* if it has the smallest FNR among all FDR procedures at level α.

The problem of combining the tests from several large groups is conceptually complicated in an FDR analysis. On the one hand, it is desirable in practice to define the FDR and FNR as global measures by pooling together all tests from different groups. On the other hand, it is beneficial to perform the analysis separately in some way when the groups are different. For example, the implementation of the adaptive p-value FDR procedure (Benjamini and Hochberg 2000; Genovese and Wasserman 2004) requires the information about the proportion of non-nulls, which may vary across groups.

Two natural strategies to testing grouped hypotheses have been considered in the literature. The first approach, termed as the *pooled analysis*, simply ignores the information of group labels and conducts a global analysis on the combined sample at a given FDR level α. It is argued by Efron (2008a) that a pooled FDR analysis may distort inferences made for separate groups because highly significant cases from one group may be hidden among the nulls from another group, while insignificant cases may be possibly enhanced. Another natural approach is the so-called *separate analysis* which first conducts separately the FDR analysis within each group at the same FDR level α, and then combines the testing results from individual analyses. It was shown by Efron (2008a) that the separate analysis is valid. However, the choice of identical FDR levels for all groups is somewhat arbitrary since there are many combinations of group-wise FDR levels α_i's that lead to an overall FDR level α. The choice of identical FDR levels $\alpha_i = \alpha$ for all groups is merely one of the combinations, and is not optimal in general.

This article studies the optimal procedure for testing grouped hypotheses in a compound decision theoretical framework and shows that both the pooled and separate analyses can be uniformly improved. We first introduce an *oracle procedure* in an ideal setting where the distributional information of each group is assumed to be known. It is shown that the oracle

procedure is optimal in a global sense, that is, it minimizes the overall FNR subject to a constraint on the overall FDR level. Our approach is different from conventional methods in that it is a hybrid strategy that has combined features from both pooled and separate analyses. The optimality of our new procedure is achieved by utilizing the information of group labels to create efficient rankings of all hypotheses, and adaptively weighting the FDR levels among different groups to minimize the overall FNR level.

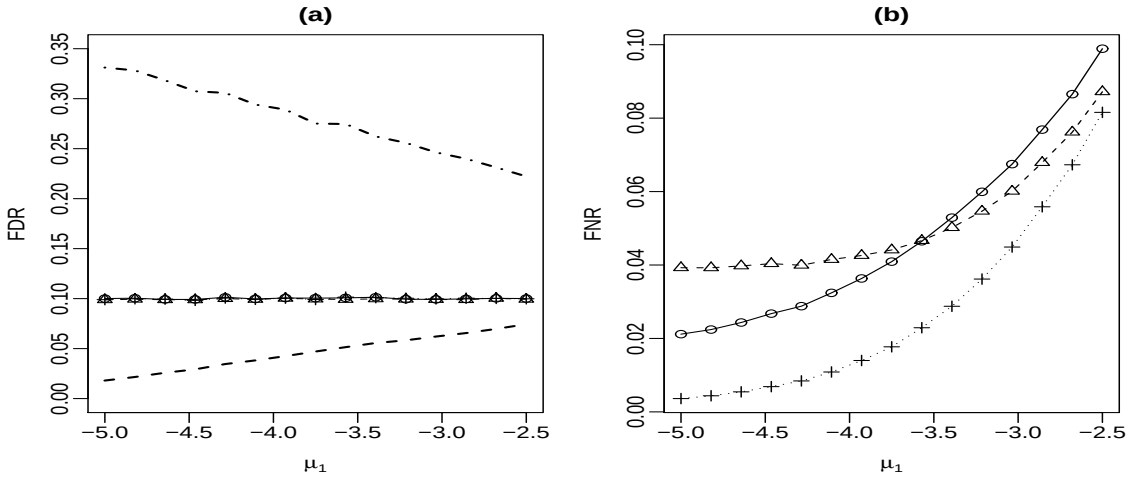

Figure 2: A comparison of the pooled, separate and optimal analyses in a two-group model: the group sizes are $m_1 = 3000$, $m_2 = 1500$; the group densities are $f_1 = 0.8N(0,1) + 0.2N(\mu_1, 1)$ and $f_2 = 0.9N(0,1) + 0.1N(2, 0.5^2)$. In panel (a), the FDR levels are plotted as functions of μ_1 ($\circ$, pooled analysis; $\triangle$, separate analysis; $+$, optimal analysis; dashed line, optimal FDR levels for Group 1; dash-dotted line, optimal FDR levels for Group 2). In Panel (b), the FNR levels are plotted as functions of μ_1 ($\circ$, pooled analysis; $\triangle$, separate analysis; $+$, optimal analysis).

Figure 2 gives a comparison of the pooled, separate and optimal testing procedures. The left panel shows that all three procedures controls the FDR at the nominal level 0.10 (the three lines are overlapped at 0.10). The right panel shows that neither the pooled nor the separate analysis is efficient, and both are uniformly dominated by the optimal procedure. The pooled analysis is inefficient because the information of group labels can be exploited to construct more efficient tests. The separate analysis with identical FDR levels is also inefficient because different group-wise FDR levels should be chosen to minimize the overall FNR level. The optimal group-wise FDR levels suggested by our new procedure are given by the dashed and dash-dotted lines in Panel (a).

We then develop a data-driven procedure that mimics the oracle procedure by plugging-in consistent estimates of the unknown parameters. It is shown that the data-driven procedure controls the overall FDR at the nominal level and attains the FNR level of the oracle procedure asymptotically. In this sense, it is *asymptotically valid and optimal*. Consistent estimates of the optimal FDR levels for separate groups are also provided based on the data-driven procedure. Simulations are conducted in Section 5, showing that our procedures enjoy superior performance and yield the most accurate results in comparison with both the pooled and separate procedures.

An important issue here is that the assumption that all hypotheses are exchangeable, which has been implicitly used in the multiple testing literature, often does not hold in practice where hypotheses are grouped. It was conjectured by Morris (2008) that when the exchangeability assumption does not hold, the resulting rankings of the hypotheses should be different; this conjecture is verified by the efficiency gain of our optimal testing procedure over the conventional methods. Generally speaking, testing procedures developed under the exchangeability assumption are *symmetric rules* (defined in Section 7); examples include the BH step-up procedure (Benjamini and Hochberg 1995), Efron's local FDR procedure (Efron et al. 2001) and Storey's optimal discovery procedure (Storey 2007). When the hypotheses are not exchangeable, even the optimal symmetric rules may suffer from substantial efficiency loss. Recent works by Efron (2008a) and Ferkinstad et al. (2008) suggest that hypotheses should be analyzed separately when they are not exchangeable. However, it is not discussed how to optimally combine the testing results from separate groups. Our new procedure not only gives the optimal rankings of all hypotheses, but also suggests an optimal way of combining testing results (where group-wise FDR levels are automatically and optimally determined and adaptively weighted among groups). Therefore it provides a convenient and efficient approach to testing grouped hypotheses.

The article is organized as follows. We begin in Section 2 with an introduction of the multiple-group model and the two natural approaches to testing grouped hypotheses. Section 3 first introduces, under an ideal setting, an oracle procedure which minimizes the overall FNR subject to a constraint on the overall FDR, and then proposes a data-driven procedure

that asymptotically mimics the oracle procedure. In Section 4, a compound decision-theoretic framework for testing grouped hypotheses is developed and the optimality of the new testing procedure is established. Simulation studies are carried out in Section 5 to investigate the numerical performance of our procedure. The methods are illustrated in Section 6 for analysis of the AYP study of California high schools. Section 7 discusses the main findings of the article as well as some open problems. The proofs are given in the Appendix.

2 Pooled and Separate FDR Analysis

The *random mixture model* provides a convenient and efficient framework for large-scale multiple testing and has been widely used in many applications, especially in DNA microarray analyses (Efron et al. 2001; Newton et al. 2001; Storey 2002). In a random mixture model, observations $x_1, \cdots, x_m$ are assumed to be generated from a mixture distribution:

$$X \sim (1-p)F_0(x) + pF_1(x), \tag{2.1}$$

where F_0 and F_1 are null and non-null distributions, and p is the proportion of non-nulls. The mixture density is denoted by $f(x) = (1-p)f_0(x) + pf_1(x)$.

We begin by reviewing the optimal and adaptive testing procedures developed in Sun and Cai (2007) under the mixture model (2.1). Then we introduce the multiple-group random mixture model that extends model (2.1) to describe grouped hypotheses. Finally we discuss two natural methods, pooled and separate FDR procedures, for testing grouped hypotheses.

2.1 Optimal testing procedures for a single group model

Conventional FDR procedures, such as the step-up procedure (Benjamini and Hochberg 1995), the adaptive p-value procedure (Benjamini and Hochberg 2000; Genovese and Wasserman 2002), and the augmentation procedure (van der laan et al. 2004), are virtually all based on thresholding the ranked p-values. However, the p-value ignores important distributional information in the sample and fails to serve as the fundamental building block in large-scale

multiple testing. Sun and Cai (2007) developed a compound decision-theoretic framework for multiple testing and showed that the optimal testing procedure is a thresholding rule based on the local false discovery rate (Lfdr, Efron et al. 2001). The Lfdr, defined as $\text{Lfdr}(x) = p_0 f_0(x)/f(x)$, is the posterior probability that a case is null given the observed statistic. Sun and Cai (2007) showed that the Lfdr produces more efficient rankings of hypotheses than the p-value, and the efficiency gain is substantial when the non-null distribution is concentrated or skewed. Let $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m) \in \{0,1\}^m$ be a general decision rule, where $\delta_i = 1$ if we claim that case i is non-null and $\delta_i = 0$ otherwise. Sun and Cai (2007) showed that when all distributional information is known, the *oracle testing procedure* at FDR level α that minimizes the FNR is

$$\boldsymbol{\delta}(\mathbf{Lfdr}, c_{OR}(\alpha)\mathbf{1}) = [I\{\text{Lfdr}(x_i) < c_{OR}(\alpha)\} : i = 1, \cdots, m], \tag{2.2}$$

where $c_{OR}(\alpha) = \sup\{c \in (0,1), \text{FDR}(c) \leq \alpha\}$ is the optimal cutoff for the Lfdr statistic at FDR level α. However, the oracle procedure (2.2) is difficult to implement because the cutoff $c_{OR}(\alpha)$ is hard to compute. Denote by $\text{Lfdr}_{(1)}, \cdots, \text{Lfdr}_{(m)}$ the ranked Lfdr values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. An asymptotically equivalent version of the oracle procedure (2.2) is the following procedure:

$$\text{Reject all } H_{(i)},\ i = 1, \cdots, l, \quad \text{where } l = \max\left\{ i : (1/i)\sum_{j=1}^{i} \text{Lfdr}_{(j)} \leq \alpha \right\}. \tag{2.3}$$

The Lfdr procedure (2.3) is asymptotically valid and optimal in the sense that it attains both the FDR and FNR levels of the oracle procedure (2.2) asymptotically.

Implementation of the Lfdr procedure requires the knowledge of population parameters such as the null density f_0 and proportion of non-nulls p, which may not be known in practice. Estimates of these unknown parameters for a normal mixture model have been developed in the literature, see Efron (2004), and Jin and Cai (2007). Let $\hat{p}$, $\hat{f}_0$ and $\hat{f}$ be estimates of the unknown parameters and define the estimated Lfdr as $\widehat{\text{Lfdr}}(x) = \hat{p}\hat{f}_0(x)/\hat{f}(x)$. An *adaptive* procedure was proposed in Sun and Cai (2007) which replaces the Lfdr statistics in (2.3) by their estimates. It was shown that the adaptive procedure is asymptotically valid and optimal

when consistent estimates (e.g., Jin and Cai (2007)'s estimates) are used to construct the tests. Numerical results show that conventional p-value procedures can be substantially improved by the adaptive procedure.

2.2 The multiple-group model

The multiple-group random mixture model (Efron 2008; see Figure 3) extends the previous random mixture model (2.1) (for a single group) to cover the situation where the m cases can be divided into K groups. It is assumed that within each group, the random mixture model (2.1) holds separately.

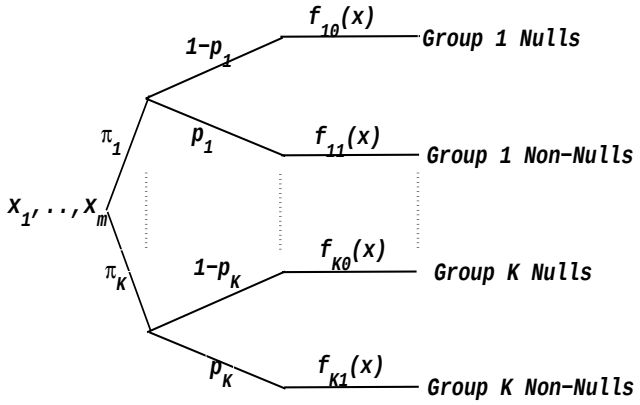

Figure 3: The multiple group model: the m hypotheses are divided into K groups with prior probability π_k; the random mixture model (2.1) holds separately within each group, with possibly different p_k, f_{k0} and f_{k1}.

Let $\boldsymbol{g} = (g_1, \cdots, g_K)$ be a multinomial variable with associated parameters $\{\pi_1, \cdots, \pi_K\}$, where $g_i = k$ indicates that case i belongs to group k. We assume that prior to analysis, the group labels $\boldsymbol{g}$ have been determined by external information derived from other data or *a priori* knowledge. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ be Bernoulli variables, where $\theta_i = 1$ indicates that case i is a non-null and $\theta_i = 0$ otherwise. Given $\boldsymbol{g}$, $\boldsymbol{\theta}$ can be grouped as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) = \{(\theta_{k1}, \cdots, \theta_{km_k}) : k = 1, \cdots, K\}$, where m_k is the number of hypotheses in group k. Different

from $\boldsymbol{g}$, $\boldsymbol{\theta}$ are unknown and need to be inferred from observations $\boldsymbol{x}$. Let $\theta_{ki}, i = 1, \cdots, m_k$, be independent Bernoulli (p_k) variables and $\boldsymbol{X} = (X_{ki})$ be generated conditional on $\boldsymbol{\theta}$:

$$X_{ki}|\theta_{ki} \sim (1-\theta_{ki})F_{k0} + \theta_{ki}F_{k1},\ i = 1, \cdots, m_k,\ k = 1, \cdots, K. \tag{2.4}$$

Hence within group k, the X_{ki}'s, $i = 1, \cdots, m_k$, are i.i.d. observations with mixture distribution $F_k = (1-p_k)F_{k0} + p_kF_{k1}$. Denote for group k the mixture density by f_k, the null and non-null densities by f_{k0} and f_{k1}, respectively. Then $f_k = (1-p_k)f_{k0} + p_kf_{k1}$.

We first consider the problem in an ideal setting where all distributional information is assumed to be known. Pooled and separate analyses are discussed in Section 2.3 and 2.4, respectively.

2.3 Pooled FDR analysis

A natural and naive approach to testing grouped hypotheses is to simply ignore the group labels and combine all cases into a pooled sample. Denote by f the mixture density,

$$f = \sum_k \pi_k[(1-p_k)f_{k0} + p_kf_{k1}] = (1-p)f_0^* + pf_1^*,$$

where $p = \sum_k \pi_k p_k$ is the non-null proportion of the pooled sample, and $f_0^* = \sum_k[(\pi_k - \pi_k p_k)/(1-p)]f_{k0}$ and $f_1^* = \sum_k(\pi_k p_k/p)f_{k1}$ are the pooled or global null and non-null densities, respectively. Denote the pooled null distribution by $F_0^* = \sum_k[(\pi_k - \pi_k p_k)/(1-p)]F_{k0}$.

In a pooled analysis, the group labels are ignored and one tests against the common pooled null distribution F_0^* in all individual tests. Define the pooled Lfdr statistic (PLfdr) by

$$\text{PLfdr}(x_i) = \frac{(1-p)f_0^*(x_i)}{f(x_i)},\ i = 1, \cdots, m. \tag{2.5}$$

The results in Sun and Cai (2007) imply that among all testing procedures that adopt the pooled-analysis strategy, the optimal one is

$$\boldsymbol{\delta}(\mathbf{PLfdr}, c_{OR}(\alpha)\mathbf{1}) = [I\{\text{PLfdr}(x_i) < c_{OR}(\alpha)\} : i = 1, \cdots, m], \tag{2.6}$$

where $c_{OR}(\alpha)$ is the largest cutoff for the PLfdr statistic that controls the overall FDR at level α. Let $\text{PLfdr}_{(1)}, \cdots, \text{PLfdr}_{(m)}$ be the ranked PLfdr values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses. An asymptotically equivalent version of (2.6) is the *PLfdr procedure*:

$$\text{Reject all } H_{(i)},\ i = 1, \cdots, l, \quad \text{where } l = \max \left\{ i : (1/i) \sum_{j=1}^{i} \text{PLfdr}_{(j)} \leq \alpha \right\}. \tag{2.7}$$

The following theorem shows that the PLfdr procedure is valid for FDR control when testing against the pooled null distribution F_0^*.

Theorem 1 *Consider the mixture model (2.4). Let $PLfdr_{(i)}, i = 1, \cdots, m$, be the ranked PLfdr values defined in (2.5). Then the PLfdr procedure (2.7) controls the FDR at level α when testing against the pooled null distribution F_0^*.*

Remark 1 We should emphasize here that a pooled analysis makes sense only when the null distributions F_{k0} are the same for all groups, in which case F_0^* coincides with the common group null. When F_{k0} are different across groups, in general the pooled null distribution F_0^* differs from any of the group null F_{k0}. In this case a pooled analysis is not appropriate at all because for each individual case a rejection against F_0^* does not imply rejection against the null distribution F_{k0} for a given group. To further illustrate this important point, let us take the most extreme case. Consider two groups where the null distribution of the first group is the alternative distribution of the second, and vice versa. It is then impossible to decide whether a case is a null or non-null without knowing the grouping information. In this case F_0^* is not the right null distribution to test against for any individual tests and therefore it is entirely inappropriate to perform a pooled analysis.

2.4 Separate FDR analysis

Another natural approach to testing grouped hypotheses is the *separate analysis* where each group is analyzed separately at the same FDR level α. Define the conditional Lfdr for group k as

$$\text{CLfdr}^k(x_{ki}) = \frac{(1 - p_k) f_{k0}(x_{ki})}{f_k(x_{ki})}, i = 1, \cdots, m_k;\ k = 1, \cdots, K. \tag{2.8}$$

Again implied by the results in Sun and Cai (2007), the optimal procedure for testing hypotheses from group k is of the form

$$\boldsymbol{\delta}^k(\mathbf{CLfdr}^k, c^k_{OR}(\alpha)\mathbf{1}) = [I\{\text{CLfdr}^k(x_{ki}) < c^k_{OR}(\alpha)\} : i = 1, \cdots, m_k], k = 1, \cdots, K, \qquad (2.9)$$

where $c^k_{OR}(\alpha)$ is the largest cutoff for CLfdr statistic that controls the FDR of group k at level α. By combining testing results from separate groups together, we have $\boldsymbol{\delta} = (\boldsymbol{\delta}^1, \cdots, \boldsymbol{\delta}^K)$.

Similarly we can propose the separated Lfdr *(SLfdr) procedure* that is asymptotically equivalent to (2.9). Denote by $\text{CLfdr}^k_{(1)}, \cdots, \text{CLfdr}^k_{(m_k)}$ the ranked CLfdr values in group k and $H^k_{(1)}, \cdots, H^k_{(m_k)}$ the corresponding hypotheses. The testing procedure for group k is:

$$\text{Reject all } H^k_{(i)},\ i = 1, \cdots, l_k, \text{ where } l_k = \max\left\{ i : (1/i)\sum_{j=1}^{i} \text{CLfdr}^k_{(j)} \le \alpha \right\}. \qquad (2.10)$$

The final rejection set of the SLfdr procedure is obtained by combining the K rejection sets from all separate analyses: $\mathcal{R}_{\text{SLfdr}} = \cup_{k=1}^{K}\{H^k_{(i)} : i = 1, \cdots, l_k\}$. The next theorem shows that the SLfdr procedure is also valid for global FDR control.

Theorem 2 *Consider the random mixture model (2.4). Let* $CLfdr^k_{(i)}, i = 1, \cdots, m_k$, $k = 1, \cdots, K$, *be the ranked CLfdr values defined by (2.8) for group* k. *Then the SLfdr procedure (2.10) controls the global FDR at level* α.

3 Optimal FDR Procedures for Testing Grouped Hypotheses

In Section 2 we discussed two natural approaches to testing grouped hypotheses: the pooled analysis and the separate analysis. Although both procedures are valid, they are inefficient in reducing the overall FNR. In this section, we begin by considering an ideal setting where all distributional information is known and propose an optimal (oracle) FDR procedure that uniformly outperforms both the pooled and separate procedures. We then turn to the situation where the distributions are unknown and introduce a data-driven procedure that is asymptotically valid and optimal.

3.1 Oracle procedure

In Section 4, we will show that the optimal testing procedure that minimizes the overall FNR subject to a constraint on the overall FDR level is the following *oracle procedure:*

$$\boldsymbol{\delta}[\mathbf{CLfdr}, c_{OR}(\alpha)\mathbf{1}] = [I\{\text{CLfdr}^k(x_{ki}) < c_{OR}(\alpha)\} : i = 1, \cdots, m_k, k = 1, \cdots, K], \tag{3.1}$$

where $c_{OR}(\alpha) = \sup\{c \in (0,1) : \text{FDR}(c) \leq \alpha\}$ is the optimal cutoff for the CLfdr statistic that controls the overall FDR at a given level α. Note that different from (2.9), the oracle procedure (3.1) suggests using a universal cutoff for all CLfdr statistics regardless of their group identities.

However, for a given FDR level, it is difficult to calculate the optimal cutoff $c_{OR}(\alpha)$ directly. An asymptotically equivalent procedure to (3.1) is the *CLfdr procedure* derived in Section 4.3. The CLfdr procedure involves the following three steps:

1. Calculate the CLfdr values for separate groups based on (2.8).

2. Combine and rank the CLfdr values from all groups. Denote by $\text{CLfdr}_{(1)}, \cdots, \text{CLfdr}_{(m)}$ the ranked CLfdr values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses.

3. Reject all $H_{(i)}$, $i = 1, \cdots, l$, where $l = \max\left\{i : (1/i)\sum_{j=1}^{i} \text{CLfdr}_{(j)} \leq \alpha\right\}$.

Remark 2 It is important to note that in Step 1, the external information of group labels is utilized to calculate the CLfdr statistic; this is the feature from a separate analysis. However, in Steps 2 and 3, the group labels are dropped and the rankings of all hypotheses are determined globally; this is the feature from a pooled analysis. Therefore the CLfdr procedure is a hybrid strategy that enjoys features from both pooled and separate analyses.

Remark 3 Unlike for the separate analysis, the group-wise FDR levels of the CLfdr procedure are in general different from α. In addition to its validity for a pooled FDR analysis, one may be interested in knowing the actual group-wise FDR levels FDR^k yielded by the CLfdr procedure; this can be conveniently obtained based on the quantities that we have already calculated. Specifically, let R_k be the number of rejections in group k. The actual FDR^k's can

be consistently estimated by

$$\widehat{\mathrm{FDR}}^k = \frac{1}{R_k}\sum_{i=1}^{R_k} \mathrm{CLfdr}^k_{(i)}. \tag{3.2}$$

The result is formally stated in Theorem 4.

Theorem 3 shows that the CLfdr is a valid procedure for global FDR control.

Theorem 3 *Consider the random mixture model (2.4). Then the CLfdr procedure controls the global FDR at level α.*

The next theorem, together with Theorem 3, shows that the CLfdr procedure is asymptotically equivalent to the oracle procedure (3.1).

Theorem 4 *Consider the random mixture model (2.4) and the CLfdr procedure, then:*

(i). The FNR level yielded by the CLfdr procedure at FDR level α is $FNR_{OR}(\alpha)+o(1)$, where $FNR_{OR}(\alpha)$ is the mFNR level of the oracle procedure (3.1).

(ii). The group-wise FDR levels of the CLfdr procedure can be consistently estimated by $\widehat{FDR}^k = FDR^k_{OR}+o_p(1)$, where $\widehat{FDR}^k$ is defined by (3.2), and FDR^k_{OR} is the group-wise FDR level of the oracle procedure (3.1).

3.2 Data-driven procedure

The CLfdr oracle procedure requires the distributional information of all individual groups. However, this information is usually unknown in practice. A commonly used strategy is to first estimate the unknown distributions and then plug-in the estimates. Estimates of the null distribution and proportion of non-nulls in a normal mixture model are provided in Efron (2004) and Jin and Cai (2007). Consider the following normal mixture model

$$X_i \sim (1-p)N(\mu_0,\sigma_0^2) + pN(\mu_i,\sigma_i^2), \tag{3.3}$$

where (μ_i,σ_i^2) follows some bivariate distribution $F(\mu,\sigma^2)$. This model can be used to approximate many mixture distributions and is found in a wide range of applications, see, e.g.,

Magder and Zeger (1996). Jin and Cai (JC, 2007) developed a procedure for estimating both the null distribution $N(\mu_0, \sigma_0^2)$ and proportion of non-null effects p in model (3.3) based on the empirical characteristic function and Fourier analysis. JC's method can be applied to separate groups directly, and the estimates are uniformly consistent over a wide class of parameters. Let $\hat{p}_k$, $\hat{f}_{k0}$ and $\hat{f}_k$ be estimates obtained for separate groups, the data-driven procedure is given as follows:

1. Calculate the plug-in CLfdr statistic $\widehat{\text{CLfdr}}^k(x_{ki}) = (1-\hat{p}_k)\hat{f}_{k0}(x_{ki})/\hat{f}_k(x_{ki})$.

2. Combine and rank the plug-in CLfdr values from all groups. Denote by $\widehat{\text{CLfdr}}_{(1)}, \cdots, \widehat{\text{CLfdr}}_{(m)}$ the ranked values and $H_{(1)}, \cdots, H_{(m)}$ the corresponding hypotheses.

3. Reject all $H_{(i)}$, $i = 1, \cdots, l$, where $l = \max\left\{i : (1/i)\sum_{j=1}^{i} \widehat{\text{CLfdr}}_{(j)} \leq \alpha\right\}$.

The actual group-wise FDR levels of the data-driven procedure can be consistently estimated as $\widehat{\text{FDR}}^k = (1/R_k)\sum_{i=1}^{R_k} \widehat{\text{CLfdr}}^k_{(i)}$, where R_k is the number of rejections in Group k.

The next theorem shows that the data-driven procedure is *asymptotically valid and optimal* in the sense that both the FDR and FNR levels of the oracle procedure are asymptotically achieved by the data-driven procedure.

Theorem 5 *Consider the multiple group model (2.4). Let $\hat{p}_k$, $\hat{f}_{k0}$ and $\hat{f}_k$ be consistent estimates of p_k, f_{k0} and f_k such that $\hat{p_k} \xrightarrow{p} p_k$, $E\|\hat{f}_{k0} - f_{k0}\|^2 \to 0$, $E\|\hat{f}_k - f_k\|^2 \to 0$, $k = 1, \cdots, K$. Let $\widehat{CLfdr}^k(x_{ki}) = (1-\hat{p})\hat{f}_0(x_{ki})/\hat{f}(x_{ki})$, $i = 1, \cdots, m_k, k = 1, \cdots, K$. Combine all test statistics from separate groups and let $\hat{CLfdr}_{(1)}, \cdots, \hat{CLfdr}_{(m)}$ be the ranked values. Then*

(i). The mFDR and mFNR levels of the data-driven procedure are respectively $\alpha + o(1)$ and $mFNR_{OR} + o(1)$, where $mFNR_{OR}$ is the mFNR level of the oracle procedure (3.1).

(ii). The mFDR level of the data driven procedure in group k can be consistently estimated as $\widehat{FDR}^k = (1/R_k)\sum_{i=1}^{R_k} \widehat{CLfdr}^k_{(i)}$. In addition, $\widehat{FDR}^k = mFDR^k_{OR} + o(1)$, where $mFDR^k_{OR} + o(1)$ is the mFDR level of the oracle procedure (3.1) in group k.

4 Compound Decision Theory for Simultaneous Testing of Grouped Hypotheses

In this section, we develop a compound decision theoretic framework for testing hypotheses arising from the multiple group model (2.4), and derive the optimal (oracle) testing procedure. We then show that the CLfdr procedure is an asymptotically equivalent version of the optimal procedure, hence the superiority of the CLfdr procedure is justified.

4.1 Compound decision problem

Consider an inference problem for the multiple group model (2.4) where the goal is to select interesting cases from each group with the overall FDR level controlled at α and the overall FNR level minimized. A solution to this problem can be represented by a general decision rule $\boldsymbol{\delta} = (\delta_{ki}) \in \{0,1\}^m$, where $\delta_{ki} = 1$ indicates that we claim case i in group k is a non-null and $\delta_{ki} = 0$ otherwise. In an FDR analysis, the m decisions are combined and evaluated as a whole; this is referred to as a *compound decision problem* (Robbins 1951).

Since hypotheses within each group are exchangeable, we consider a decision rule defined in terms of statistic $\boldsymbol{T} = \{T_k(x_{ki}) : k = 1, \cdots, K; i = 1, \cdots, m_k\}$ and threshold t such that $\boldsymbol{\delta}(\boldsymbol{T}, t) = (\delta_{ki}) = (I\{T_k(x_{ki}) < t\} : k = 1, \cdots, K; i = 1, \cdots, m_k)$, where the function T_k is the same for all observations in group k but may be different across groups. We allow T_k to depend on unknown quantities, such as the non-null proportion and null, non-null distributions in group k. In addition, T_k are standardized so that the threshold t is universal for all tests.

The multiple testing problem is closely connected to a *weighted classification problem*. Suppose the relative cost of a false positive (type I error) to a false negative (type II error) is known to be λ. Let $\boldsymbol{\delta} = (\delta_{ki} : k = 1, \cdots, K; i = 1, \cdots, m_k) \in \{0,1\}^m$ be a classification rule, where $\delta_{ki} = 1$ indicates that we classify case i of the kth group as a non-null and $\delta_{ki} = 0$ otherwise. Define the loss function

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = (1/m) \sum_{k=1}^{K} \sum_{i=1}^{m_k} \lambda(1 - \theta_{ki})\delta_{ki} + \theta_{ki}(1 - \delta_{ki}). \tag{4.1}$$

The goal in a weighted classification problem is to find $\boldsymbol{\delta} \in \{0,1\}^m$ that minimizes the classification risk $E[L_\lambda(\boldsymbol{\theta}, \boldsymbol{\delta})]$. We will show that the optimal procedure for testing grouped hypotheses can be conveniently derived by studying the optimal classification rule for the multiple group model (2.4).

Note that $\{X_{ki} : i = 1, \cdots, m_k\}$ are i.i.d. random variables, we assume that $T_k(X_{ki}) \sim G_k = (1-p_k)G_{k0} + p_k G_{k1}$, where G_{k0} and G_{k1} are the conditional cdf's of $T_k(X_{ki})$ under the null distribution F_{k0} and alternative distribution F_{k1}, respectively. The pdf of $T_k(X_{ki})$ is $g_k = (1-p_k)g_{k0} + p_k g_{k1}$, with g_{k0} and g_{k1} the corresponding conditional pdf's. Let $\tilde{G}_k(t) = 1 - G_k(t)$, $\tilde{G}_{k1}(t) = 1 - G_{k1}(t)$, $\bar{g}_0(t) = \sum_k [\pi_k(1-p_k)/(1-p)] g_{k0}(t)$, and $\bar{g}_1(t) = \sum_k (\pi_k p_k / p) g_{k1}(t)$. For a given test statistic $\boldsymbol{T}$, the mFDR and mFNR are functions of the threshold t:

$$\text{mFDR}(t) = \frac{\sum_k \pi_k (1-p_k) G_{k0}(t)}{\sum_k \pi_k G_k(t)} \text{ and } \text{mFNR}(t) = \frac{\sum_k \pi_k p_k \tilde{G}_{k1}(t)}{\sum_k \pi_k \tilde{G}_k(t)}, \tag{4.2}$$

We consider a class of test statistics $\mathcal{T}$ satisfying the monotone ratio condition (MRC):

$$\bar{g}_1(t)/\bar{g}_0(t) \text{ is decreasing in } t. \tag{4.3}$$

The following shows that the MRC is a desirable condition.

Proposition 1 *Consider the random mixture model (2.4). Let $\boldsymbol{T} = \{T_k(x_{ki})\}$ be a statistic that satisfies the MRC (4.3).*

(i) Suppose $\boldsymbol{T}$ is used for the multiple testing problem, then the mFDR (mFNR) level of testing procedure $\boldsymbol{\delta} = I(\boldsymbol{T} < t\mathbf{1})$ increases (decreases) in the threshold t. Therefore the mFNR is decreasing in the mFDR.

(ii) Suppose $\boldsymbol{T}$ is used for the weighted classification problem, then $c(\lambda)$, the optimal cutoff for $\boldsymbol{T}$ that minimizes the classification risk, is decreasing in λ, where λ is the relative weight of a false positive to a false negative.

The first part of Proposition 1 implies that in a multiple testing problem, we shall choose the largest mFDR/cutoff to minimize the mFNR level when the MRC holds. This property

is useful to determine a cutoff for this constrained minimization problem, and is conceptually reasonable as a requirement for a multiple testing procedure. In addition, the MRC class $\mathcal{T}$ is fairly general. For example, the condition in Genovese and Wasserman (2002 and 2004) and Storey (2002) that the non-null cdf of p-value is concave implies that the p-value vector $\boldsymbol{P} = (P_1, \cdots, P_m)$ belong to the MRC class $\mathcal{T}$. See Sun and Cai (2007) for more discussions about the MRC condition.

4.2 Multiple testing via weighted classification

We now connect the multiple testing and weighted classification problems by showing that the two problems are "equivalent" under mild conditions. We then derive the optimal weighted classification rule and propose the optimal testing procedure. Consider a class of decision rules $\mathcal{D}$ that are of the form $\boldsymbol{\delta} = I(\boldsymbol{T} < t\mathbf{1})$ with $\boldsymbol{T} \in \mathcal{T}$. The next theorem shows that under mild conditions, the optimal weighted classification rule is also optimal for multiple testing.

Theorem 6 *Consider the random mixture model (2.4). Suppose the classification risk with the loss function* $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = (1/m)\sum_{k=1}^{K}\sum_{i=1}^{m_k} \lambda(1-\theta_{ki})\delta_{ki} + \theta_{ki}(1-\delta_{ki})$ *is minimized by* $\boldsymbol{\delta}^{\lambda}(\boldsymbol{T}, c(\lambda)) = I(\boldsymbol{T} < c(\lambda)\mathbf{1})$*, so that* $\boldsymbol{T}$ *is the optimal statistic in the weighted classification problem. If* $\boldsymbol{T}$ *belongs to* $\mathcal{T}$*, then* $\boldsymbol{T}$ *is also the optimal statistic in the multiple-testing problem in the sense that for each global mFDR level* α*, there exists a unique* $c(\alpha)$ *such that* $\boldsymbol{\delta}^{\alpha}(\boldsymbol{T}, c(\alpha)) = I(\boldsymbol{T} < c(\alpha)\mathbf{1})$ *controls the global mFDR at level* α *with the smallest global mFNR among all decision rules in* $\mathcal{D}$ *at global mFDR level* α*.*

We consider an ideal setting where an oracle knows p_k, f_{k0} and f_{k1}, $k = 1, \cdots, K$. In this case, the optimal classification rule is given by the next theorem.

Theorem 7 *Consider the random mixture model (2.4). Suppose* p_k*,* f_{k0}*,* f_{k1} *are known. Then the classification risk with loss function (4.1) is minimized by* $\boldsymbol{\delta}^{\lambda} = (\delta_{ki})$*, where*

$$\delta_{ki} = I\left\{\Lambda^k(x_{ki}) = \frac{(1-p_k)f_{k0}(x_{ki})}{p_k f_{k1}(x_{ki})} < \frac{1}{\lambda}\right\}. \tag{4.4}$$

Note that $\Lambda^k(x) = \text{CLfdr}^k(x)/[1 - \text{CLfdr}^k(x)]$ is strictly increasing in $\text{CLfdr}^k(x)$, where $\text{CLfdr}^k(x)$ is the conditional local false discovery rate defined in (2.8), an equivalent optimal test statistic is $\mathbf{CLfdr} = [\text{CLfdr}^k(x_{ki}) : i = 1, \cdots, m_k, k = 1, \cdots, K]$. Theorems 6 and 7 together imply that the optimal testing procedure is of the form $\boldsymbol{\delta}[\mathbf{CLfdr} < c(\alpha)]$. Proposition 1 indicates that the cutoff should be chosen as $c_{OR}(\alpha) = \sup\{c \in (0,1) : \text{mFDR}(c) \leq \alpha\}$. Therefore the optimal (oracle) procedure for multiple group hypothesis testing is

$$\boldsymbol{\delta}[\mathbf{CLfdr}, c_{OR}(\alpha)] = [I\{\text{CLfdr}^k(x_{ki}) < c_{OR}(\alpha)\} : i = 1, \cdots, m_k, k = 1, \cdots, K]. \tag{4.5}$$

For a given FDR level, it is difficult to calculate the optimal cutoff $c_{OR}(\alpha)$ directly. The difficulty can be circumvented by using the *CLfdr procedure* proposed in Section 3.1, where c_{OR} is estimated consistently based on a simple step-up procedure.

4.3 The derivation of the CLfdr procedure

This section demonstrates that the CLfdr procedure can be used to approximate the oracle procedure (3.1). The essential idea in the derivation is to first evaluate the distributions of the CLfdr statistic empirically, then estimate the mFDR for a given cutoff, and finally choose the largest cutoff c subject to the constraint $\widehat{\text{mFDR}}(c) \leq \alpha$. Let G_k and G_{k0} be the marginal cdf and null cdf of $\text{CLfdr}^k(X_{ki})$, then the mFDR of testing rule $\boldsymbol{\delta}(\mathbf{CLfdr}, c\mathbf{1})$ is

$$\text{mFDR}(c) = \frac{\sum_k \pi_k (1 - p_k) G_{k0}(c)}{\sum_k \pi_k G_k(c)}. \tag{4.6}$$

Note that $\sum_k \pi_k G_k(c) = \sum_k \pi_k \int I[\text{CLfdr}^k(x) < c] f_k(x) dx$; hence it can be estimated by

$$\sum_{k=1}^{K} \pi_k \frac{1}{m_k} \sum_{i=1}^{m_k} I[\text{CLfdr}^k(x_{ki}) < c] = (1/m) \sum_{k=1}^{K} \sum_{i=1}^{m_k} I[\text{CLfdr}^k(x_{ki}) < c].$$

Next note that

$$\begin{aligned}\sum_k \pi_k(1-p_k)G_{k0}(c) &= \sum_{k=1}^K \pi_k \int I[\text{CLfdr}^k(x) < c](1-p_k)f_{k0}(x)dx \\ &= \sum_{k=1}^K \pi_k \int I[\text{CLfdr}^k(x) < c]\text{CLfdr}^k(x)f_k(x)dx,\end{aligned}$$

which can be estimated by $(1/m)\sum_{k=1}^K \sum_{i=1}^{m_k} I[\text{CLfdr}^k(x_{ki}) < c]\text{CLfdr}^k(x_{ki})$. Therefore the mFDR($c$) can be estimated by

$$\widehat{\text{mFDR}}(c) = \frac{\sum_{k=1}^K \sum_{i=1}^{m_k} I[\text{CLfdr}^k(x_{ki}) < c]\text{CLfdr}^k(x_{ki})}{\sum_{k=1}^K \sum_{i=1}^{m_k} I[\text{CLfdr}^k(x_{ki}) < c]}. \tag{4.7}$$

Suppose a total of R hypotheses are rejected, then (4.7) reduces to $\widehat{\text{mFDR}} = (1/R)\sum_{j=1}^R \text{CLfdr}_{(j)}$, where $\text{CLfdr}_{(1)}, \cdots, \text{CLfdr}_{(m)}$ are obtained by the ranking all m CLfdr values (calculated for separate groups): $\{\text{CLfdr}^k(x_{ki}) : i = 1, \cdots, m_k, k = 1, \cdots, K\}$. The group labels are no longer needed and hence dropped. Note $\widehat{\text{mFDR}}(R) = (1/R)\sum_{j=1}^R \text{CLfdr}_{(j)}$ is strictly increasing in R (since $\widehat{\text{mFDR}}(R+1) - \widehat{\text{mFDR}}(R) = [1/(R^2+R)]\sum_{j=1}^R(\text{CLfdr}_{(R+1)} - \text{CLfdr}_{(j)}) > 0$), we choose the largest R such that the mFDR is controlled at level α. Hence a natural testing procedure is:

$$\text{Reject all } H_{(i)},\ i = 1, \cdots, l, \quad \text{where } l = \max\Big\{i : (1/i)\textstyle\sum_{j=1}^i \text{CLfdr}_{(j)} \le \alpha\Big\}.$$

Thus we have derived the CLfdr procedure from the oracle procedure (3.1).

5 Numerical Results

Now we turn to the numerical performance of the PLfdr, SLfdr and CLfdr procedures. Section 5.1 compares the three procedures under a two-group model in an oracle setting. More complicated settings are considered in Section 5.2 where (i) the number of groups is greater than two, and (ii) the null and non-null distributions are unknown and need to be estimated. A real data analysis is discussed in Section 6.

5.1 Comparison of oracle procedures in a two-group model

Consider the following two-group normal mixture model:

$$X_{ki} \sim (1-p_k)N(\mu_{k0}, \sigma_{k0}^2) + p_k N(\mu_k, \sigma_k^2), \quad k = 1, 2. \tag{5.1}$$

The numerical performances of the PLfdr, SLfdr and CLfdr procedures are investigated in the next two simulation studies. The nominal global FDR level is 0.10 for all simulations.

Simulation Study 1 *Identical null distributions.* The null distributions of both groups are fixed as $N(0,1)$. Three simulation settings are considered: (i) The group sizes are $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = (1-p_1)N(0,1) + p_1 N(-2,1)$ and $f_2 = 0.9N(0,1) + 0.1N(4,1)$. We vary p_1, the proportion of non-nulls in group 1, and plot the FDR and FNR levels as functions of p_1. (ii) The groups sizes are also $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = 0.8N(0,1) + 0.2N(\mu_1, 1)$ and $f_2 = 0.9N(0,1) + 0.1N(2, 0.5^2)$. The FDR and FNR levels are plotted as functions of μ_1. (iii) The marginal pdf's are $f_1 = 0.8N(0,1) + 0.2N(-2, 0.5^2)$ and $f_2 = 0.9N(0,1) + 0.1N(4,1)$. The sample size of group 2 is fixed at $m_2 = 1500$, the FDR and FNR levels are plotted as functions of m_1. The simulation results with 500 replications are given in Figure 4. The top row compares the actual FDR levels of the three procedures; the results for setting (i), (ii) and (iii) are shown in Panel (a), (b) and (c), respectively. The group-wise FDR levels of the CLfdr procedure are also provided (the dashed line for group 1 and dotted line for group 2). The bottom row compares the FNR levels of the three procedures; the results for setting (i), (ii) and (iii) are shown in Panel (d), (e) and (f), respectively.

We can see that all three procedures control the global FDR level at the nominal level 0.10, indicating that all three procedures are valid. It is important to note that the CLfdr procedure chooses group-wise FDR levels automatically (dashed and dotted lines in Panel (a)-(c)), and the levels are in general different from the nominal level 0.10. The relative efficiency of PLfdr versus SLfdr is inconclusive (depends on simulation settings). For example, the SLfdr procedure yields lower FNR levels in Panel (d), but higher FNR levels in Panel (f). However,

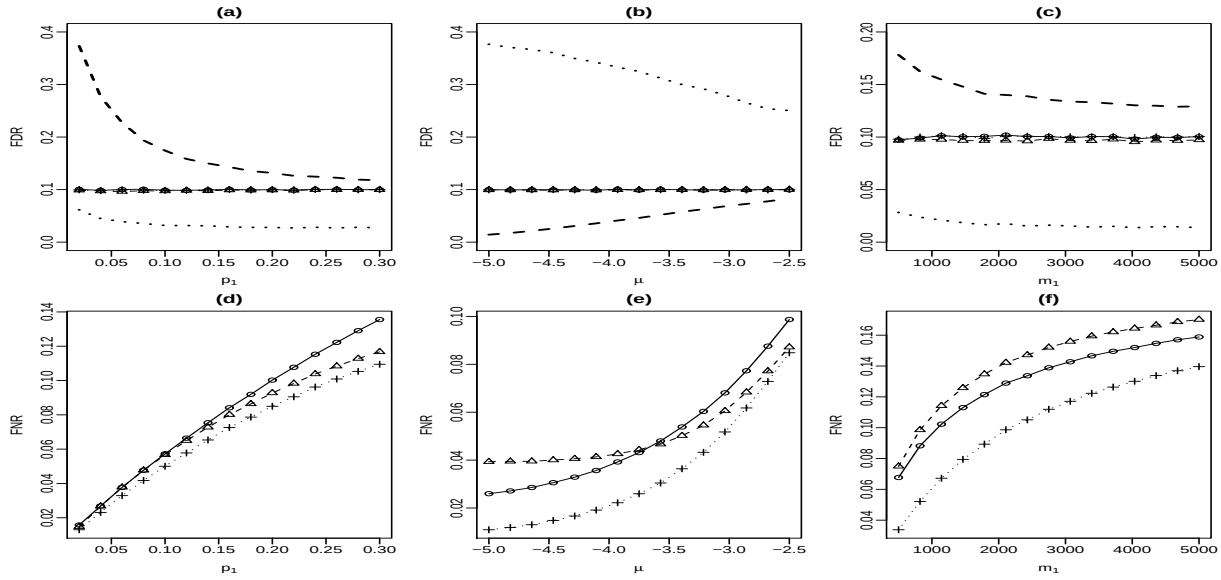

Figure 4: Results for Simulation Study 1: the top row compares the FDR levels and the bottom row compares the FNR levels ($\circ$, PLfdr; $\triangle$, SLfdr; $+$, CLfdr). The optimal group-wise FDR levels suggested by the CLfdr procedure are provided together with the global FDR levels (dashed line, Group1; dotted line, Group2).

all simulations show that both the PLfdr and SLfdr procedures are uniformly dominated by the CLfdr procedure.

Next we consider the situation where the null distributions of the observations are different. It was argued by Efron (2008a) that in this case a pooled FDR analysis becomes problematic since highly significant non-null cases from one group may be hidden among the nulls from the other group. See Remark 1 in Section 2.3.

Simulation Study 2 *Disparate null distributions.* We consider three situations where the null distributions of the two groups may differ: (i) The null means are different. The group sizes are $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = 0.8N(\mu_{10}, 1) + 0.2N(-2, 1)$ and $f_2 = 0.9N(0, 1) + 0.1N(2, 0.5^2)$. (ii) The null means are the same, but one null is more dispersed. The group sizes are $m_1 = 3000$ and $m_2 = 1500$; the group mixture pdf's are $f_1 = 0.8N(0, \sigma_{10}^2) + 0.2N(-4, 1)$ and $f_2 = 0.9N(0, 1) + 0.1N(2, 1)$. (iii) Both the null means and null variances differ. The group sizes are m_1 and 2000; the group pdf's are $f_1 = 0.8N(1.5, 1) + 0.2N(-2, 1)$ and $f_2 = 0.9N(0, 0.8^2) + 0.1N(2, 0.5^2)$. The simulation results

with 500 replications are shown in Figure 5. The top row compares the actual FDR levels of the three procedures and the bottom row compares the FNR levels of the three procedures. Again, the group-wise FDR levels of the CLfdr procedure are provided together with the global FDR levels. Results for setting (i), (ii) and (iii) are displayed in column 1, 2 and 3, respectively.

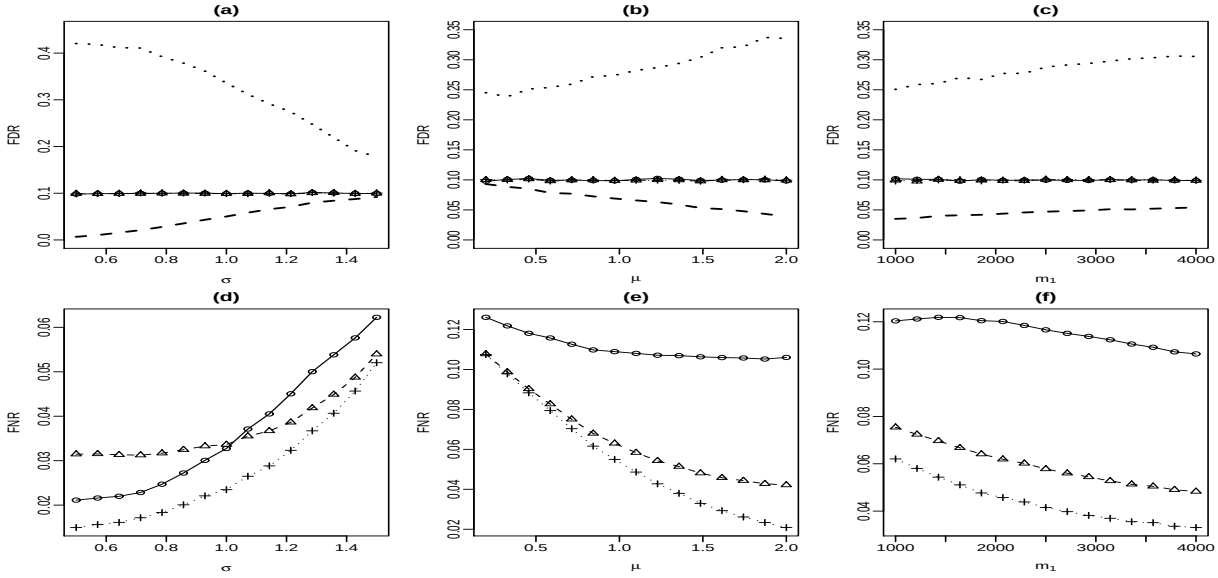

Figure 5: Results for Simulation Study 2: the top row compares the FDR levels and the bottom row compares the FNR levels ($\circ$, PLfdr; $\triangle$, SLfdr; $+$, CLfdr). The optimal group-wise FDR levels suggested by the CLfdr procedure are provided together with the global FDR levels (dashed line, Group1; dotted line, Group2).

From Figure 5 it can be similarly seen that (i) all three procedures are valid in terms of the FDR control, (ii) the group-wise FDR levels of the CLfdr procedure are different from the nominal level and from each other, and (iii) both the PLfdr and SLfdr procedures are uniformly dominated by the CLfdr procedure. It is interesting to note that in Panel (d), the PLfdr procedure is at first more efficient than the SLfdr for small σ_{10}, but becomes less and less efficient as σ_{10} increases (since more and more non-null cases from Group 2 are hidden in the nulls from Group 1). It is imporant to note that in this case the PLfdr procedure and the SLfdr procedure are testing against different null distributions.

5.2 Extended comparisons

In practice, the number of groups is often greater than two, and the distributional information for individual groups may be unknown. This section extends our previous comparisons to cover these new situations.

Simulation Study 3 *(i) More groups.* The number of groups is chosen to be $K = 5$. We consider three simulation settings, whose distributional information is summarized in the top half of Table 2. We apply the PLfdr, SLfdr and CLfdr procedures to the simulated data and obtain the FDR and FNR levels. *(ii). Unknown distributions.* The number of groups is chosen to be $K = 5$. The non-null proportions and mixture densities are unknown. We first take the approach in Jin and Cai (2007) to estimate the unknown quantities and then apply the data-driven procedures. We consider three simulation settings, whose distributional information is summarized in the bottom half of Table 2. The simulation results are displayed in Figure 6.

Table 2: Settings for simulation study 3: (a)-(e) the sample sizes of all individual groups are 2000; (f) the sample sizes are $m_1 = m_2 = m_3 = 2000$, $m_4 = m_5 = m_k$. The number of replications is 500.

Group	Panel (a)	Panel (b)	Panel (c)
1	$0.7N(0,\sigma^2)+0.3N(-4,1)$	$0.9N(0,0.5^2)+0.1N(-4,1)$	$0.7N(0,1)+0.3N(-4,1)$
2	$0.8N(0,\sigma^2)+0.2N(-2,1)$	$0.85N(0,0.5^2)+0.15N(-2,1)$	$0.8N(0,1)+0.2N(-2,1)$
3	$0.8N(0,\sigma^2)+0.2N(-1,1)$	$0.8N(0,0.5^2)+0.2N(-1,1)$	$0.8N(0,1)+0.2N(1,1)$
4	$0.9N(0,0.5^2)+0.1N(1,1)$	$0.75N(0,\sigma^2)+0.25N(1,1)$	$0.9N(\mu,0.5^2)+0.1N(1,1)$
5	$0.7N(0,0.5^2)+0.3N(2,1)$	$0.7N(0,\sigma^2)+0.3N(2,1)$	$0.7N(\mu,0.5^2)+0.3N(2,1)$
	Panel (d)	Panel (e)	Panel (f)
1	$0.7N(0,1)+0.3N(-4,1)$	$0.7N(0,1)+0.3N(-4,\sigma^2)$	$0.7N(0,1)+0.3N(-4,1)$
2	$0.8N(0,1)+0.2N(-2,1)$	$0.8N(0,1)+0.2N(-2,\sigma^2)$	$0.8N(0,1)+0.2N(-2,1)$
3	$0.8N(0,1)+0.2N(-1,1)$	$0.8N(0,1)+0.1N(-1,\sigma^2)$	$0.8N(0,1)+0.2N(1,1)$
4	$0.9N(0,1)+0.1N(\mu,1)$	$0.9N(0,1)+0.2N(1,0.5^2)$	$0.9N(0,1)+0.1N(1,0.5^2)$
5	$0.7N(0,1)+0.3N(\mu+1,1)$	$0.7N(0,1)+0.3N(2,0.5^2)$	$0.7N(0,1)+0.3N(2,0.5^2)$

Similar to the two-group case, all three procedures control the FDR at the nominal level 0.1 when we have more groups. The FNR levels of the three procedures for the three settings considered in the first row of Table 2 are displayed in Panel (a), (b) and (c) of Figure 6, respectively. We can see that both the PLfdr and SLfdr procedures are dominated by the CLfdr procedure. When the distributions are unknown and data-driven procedures are used,

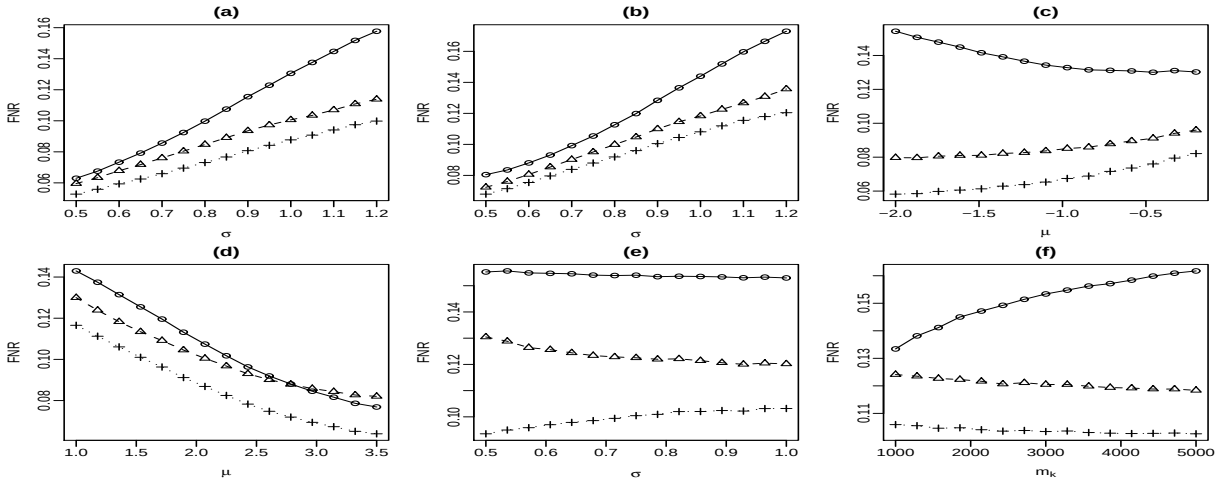

Figure 6: Results for Simulation Study 3. The FNR levels of PLfdr, SLfdr and CLfdr procedures are plotted as functions of model parameters ($\circ$, PLfdr; $\triangle$, SLfdr; $+$, CLfdr). The first row compares oracle procedures and the second row compares data-driven procedures. Detailed simulation settings are summarized in Table 2.

the FDR levels of the three procedures are approximately 0.10. The FNR levels of the three procedures for the three settings considered in the second row of Table 2 are displayed in Panel (d), (e) and (f), respectively. Again we can see that at the same FDR level 0.10, the FNR of the CLfdr procedure is uniformly smaller than those of the other two procedures.

6 Applications to the Adequate Yearly Progress Study of California High Schools

We now return to the adequate yearly progress (AYP) study mentioned in the introduction. In this section, we analyze the data collected from $m = 7867$ of California high schools (Rogosa 2003) by using the three multiple testing procedures discussed in detail in earlier sections, namely the PLfdr, SLfdr and CLfdr procedures.

The association between social-economic status (SES) and academic performance of students is an important topic in sociological research (Sparkes 1999; Considine and Zappalà 2002). One goal of the AYP study is to compare the success rates in Math exams of social-economically advantaged (SEA) versus social-economically disadvantaged (SED) students. Since the average success rates of the SEA students are in general (7370 out of 7867 schools)

higher that the SED students, it is of interest to identify a subset of schools in which the advantaged-disadvantaged performance differences are unusually small or large. Given the limited financial and educational resources, the correct identification of these "unusual" schools is important for making policies to reduce social exclusion and promote the overall performance of all students.

Denote by X_i and Y_i the success rates, and n_i and n_i' the numbers of scores reported for SEA and SED students in school i, $i = 1, \cdots, m$. Define the centering constant $\Delta = \text{median}(X_i) - \text{median}(Y_i)$. A z-value for comparing the SEA students versus the SED students can be computed for each school:

$$z_i = \frac{X_i - Y_i - \Delta}{\sqrt{X_i(1-X_i)/n_i + Y_i(1-Y_i)/n_i'}}, \tag{6.1}$$

for $i = 1, \cdots, m$. We claim school i is "interesting" if the observed $|z_i|$ is large.

The AYP data has been analyzed by Efron (2007 and 2008b), where he first estimated the global null density $\hat{f}_0$, then searched for interesting cases in the tail areas of $\hat{f}_0$. This pooled-analysis strategy ignores the fact that the hypotheses formed for different schools are not exchangeable. In particular, the number of scores reported by each school varies from less than a hundred to more than ten thousands. A pooled analysis tends to over-select too many large schools, which often express themselves as "very significant" in the tail areas due to small denominators in (6.1). In contrast, small schools are likely to be hidden in the central area of $\hat{f}_0$ and appear "uninteresting". This is not desirable because, in practice, investigators are interested in identifying significant differences from all schools, not only from large schools. As we shall see, an important feature of the AYP data is that the empirical null distributions of the z-values are substantially different for small and large schools, therefore a pooled analysis is inappropriate and one should perform a separate analysis to take into account the effect of school size. Based on a preliminary cluster analysis, we divide all schools into three groups according to the number of scores reported ($n_i + n_i'$): small schools ($n_i + n_i' \le 120$), medium schools ($120 < n_i + n_i' \le 900$) and large schools ($n_i + n_i' > 900$). The group characteristics are summarized in Table 3, where the empirical null distributions are estimated using Jin and Cai

(2007)'s method. Note that the variance of the empirical null distribution for the scores from the large schools is more than four times than those for the scores from the other two groups. See also Figure 1 in the introduction.

Table 3: Group characteristics in the AYP data: 7867 schools in total. The global null density is $\hat{f}_0 = N(-0.59, 1.59^2)$

Group	Group Definition	Group Size	Proportion	Empirical Null
Small	$n_i + n_i' \leq 120$	516	6.6%	$\hat{f}_{10} = N(-0.51, 1.27^2)$
Medium	$120 < n_i + n_i' \leq 900$	6514	80.6%	$\hat{f}_{20} = N(-0.61, 1.54^2)$
Large	$n_i + n_i' > 900$	837	12.8%	$\hat{f}_{30} = N(-0.95, 3.16^2)$

We then apply the PLfdr, SLfdr and CLfdr procedures to the AYP data at different FDR levels. The results are summarized in Table 4.

Table 4: Numbers of Interesting Cases Identified by PLfdr, SLfdr and CLfdr Procedures

	From Small Group			From Medium Group			From Large Group			Total		
FDR	PLfdr	SLfdr	CLfdr	PLfdr	SLfdr	CLfdr	PLfdr	SLfdr	CLfdr	PLfdr	SLfdr	CLfdr
0.01	2	6	6	59	47	51	171	42	39	232	95	96
0.025	6	7	6	89	67	75	203	54	50	298	128	131
0.04	6	9	9	123	89	98	215	64	58	344	162	165
0.055	6	10	10	152	109	122	222	71	64	380	190	196
0.07	7	12	12	176	130	146	233	76	66	416	218	224
0.085	7	13	12	203	150	173	241	80	69	451	243	254
0.10	7	15	15	230	173	195	249	85	72	486	273	282
0.115	8	16	18	253	194	217	259	90	75	520	300	310

The PLfdr procedure claims the most discoveries, followed by the CLfdr and then SLfdr procedure. It is important to emphasize that the PLfdr procedure is inappropriate here because the pooled null distribution is not the correct null to test against. See Remark 1 in Section 2.3. The PLfdr procedure is too liberal for the large group yet too conservative for the small group: around 50%-70% significant schools come from the large group, although its population proportion is only 13%; in contrast, only around 1% interesting cases come from the small group, although its population proportion is more than 6%. The SLfdr procedure considers the groups separately; large schools are no longer over-selected and more small schools are identified. The CLfdr procedure further improves the SLfdr procedure by efficiently exploiting

the important grouping information and weighting the numbers of discoveries among groups. The optimal group-wise FDR levels estimated by the CLfdr procedure at different nominal FDR levels are plotted in Figure 7, suggesting that we should choose higher FDR levels for the medium group and lower FDR level for the large group. Note that the SLfdr procedure uses the same FDR level for all groups, the CLfdr procedure usually identifies more cases from the medium group, but fewer cases from the large group.

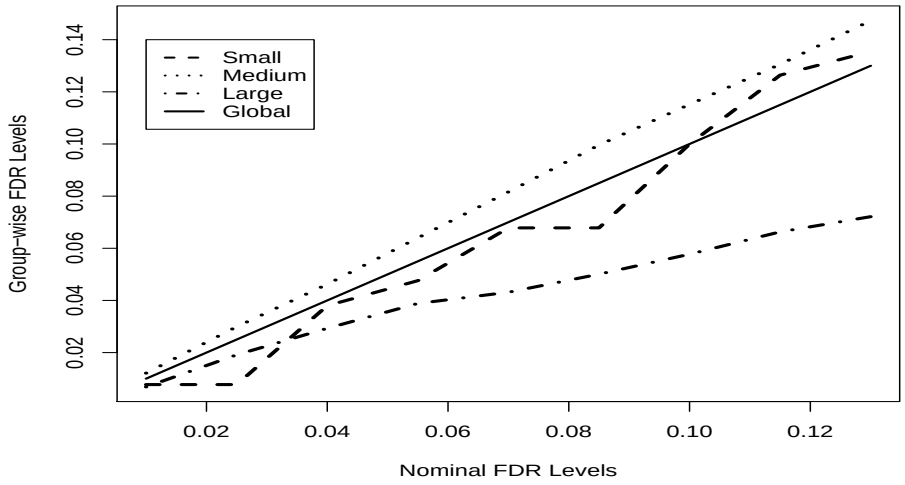

Figure 7: AYP study. Optimal group-wise FDR levels estimated by the CLfdr procedure.

7 Discussion

We have developed a compound decision theoretic framework for testing grouped hypotheses arising from the random mixture model (2.4). Both numerical and theoretical results demonstrate that conventional testing procedures can be substantially improved. In this section, we discuss how the rankings of hypotheses may be affected when hypotheses are not exchangeable; this provides additional insights on the efficiency gain of the optimal CLfdr procedure.

The concept of *symmetric rules* in compound decision theory is closely connected to the exchangeability assumption in multiple testing. Let $\boldsymbol{\delta}(\boldsymbol{x})$ be a general decision rule. Then $\boldsymbol{\delta}$ is symmetric if $\boldsymbol{\delta}[\tau(\boldsymbol{x})] = \tau[\boldsymbol{\delta}(\boldsymbol{x})]$ for all permutation operator τ (Copas 1974). Most multiple

testing procedures in the literature are symmetric rules, which implicitly assume that the hypotheses are exchangeable. However, when the hypotheses are not exchangeable (e.g., groups are heterogeneous), even the optimal symmetric rule may suffer from severe efficiency loss. To further illustrate the point, let us consider the following "optimal" procedures in the multiple testing literature.

Lfdr. The Lfdr, which corresponds to the PLfdr defined in (2.5), was shown to provide optimal rankings when all hypotheses are independent and exchangeable (Sun and Cai, 2007). In practical situations such as the AYP study, the exchangeability assumption obviously fails to hold. However, testing procedures that threshold the PLfdr statistic are symmetric rules, implying that the hypotheses are ranked only based on observed z-values. This is inappropriate in the AYP study, where the same observed z-values from small and large schools may indicate different significance levels (since the null distributions are different). The PLfdr statistic is no longer optimal because the grouping information can be exploited to construct more efficient rankings of all hypotheses.

Storey's ODP. Suppose the null hypotheses are true for tests $i = 1, \cdots, m_0$ and the alternative is true for tests $i = m_0 + 1, \cdots, m$. The null densities and non-null densities are denoted respectively by $f_1, \cdots, f_{m_0}$ and $g_{m_0+1}, \cdots, g_m$. Storey's optimal discovery procedure (ODP, Storey 2008) rejects hypothesis i if $S_{ODP}(x_i) \geq \lambda$, where

$$S_{ODP}(x) = \frac{g_{m_0+1}(x) + \cdots + g_m(x)}{f_1(x) + \cdots + f_{m_0}(x)}. \tag{7.1}$$

It was shown by Storey that the ODP maximizes the expected number of true positives (ETP) for each fixed expected number of false positives (EFP) among all single-thresholding procedures (STPs). It can be shown that all STPs (hence the ODP) are symmetric rules; therefore the optimality of the ODP is only claimed for a subclass of decision rules (symmetric rules), and can be outperformed by other asymmetric rules when hypotheses come from heterogeneous groups. There are other issues for the ODP procedure, and we briefly point out the following: (i) One needs to know which hypotheses are true and the densities of the individual test statistics in (7.1); this assumption is extremely impractical. (ii) The ODP depends on

unknown quantities that cannot be estimated from the data. The actual ODP procedure is based on an ad-hoc estimate of the thresholding function, where the optimality is lost at the estimation step. (iii) The optimal threshold is difficult to determine for a given FDR/ETP level.

Spjøtvoll's optimal procedure. Spjøtvoll (1972, Theorem 1) proposed an "optimal" procedure that maximizes the "ETP" subject to a constraint on the "EFP". In our setting, the testing procedure is reduced to the following form

$$\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m) = [I\{f_{k0}(x) > cf_{k1}(x)\} : i = 1, \cdots, m]. \tag{7.2}$$

Spjøtvoll's procedure (7.2) is not symmetric, and suggests universal thresholding of the likelihood ratio (LR) statistic. In contrast, the CLfdr procedure suggests thresholding a constant $(1 - p_k)$ times the LR. The two procedures are different when p_k varies across groups. As argued by Storey (2007), the setting considered in Spjøtvoll (1972) is problematic because of the wrong definitions of "ETP" and "EFP", which do not represent the underlying reality. It can be shown that this wrong formulation naturally leads to a procedure that ignores p_k, which provides important information for ranking the hypotheses.

Asymmetric testing rules have been recently proposed under different settings by Genovese et al. (2006), Efron (2008) and Sun and Cai (2008), among others. These works indicate that it is beneficial to treat hypotheses differently when some prior or structural information of the sample are available. The efficiency of a testing procedure can thus be improved by weighting the conventional test statistics (such as the weighted p-values, Genovese et al. 2006) or introducing new test statistics to incorporate the prior/structural information (the CLfdr statistic; the local index of significance, Sun and Cai 2008). This article studies the multiple testing problem from a compound decision theoretical perspective, which provides additional insights on the benefit of extending one's attention to a wider class of decision rules when hypotheses are grouped.

Appendix I: Proofs of Main Results

We shall prove here the main results, Theorems 1-5. The proofs of other results are given in Appendix II.

Proof of Theorem 1. Denote the group labels by $\boldsymbol{g} = (g_1, \cdots, g_m)$, i.e., $g_i = k$ if case i comes from group k. Suppose the group labels are unknown, we have

$$\begin{aligned} P(\theta_i|x_i) &= \sum_k P(\theta_i|g_i = k, x_i) P(g_i = k|x_i) \\ &= \sum_k \frac{(1-p_k) f_{k0}(x_i)}{f_k(x_i)} \frac{\pi_k f_k(x_i)}{\sum_k \pi_k f_k(x_i)} \\ &= \frac{\sum \pi_k (1-p_k) f_{k0}(x_i)}{\sum_k \pi_k f_k(x_i)} \equiv \text{PLfdr}(x_i). \end{aligned}$$

Let R and N_{10} be the number of rejections and number of false positives of the PLfdr procedure. Note $E(N_{10}|\boldsymbol{x}) = \sum_{i=1}^m I(\delta_i = 1) P(\theta_i = 0|\boldsymbol{x}) = \sum_{i=1}^R \text{PLfdr}_{(i)}$, we have

$$\begin{aligned} \text{FDR}_{\text{PLfdr}} &= E\left(N_{10}/R\right) P(R > 0) = E\left[(1/R) E(N_{10}|\boldsymbol{x})\right] P(R > 0) \\ &= E\left[\frac{1}{R} \sum_{i=1}^R \text{PLfdr}_{(i)}\right] P(R > 0). \end{aligned}$$

The claim follows by noting that the PLfdr procedure guarantees that $(1/R) \sum_{i=1}^R \text{PLfdr}_{(i)} \leq \alpha$ for all realizations of $\boldsymbol{x}$. ∎

Proof of Theorem 2. Let R_k and N_{10k} be the number of rejections and the number of false positives in group k. Note that $E(N_{10k}|\boldsymbol{g}, \boldsymbol{x}) = \sum_{i=1}^{m_k} I(\delta_{ki} = 1) P(\theta_{ki} = 0|x_{ki}) = \sum_{i=1}^{R_k} \text{CLfdr}_{(i)}^k$, and that the SLfdr procedure guarantees $(1/R_k) \sum_{i=1}^{R_k} \text{CLfdr}_{(i)}^k \leq \alpha$ for all realizations of $\boldsymbol{x}$, we have

$$\begin{aligned} \text{FDR}_{\text{SLfdr}} &= E\left(\frac{\sum_k N_{10k}}{\sum_k R_k}\right) P(\sum_k R_k > 0) \\ &= E\left\{\frac{1}{\sum_k R_k} \sum_k E(N_{10k}|\boldsymbol{g}, \boldsymbol{x})\right\} P(\sum_k R_k > 0) \\ &\leq E\left\{\frac{1}{\sum_k R_k} (\sum_k \alpha R_k)\right\} P(\sum_k R_k > 0) \leq \alpha. \quad ∎ \end{aligned}$$

Proof of Theorem 3. Let R and N_{10} be the number of rejections and number of false positives. Then $E(N_{10}|\boldsymbol{g},\boldsymbol{x}) = \sum_{i=1}^{m} I(\delta_i = 1)P(\theta_i = 0|x_i, g_i) = \sum_{i=1}^{R} \text{CLfdr}_{(i)}$. The SLfdr procedure guarantees that $(1/R)\sum_{i=1}^{R} \text{CLfdr}_{(i)} \leq \alpha$ for all realizations of $\boldsymbol{x}$, it follows that

$$\begin{aligned} \text{FDR}_{\text{CLfdr}} &= E(N_{10}/R)P(R>0) = E\left\{E(N_{10}/R|\boldsymbol{g},\boldsymbol{x})\right\}P(R>0) \\ &\leq E\left\{\frac{1}{R}\sum_{i=1}^{R}\text{CLfdr}_{(i)}\right\}P(R>0) \leq \alpha. \quad \blacksquare \end{aligned}$$

Proof of Theorem 4. (i) In the CLfdr procedure, the group labels were only used for calculating the CLfdr statistics and then dropped afterwards. Hence when the interest is to evaluate global FDR and FNR levels of the CLfdr procedure, the group labels provide no information, i.e., let $T_i = \text{CLfdr}(x_i)$, then $\{T_i : i = 1, \cdots, m\}$ can be viewed as a random sample from G_{OR}, the cdf of the pooled sample. The null, non-null cdf's of T_i are denoted by G_0^{OR} and G_1^{OR}, respectively. Let t_{OR} and $\hat{t}_{OR}$ be the thresholds of the oracle procedure and CLfdr procedure, respectively. Note

$$\text{mFNR}_{OR} = \frac{pP_{H_1}(T_i > t_{OR})}{P(T_i > t_{OR})} \text{ and } \text{mFNR}_{\text{CLfdr}} = \frac{pP_{H_1}(T_i > \hat{t}_{OR})}{P(T_i > \hat{t}_{OR})},$$

It is sufficient to show that $\hat{t}_{OR} \xrightarrow{p} t_{OR}$.

Let $Q_{OR}(t) = (1-p)G_0(t)/G(t)$ and $\hat{Q}_{OR}(t) = \{\sum_i I(T_i < t)T_i\}/\{\sum_i I(T_i < t)\}$. Applying law of large numbers, we have $(1/m)\sum_i I(T_i < t) \xrightarrow{p} E(T_i < t) = G^{OR}(t)$ and $(1/m)\sum_{i=1}^{m} I(T_i < t)T_i \xrightarrow{p} E\{I(T_i < t)T_i\} = E[E\{I(T_i < t)T_i|g_i\}] = \pi_k(1-p_k)G_{k0}^{OR}(t) = (1-p)G_0^{OR}(t)$. It follows that $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$. Subsequent arguments are similar to the proof of Lemma A.5 in Sun and Cai (2007). Note that $\hat{Q}_{OR}(t)$ is a step function with jump at $T_{(i)}$. For $T_{(k)} < t < T_{(k+1)}$, we construct an envelope for $\hat{Q}_{OR}(t)$ using two monotone continuous functions:

$$\begin{aligned} \hat{Q}_{OR}^{-}(t) &= \frac{T_{(k+1)}-t}{T_{(k+1)}-T_{(k)}}\hat{Q}_{OR}(T_{(k-1)}) + \frac{t-T_{(k)}}{T_{(k+1)}-T_{(k)}}\hat{Q}_{OR}(T_{(k)}); \\ \hat{Q}_{OR}^{+}(t) &= \frac{T_{(k+1)}-t}{T_{(k+1)}-T_{(k)}}\hat{Q}_{OR}(T_{(k)}) + \frac{t-T_{(k)}}{T_{(k+1)}-T_{(k)}}\hat{Q}_{OR}(T_{(k+1)}). \end{aligned}$$

It can be shown that (i) $\hat{Q}_{OR}^{+}(t) \geq \hat{Q}_{OR}(t) \geq \hat{Q}_{OR}^{-}(t)$, (ii) $\hat{Q}_{OR}^{+}(t)$ and $\hat{Q}_{OR}^{-}(t)$ are strictly

increasing in t, and (iii) $|\hat{Q}^+_{OR}(t) - \hat{Q}^-_{OR}(t)| \leq 1/R(t) \xrightarrow{p} 0$. Note that $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$, we have $\hat{Q}^-_{OR}(t) \xrightarrow{p} Q_{OR}(t)$ and $\hat{Q}^+_{OR}(t) \xrightarrow{p} Q_{OR}(t)$.

Now define $\hat{t}^-_{OR} = \sup\{t \in (0,1) : \hat{Q}^-_{OR}(t) \leq \alpha\}$ and $\hat{t}^+_{OR} = \sup\{t \in (0,1) : \hat{Q}^+_{OR}(t) \leq \alpha\}$; then $\hat{t}^+_{OR} \leq \hat{t}_{OR} \leq \hat{t}^-_{OR}$. We claim that $\hat{t}^-_{OR} \xrightarrow{p} t_{OR}$. If not, there exists ϵ_0 and η_0 such that for any $M > 0$, $P(|\hat{t}^-_{OR} - t_{OR}| > \epsilon_0) \geq 4\eta_0$ holds for some $\mathbb{Z}^+ \ni m \geq M$. Suppose $P(K^1_m) = P(\hat{t}^-_{OR} t_{OR} > \epsilon_0) \geq 2\eta_0$. The MRC implies that $Q_{OR}(t)$ is strictly increasing in t and $Q_{OR}(t_{OR}) = \alpha$. Let $2\delta_0 = Q_{OR}(t_{OR} + \epsilon_0) - \alpha$. $\hat{Q}_{OR}(t) \xrightarrow{p} Q_{OR}(t)$ implies that there exists M such that $P(K^2_m) = P(|\hat{Q}^-_{OR}(t_{OR} + \epsilon_0) - Q_{OR}(t_{OR} + \epsilon_0)| < \delta_0) \geq 1 - \eta_0$ holds for all $m \geq M$. Consider $K_m = K^1_m \cap K^2_m$, then there exists $m \in \mathbb{Z}^+$ such that $P(K_m) \geq \eta_0$. However, note $\hat{Q}^-_{OR}(t)$ is strictly increasing in t, on K_m we must have $\alpha = \hat{Q}^-_{OR}(\hat{t}^-_{OR}) > \hat{Q}^-_{OR}(t_{OR} + \epsilon_0) > Q_{OR}(t_{OR} + \epsilon_0) - \eta_0 = \alpha + \delta_0$. Hence K_m cannot have positive measure. This is a contradiction. Therefore we must have $\hat{t}^-_{OR} \xrightarrow{p} t_{OR}$. Similarly we can show that $\hat{t}^+_{OR} \xrightarrow{p} t_{OR}$. Note $\hat{t}^+_{OR} \leq \hat{t}_{OR} \leq \hat{t}^-_{OR}$, it follows that $\hat{t}_{OR} \xrightarrow{p} t_{OR}$.

(ii) Since $\hat{t}_{OR} \xrightarrow{p} t_{OR}$, the group-wise FDR level yielded by the CLfdr procedure converges in probability to the FDR level of the oracle procedure. Next, let $0 < \lambda < 1$ be a threshold and R_k be the number of rejections in group k. Then $(1/R_k)\sum_{i=1}^{R_k} \text{CLfdr}^k_{(i)} = [\sum_{i=1}^{m_k} I(\text{CLfdr}_i > \lambda)\text{CLfdr}_i]/[\sum_{i=1}^{m_k} I(\text{CLfdr}_i > \lambda)] \xrightarrow{p} E[I(\text{CLfdr}_i < \lambda)\text{CLfdr}_i]/E[I(\text{CLfdr}_i < \lambda)]$. In Section 4.3 we have shown that $(1-p_k)G_{k0}(\lambda) = E[I(\text{CLfdr}_i < \lambda)\text{CLfdr}_i]$ and $G_k(\lambda) = E[I(\text{CLfdr}_i < \lambda)]$. Therefore $(1/R_k)\sum_{i=1}^{R_k} \text{CLfdr}^k_{(i)} \xrightarrow{p} (1-p_k)G_{k0}(\lambda)/G_k(\lambda) = \text{mFDR}^k_{OR} = \text{FDR}^k_{OR} + o(1)$. ∎

Proof of Theorem 5. (i). Define the plug-in statistic $\hat{T}_i = \widehat{\text{CLfdr}}(x_i)$. Let $\hat{Q}_{PI}(t) = \{\sum_i I(\hat{T}_i < t)T_i\}/\{\sum_i I(\hat{T}_i < t)\}$. The threshold of the data-driven procedure can be defined as $\hat{t}_{PI} = \sup\{t \in (0,1) : \hat{Q}_{PI}(t) \leq \alpha\}$. Lemma A.1 in Sun and Cai (2007) implies that $\hat{T}_i \xrightarrow{p} T_i$. We only need to show that $\hat{t}_{PI} \xrightarrow{p} t_{OR}$. Similar to Lemma A.4 of Sun and Cai (2007), it can be shown that $\hat{Q}_{PI}(t) \xrightarrow{p} Q_{OR}(t)$. Then by using the same constructions and arguments as in Theorem 6, we can obtain that $\hat{t}_{PI} \xrightarrow{p} t_{OR}$. (ii) Note that $\hat{T}_i \xrightarrow{p} T_i$ and $\hat{t}_{PI} \xrightarrow{p} t_{OR}$, the proof follows similar lines to the part (ii) of Theorem 4. ∎

References

[1] 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, 289-300.

[2] Considine, G., and Zappalà, G. (2002), "The Influence of Social and Economic Disadvantage in the Academic Performance of School Students in Australia", *Journal of Sociology*, 38, 129-148.

[3] Copas, J. (1974), "On Symmetric Compound Decision Rules for Dichotomies", *The Annals of statististics,* 2, 199-204.

[4] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association,* 96, 1151-1160.

[5] Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: the Choice of a Null Hypothesis," *Journal of the American Statistical Association,* 99, 96-104.

[6] Efron, B. (2007), "Doing Thousands of Hypothesis Tests at the Same Time", *METRON Internal Journal of Statistics*, 65, 3-21.

[7] Efron, B. (2008a), "Simultaneous inference: When Should Hypothesis Testing Problems Be Combined?", *The Annals of Applied Statistics*, 1, 197-223.

[8] Efron, B. (2008b), "Microarrays, Empirical Bayes, and the Two-Groups Model", *Statistical Science*, 23, 1-22.

[9] Ferkinstad, E., Frigessi, A., Thorleifsson, G. and Kong, A. (2008), "Unsupervised Empirical Bayesian Multiple Testing with External Covariates", *The Annals of Applied Statistics*, 2, 714-735.

[10] Genovese, C., and Wasserman, L. (2002), "Operating Characteristic and Extensions of the False Discovery Rate Procedure, " *Journal of the Royal Statistical Society ,* Series B, 64, 499-517.

[11] Genovese, C., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics,* 32, 1035-1061.

[12] Genovese, C., Roeder, K., Wasserman, L. (2006), "False Discovery Control with p value Weighting. " *Biometrika*, 93 (3), 509-524.

[13] Jin, J. and Cai, T. (2007), "Estimating the Null and the Proportion of Non-Null Effects in Large-scale Multiple Comparisons," *Journal of the American Statistical Association*, 102, 495-506.

[14] Morris, B. (2008), "Discussion of Efron's 'Microarrays, Empirical Bayes, and the Two-Groups Model'" *Statistical Science*, 23, 34-40.

[15] Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric hierachical mixture method," *Biostatstics*, 5, 155-176.

[16] Robbins, H. (1951), "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," in *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, CA: University of California Press, 131-148.

[17] Rogasa, D. (2003), "Accuracy of API index and school base report elements: 2003 Academic Performance Index, California Department of Education", Technical Report, Department of Statistics and School of Education, Stanford University, available at "http://www.cde.ca.gov/ta/ac/ap/researchreports.asp".

[18] Sparkes, J. (1999), "Schools, Education and Social Exclusion," CASE Paper 29, Centre for Analysis of Social Exclusion, London School of Economics, London.

[19] Spjøtvoll, E. (1972), "On the Optimality of Some Multiple Comparison Procedures," *The Annals of Mathematical Statistics,* 43, 398-411.

[20] Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479-498.

[21] Storey, J. (2007), "The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing," *Journal of Royal Statistical Society*, Series B, 69, 347-368.

[22] Sun, W., and Cai, T. (2007), " The Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control." *Journal of the American Statistical Association*, 102, 901-912.

[23] Sun, W., and Cai, T. (2008), "Large-Scale Multiple Testing under Dependency", *Journal of the Royal Statistical Society*, Seires B, 71, 393-424.

[24] Schwartzman, A., Dougherty R., and Taylor, J. (2008), "False Discovery Rate Analysis of Brain Diffusion Direction Maps", *The Annals of Applied Statistics,* 2, 153-175.

[25] van der Laan, M., Dudoit, S. and Pollard, K. (2004). "Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives", *Statistical Applications in Genetics and Molecular Biology*, 3, Article 15.

Appendix II: Proof of Other Results

Proof of Proposition 1 (i). Define G_0 and G_1 as before. Let $G = (1-p)G_0 + pG_1$. It follows that $\text{mFDR}(t) = (1-p)G_0(t)/G(t)$ and $\text{mFNR}(t) = (1-p)\tilde{G}_0(t)/\tilde{G}(t)$. Note that the MRC implies that

$$\frac{G_0(t)}{G_1(t)} = \frac{\int_0^t \bar{g}_0(s)ds}{\int_0^t \bar{g}_1(s)ds} = \frac{\int_0^t \bar{g}_0(s)ds}{\int_0^t \{\bar{g}_1(s)/\bar{g}_0(s)\}g_0(s)ds} < \frac{\int_0^t \bar{g}_0(s)ds}{\int_0^t \{\bar{g}_1(t)/\bar{g}_0(t)\}g_0(s)ds} = \frac{\bar{g}_0(t)}{\bar{g}_1(t)}.$$

Hence $\bar{g}_0 G_1 > \bar{g}_1 G_0$. Likewise, $\bar{g}_0\tilde{G}_1 < \bar{g}_1\tilde{G}_0$. The result follows by taking derivatives:

$$\text{mFDR}'(t) = \frac{p(1-p)(\bar{g}_0 G_1 - \bar{g}_1 G_0)}{G^2(t)} > 0, \text{ and } \text{mFNR}'(t) = \frac{p(1-p)(\bar{g}_0\tilde{G}_1 - \bar{g}_1\tilde{G}_0)}{\tilde{G}^2(t)} < 0.$$

(ii). For classification rule $\boldsymbol{\delta} = \{I(T_{ki} < c)\}$, we have $E[(1-\theta_{ki})\delta_{ki}] = E[(1-\theta_{ki})E\{\delta_{ki}|\theta_{ki}\}] =$

$(1-p_k)G_{k0}(c)$. Similarly, $E[\theta_{ki}(1-\delta_{ki})] = p_k\tilde{G}_{k1}(c)$. Then the classification risk is

$$\begin{aligned} R_\lambda &= \frac{1}{m}E\left\{\sum_{k=1}^{K}\sum_{i=1}^{m_k}\lambda(1-\theta_{ki})\delta_{ki}+\theta_{ki}(1-\delta_{ki})\right\} \\ &= \lambda\sum_k \pi_k(1-p_k)G_{k0}(c)+\sum_k \pi_k p_k\tilde{G}_{k1}(c) \\ &= \lambda(1-p)G_0(c)+pG_1(c), \end{aligned}$$

where $G_0 = \{1/(1-p)\}\sum_k \pi_k(1-p_k)G_{k0}$ and $G_1 = (1/p)\sum_k \pi_k p_k G_{k1}$. The cutoff $c(\lambda)$ for $\boldsymbol{T}$ that minimizes R_λ satisfies $\lambda(1-p)\bar{g}_0(c) = p\bar{g}_1(c)$. Suppose $\lambda_1 < \lambda_2$, and c_i solves the previous equation when λ_i is chosen, $i = 1, 2$. It is sufficient to show that $c_1 > c_2$. This must be true since otherwise we have $c_1 \le c_2$ and hence $\lambda_1 = \{p/(1-p)\}\{\bar{g}_1(c_1)/\bar{g}_0(c_1)\} \ge \{p/(1-p)\}\{\bar{g}_1(c_2)/\bar{g}_0(c_2)\} = \lambda_2$, which contradicts $\lambda_1 < \lambda_2$. ∎

Proof of Theorem 6. Proposition 2 implies that for any $\boldsymbol{T} \in \mathcal{T}$ and a given α, there exists a unique $t(\alpha)$ such that the mFDR level of $\boldsymbol{\delta}(\boldsymbol{T}, t(\alpha)\mathbf{1})$ is α. Let $r(\alpha)$ be the expected number of rejections of procedure $\boldsymbol{\delta}(\boldsymbol{T}, t\mathbf{1})$. Now consider the optimal classification statistic $\boldsymbol{\Gamma}$. Proposition 1 implies that the optimal cutoff γ for $\boldsymbol{\Gamma}$, and hence the expected number of rejections r, is decreasing in λ. Therefore for a given $r(\alpha)$, there exists a unique $\lambda(\alpha)$ that defines a weighted classification problem whose classification risk is minimized by $\boldsymbol{\delta}\{\boldsymbol{\Gamma}, \gamma(\alpha)\mathbf{1}\}$.

Suppose that $\boldsymbol{\delta}\{\boldsymbol{\Gamma}, \gamma(\alpha)\mathbf{1}\}$ is used in the multiple testing problem. Let $v_{\boldsymbol{\Gamma}}$ and $u_{\boldsymbol{\Gamma}}$ ($v_{\boldsymbol{T}}$ and $u_{\boldsymbol{T}}$) be the expected number of false and true positives of $\boldsymbol{\delta}\{\boldsymbol{\Gamma}, \gamma(\alpha)\mathbf{1}\}$ ($\boldsymbol{\delta}(\boldsymbol{T}, t(\alpha)\mathbf{1})$). Then for the weighted classification problem with weight $\lambda(\alpha)$, the classification risks of $\boldsymbol{\delta}\{\boldsymbol{\Gamma}, \gamma(\alpha)\mathbf{1}\}$ and $\boldsymbol{\delta}(\boldsymbol{T}, t(\alpha)\mathbf{1})$ are $R_{\boldsymbol{\Gamma}} = p + (1/m)\{\lambda(\alpha)v_{\boldsymbol{\Gamma}} - u_{\boldsymbol{\Gamma}}\}$ and $R_{\boldsymbol{T}} = p + (1/m)\{\lambda(\alpha)v_{\boldsymbol{T}} - u_{\boldsymbol{T}}\}$, respectively. Note that by construction, we have $r(\alpha) = v_{\boldsymbol{\Gamma}} + u_{\boldsymbol{\Gamma}} = v_{\boldsymbol{T}} + u_{\boldsymbol{T}}$; hence $v_{\boldsymbol{\Gamma}} \le v_{\boldsymbol{T}}$ and $u_{\boldsymbol{\Gamma}} \ge u_{\boldsymbol{T}}$. Therefore $\text{mFDR}_{\boldsymbol{\Gamma}} = v_{\boldsymbol{\Gamma}}/r \le v_{\boldsymbol{T}}/r = \alpha$ and $\text{mFNR}_{\boldsymbol{\Gamma}} = (m_1 - u_{\boldsymbol{\Gamma}})/(m-r) \le (m_1 - u_{\boldsymbol{T}})/(m-r) = \text{mFNR}_{\boldsymbol{T}}$. Since $\boldsymbol{T}$ can be any test statistic satisfying the MRC, we have shown the optimality of $\boldsymbol{\Gamma}$ in the multiple testing problem. ∎

Proof of Theorem 7. The posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and $\boldsymbol{g}$ is

$$P(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{g}) = \prod_{k=1}^{K}\prod_{i=1}^{m_k} P_{\theta_{ki}|X_{ki}}(\theta_{ki}|x_{ki}),$$

where $P_{\theta_{ki}|X_{ki}}(\theta_{ki}|x_{ki}) = \{(1-\theta_{ki})(1-p_k)f_{k0}(x_{ki})+\theta_{ki}p_k f_{k1}(x_{ki})\}/f_k(x_{ki})$. Hence the posterior risk is

$$\begin{aligned} E_{\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{g}}L(\boldsymbol{\theta},\boldsymbol{\delta}) &= \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m_k} E_{\theta_{ki}|X_{ki}}\{\lambda(1-\theta_{ki})\delta_{ki} + \theta_{ki}(1-\delta_{ki})\} \\ &= \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m_k}\frac{p_k f_{k1}(x_{ki})}{f_k(x_{ki})} + \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m_k}\frac{(1-p_k)f_{k0}(x_{ki}) - p_k f_{k1}(x_{ki})}{f_k(x_{ki})}\delta_{ki} \end{aligned}$$

It is easy to see that the classification risk is minimized by

$$\boldsymbol{\delta}\{\boldsymbol{\Lambda}, c(\lambda)\mathbf{1}\} = (\delta_{ki}) = \left[I\left\{\frac{(1-p_k)f_{k0}(x_{ki})}{p_k f_{k1}(x_{ki})} < \frac{1}{\lambda}\right\} : k = 1,\cdots,K; i = 1,\cdots,m_k\right]. \quad \blacksquare$$