6-26-2017

Weighted False Discovery Rate Control in Large-Scale Multiple Testing

Pallavi Basu
University of Southern California

Tony Cai
University of Pennsylvania

Kiranmoy Das
Indian Statistical Institute

Wenguang Sun
University of Southern California

Recommended Citation

Basu, P., Cai, T., Das, K., & Sun, W. (2017). Weighted False Discovery Rate Control in Large-Scale Multiple Testing. *Journal of the American Statistical Association,* http://dx.doi.org/10.1080/01621459.2017.1336443

Weighted False Discovery Rate Control in Large-Scale Multiple Testing

Abstract
The use of weights provides an effective strategy to incorporate prior domain knowledge in large-scale inference. This paper studies weighted multiple testing in a decisiontheoretic framework. We develop oracle and data-driven procedures that aim to maximize the expected number of true positives subject to a constraint on the weighted false discovery rate. The asymptotic validity and optimality of the proposed methods are established. The results demonstrate that incorporating informative domain knowledge enhances the interpretability of results and precision of inference. Simulation studies show that the proposed method controls the error rate at the nominal level, and the gain in power over existing methods is substantial in many settings. An application to genome-wide association study is discussed.

Keywords
Class weights, Decision weights, Multiple testing with groups, Prioritized subsets, Value to cost ratio, Weighted p-value

Disciplines
Business | Statistics and Probability

Weighted False Discovery Rate Control in Large-Scale Multiple Testing

Pallavi Basu[1], T. Tony Cai[2], Kiranmoy Das[3], and Wenguang Sun[1]

August 5, 2015

Abstract

The use of weights provides an effective strategy to incorporate prior domain knowledge in large-scale inference. This paper studies weighted multiple testing in a decision-theoretic framework. We develop oracle and data-driven procedures that aim to maximize the expected number of true positives subject to a constraint on the weighted false discovery rate. The asymptotic validity and optimality of the proposed methods are established. The results demonstrate that incorporating informative domain knowledge enhances the interpretability of results and precision of inference. Simulation studies show that the proposed method controls the error rate at the nominal level, and the gain in power over existing methods is substantial in many settings. An application to genome-wide association study is discussed.

Keywords: Class weights; Decision weights; Multiple testing with groups; Prioritized subsets; Value to cost ratio; Weighted p-value.

[1]Department of Data Sciences and Operations, University of Southern California. The research of Wenguang Sun was supported in part by NSF grant DMS-CAREER 1255406.

[2]Department of Statistics, The Wharton School, University of Pennsylvania. The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

[3]Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata.

1 Introduction

In large-scale studies, relevant domain knowledge, such as external covariates, scientific insights and prior data, is often available alongside the primary data set. Exploiting such information in an efficient manner promises to enhance both the interpretability of scientific results and precision of statistical inference. In multiple testing, the hypotheses being investigated often become "unequal" in light of external information, which may be reflected by differential attitudes towards the relative importance of testing units or the severity of decision errors. The use of weights provides an effective strategy to incorporate informative domain knowledge in large-scale testing problems.

In the literature, various weighting methods have been advocated for a range of multiple comparison problems. A popular scheme, referred to as the *decision weights* or loss weights approach, involves modifying the error criteria or power functions in the decision process (Benjamini and Hochberg, 1997). The idea is to employ two sets of positive constants $\boldsymbol{a} = \{a_i : i = 1, \cdots, m\}$ and $\boldsymbol{b} = \{b_i : i = 1, \cdots, m\}$ to take into account the costs and gains of multiple decisions. Typically, the choice of the weights $\boldsymbol{a}$ and $\boldsymbol{b}$ reflects the degree of confidence one has toward prior beliefs and external information. It may also be pertinent to the degree of preference that one has toward the consequence of one class of erroneous/correct decisions over another class based on various economical and ethical considerations. For example, in the spatial cluster analysis considered by Benjamini and Heller (2007), the weighted false discovery rate was used to reflect that a false positive cluster with larger size would account for a larger error. Another example arises from genome-wide association studies (GWAS), where prior data or genomic knowledge, such as prioritized subsets (Lin and Lee, 2012), allele frequencies (Lin et al., 2014) and expression quantitative trait loci information (Li et al., 2013), can often help to assess the scientific plausibility of significant associations. To incorporate such information in the analysis, a useful strategy is to up-weight the gains for the discoveries in preselected genomic regions

by modifying the power functions in respective testing units (Pẽna et al., 2011; Sun et al., 2015). We assume in this paper that the weights have been pre-specified by the investigator. This is a reasonable assumption in many practical settings. For example, weights may be assigned according to economical considerations (Westfall and Young, 1993), external covariates (Benjamini and Heller, 2007; Sun et al., 2015) and biological insights from prior studies (Xing et al., 2010).

We mention two alternative formulations for weighted multiple testing. One popular method, referred to as the *procedural weights* approach by Benjamini and Hochberg (1997), involves the adjustment of the p-values from individual tests. In GWAS, Roeder et al. (2006) and Roeder and Wasserman (2009) proposed to utilize linkage signals to up-weight the p-values in preselected regions and down-weight the p-values in other regions. It was shown that the power to detect association can be greatly enhanced if the linkage signals are informative, yet the loss in power is small when the linkage signals are uninformative. Another useful weighting scheme, referred to as the *class weights* approach, involves allocating varied test levels to different classes of hypotheses. For example, in analysis of the growth curve data (Box, 1950), Westfall and Young (1993, page 186) proposed to allocate a higher family-wise error rate (FWER) to the class of hypotheses related to the primary variable "gain" and a lower FWER to the secondary variable "shape".

We focus on the decision weights approach in the present paper. This weighting scheme is not only practically useful for a wide range of applications, but also provides a powerful framework that enables a unified investigation of various weighting methods. Specifically, the proposal in Benjamini and Hochberg (1997) involves the modification of both the error rate and power function. The formulation is closely connected to classical ideas in compound decision theory that aim to optimize the tradeoffs between the gains and losses when many simultaneous decisions are combined as a whole. Our theory reveals that if the goal is to maximize the power subject to a given error rate, then the modifications via decision

weights would lead to improved multiple testing methods with sensible procedural weights or class weights, or both. For example, in GWAS, the investigators can up-weight the power functions for discoveries in genomic regions that are considered to be more scientific plausible or biologically meaningful; this would naturally up-weight the p-values in these regions and thus yield weighting strategies similar to those suggested by Roeder and Wasserman (2009). In large clinical trials, modifying the power functions for respective rejections at the primary and secondary end points would correspond to the allocation of varied test levels across different classes of hypotheses, leading to weighting strategies previously suggested by Westfall and Young (1993).

The false discovery rate (FDR; Benjamini and Hochberg, 1995) has been widely used in large-scale multiple testing as a powerful error criterion. Following Benjamini and Hochberg (1997), we generalize the FDR to weighted false discovery rate (wFDR), and develop optimal procedures for wFDR control under the decision weights framework. We first construct an oracle procedure that maximizes the weighted power function subject to a constraint on the wFDR, and then develop a data-driven procedure to mimic the oracle and establish its asymptotic optimality. The numerical results show that the proposed method controls the wFDR at the nominal level, and the gain in power over existing methods is substantial in many settings. Our optimality result in the decision weights framework marks a clear departure from existing works in the literature that are mainly focused on the derivation of optimal procedural weights subject to the conventional FDR criterion and unweighted power function (Roeder and Wasserman, 2009; Roquain and van de Wiel, 2009).

Our research also makes a novel contribution to the theory of optimal ranking in multiple testing. Conventionally, a multiple testing procedure operates in two steps: ranking the hypotheses according to their significance levels and then choosing a cutoff along the rankings. It is commonly believed that the rankings remain the same universally at all FDR levels. For example, the ranking based on p-values or adjusted p-values in common

practice is invariant to the choice of the FDR threshold. The implication of our theory is interesting, for it claims that there does not exist a ranking that is universally optimal at all test levels. Instead, the optimal ranking of hypotheses depends on the pre-specified wFDR level. That is, the hypotheses may be ordered differently when different wFDR levels are chosen. This point is elaborated in Section 3.3.

The rest of the article is organized as follows. Section 2 discusses a general framework for weighted multiple testing. Sections 3 and 4 develop oracle and data-driven wFDR procedures and establish their optimality properties. Simulation studies are conducted in Section 5 to investigate the numerical performance of the proposed methods. An application to GWAS is presented in Section 6. Section 7 concludes the article with a discussion of related and future works. Proofs of the technical results are given in the Appendix.

2 Problem Formulation

This section discusses a decision weights framework for weighted multiple testing. We first introduce model and notation and then discuss modified error criteria and power functions.

2.1 Model and notation

Suppose that m hypotheses $H_1, \cdots, H_m$ are tested simultaneously based on observations $X_1, \cdots, X_m$. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m) \in \{0,1\}^m$ denote the true state of nature, where 0/1 indicates a null/non-null case. Assume that observations X_i are independent and distributed according to the following model

$$X_i|\theta_i \sim (1-\theta_i)F_{0i} + \theta_i F_{1i}, \tag{2.1}$$

where F_{0i} and F_{1i} are the null and non-null distributions for X_i, respectively. Denote by f_{0i} and f_{1i} the corresponding density functions. Suppose that the unknown states θ_i are Bernoulli (p_i) variables, where $p_i = P(\theta_i = 1)$. The mixture density is denoted by $f_{\cdot i} = (1 - p_i)f_{0i} + p_i f_{1i}$.

Consider the widely used random mixture model (Efron et al., 2001; Storey, 2002; Genovese and Wasserman, 2002)

$$X_i \sim F = (1 - p)F_0 + pF_1. \tag{2.2}$$

This model, which assumes that all observations are identically distributed according to a common distribution F, can sometimes be unrealistic in applications. In light of domain knowledge, the observations are likely to have different distributions. For example, in the context of a brain imaging study, Efron (2008) showed that the proportions of activated voxels are different for the front and back halves of a brain. In GWAS, certain genomic regions contain higher proportions of significant signals than other regions. In the adequate yearly progress study of California high schools (Rogasa, 2003), the densities of z-scores vary significantly from small to large schools. We develop theories and methodologies for model (2.1) for it considers different non-null proportions and densities; this allows the proposed method to be applied to a wider range of situations.

The multi-group model considered in Efron (2008) and Cai and Sun (2009), which has been widely used in applications, is an important case of the general model (2.1). The multi-group model assumes that the observations can be divided into K groups. Let $\mathcal{G}_k$ denote the index set of the observations in group k, $k = 1, \cdots, K$. For each $i \in \mathcal{G}_k$, θ_i is distributed as Bernoulli(p_k), and X_i follows a mixture distribution:

$$(X_i | i \in \mathcal{G}_k) \sim f_{\cdot k} = (1 - p_k)f_{0k} + p_k f_{1k}, \tag{2.3}$$

where f_{0k} and f_{1k} are the null and non-null densities for observations in group k. This model will be revisited in later sections. See also Ferkingstad et al. (2008) and Hu et al. (2010) for related works on multiple testing with groups.

2.2 Weighted error criterion and power function

This section discusses a generalization of the FDR criterion in the context of weighted multiple testing. Denote the decisions for the m tests by $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m) \in \{0,1\}^m$, where $\delta_i = 1$ indicates that H_i is rejected and $\delta_i = 0$ otherwise. The weighted false discovery rate (wFDR) is defined as

$$\text{wFDR} = \frac{E\left\{\sum_{i=1}^{m} a_i(1-\theta_i)\delta_i\right\}}{E\left(\sum_{i=1}^{m} a_i\delta_i\right)}, \tag{2.4}$$

where a_i is the weight indicating the severity of a false positive decision. For example, a_i is taken as the cluster size in the spatial cluster analyses conducted in Benjamini and Heller (2007) and Sun et al. (2015). As a result, rejecting a larger cluster erroneously corresponds to a more severe decision error.

Remark 1 Our definition of the wFDR is slightly different from that considered in Benjamini and Hochberg (1997), which defines the wFDR as the expectation of a ratio. The consideration of using a ratio of two expectations (or a marginal version of the wFDR) is only to facilitate our theoretical derivations. Genovese and Wasserman (2002) showed that, in large-scale testing problems, the difference between the marginal FDR (mFDR) and FDR is negligible under mild conditions. The asymptotic equivalence in the weighted case can be established similarly.

To compare the effectiveness of different weighted multiple testing procedures, we define the expected number of true positives

$$\text{ETP} = E\left(\sum_{i=1}^{m} b_i \theta_i \delta_i\right), \tag{2.5}$$

where b_i is the weight indicating the power gain when H_i is rejected correctly. The use of b_i provides a useful scheme to incorporate informative domain knowledge. In GWAS, larger b_i can be assigned to pre-selected genomic regions to reflect that the discoveries in these regions are more biologically meaningful. In spatial data analysis, correctly identifying a larger cluster that contains signal may correspond to a larger b_i, indicating a greater decision gain.

By combining the concerns on both the error criterion and power function, the goal in weighted multiple testing is to

$$\textbf{maximize the ETP subject to the constraint wFDR} \leq \alpha. \tag{2.6}$$

The optimal solution to (2.6) is studied in the next section.

3 Oracle Procedure for wFDR Control

The basic framework of our theoretical and methodological developments is outlined as follows. In Section 3.1, we assume that p_i, f_{0i}, and $f_{\cdot i}$ in the mixture model (2.1) are known by an oracle and derive an oracle procedure that maximizes the ETP subject to a constraint on the wFDR. Connections to the literature and a discussion on optimal ranking are included in Sections 3.2 and 3.3. In Section 4, we develop a data-driven procedure to mimic the oracle and establish its asymptotic validity and optimality.

3.1 Oracle procedure

The derivation of the oracle procedure involves two key steps: the first is to derive the optimal ranking of hypotheses and the second is to determine the optimal threshold along the ranking that exhausts the pre-specified wFDR level. We discuss the two issues in turn.

Consider model (2.1). Define the local false discovery rate (Lfdr, Efron et al. 2001) as

$$\text{Lfdr}_i = \frac{(1-p_i)f_{0i}(x_i)}{f_{\cdot i}(x_i)}. \tag{3.1}$$

The wFDR problem (2.6) is equivalent to the following constrained optimization problem

$$\text{maximize } E\left\{\sum_{i=1}^{m} b_i\delta_i(1-\text{Lfdr}_i)\right\} \text{ subject to } E\left\{\sum_{i=1}^{m} a_i\delta_i(\text{Lfdr}_i-\alpha)\right\} \le 0. \tag{3.2}$$

Let $S^- = \{i : \text{Lfdr}_i \le \alpha\}$ and $S^+ = \{i : \text{Lfdr}_i > \alpha\}$. Then the constraint in (3.2) can be equivalently expressed as

$$E\left\{\sum_{S^+} a_i\delta_i(\text{Lfdr}_i-\alpha)\right\} \le E\left\{\sum_{S^-} a_i\delta_i(\alpha-\text{Lfdr}_i)\right\}. \tag{3.3}$$

Consider an optimization problem which involves packing a knapsack with a capacity given by the right hand side of equation (3.3). Every available object has a known value and a known cost (of space). Clearly rejecting a hypothesis in S^- is always beneficial as it allows the capacity to expand, and thus promotes more discoveries. The key issue is how to efficiently utilize the capacity (after all hypotheses in S^- are rejected) to make as many discoveries as possible in S^+. Each rejection in S^+ would simultaneously increase the power and decrease the capacity. We propose to sort all hypotheses in S^+ in an decreasing order of the value to cost ratio (VCR). Equations (3.2) and (3.3) suggest that

$$\text{VCR}_i = \frac{b_i(1-\text{Lfdr}_i)}{a_i(\text{Lfdr}_i-\alpha)}. \tag{3.4}$$

To maximize the power, the ordered hypotheses are rejected sequentially until maximum capacity is reached.

The above considerations motivate us to consider the following class of decision rules $\boldsymbol{\delta}^*(t) = \{\delta_i^*(t) : i = 1, \cdots, m\}$, where

$$\delta_i^*(t) = \begin{cases} 1, & \text{if } b_i(1 - \text{Lfdr}_i) > t a_i(\text{Lfdr}_i - \alpha), \\ 0, & \text{if } b_i(1 - \text{Lfdr}_i) \leq t a_i(\text{Lfdr}_i - \alpha). \end{cases} \tag{3.5}$$

We briefly explain some important operational characteristics of testing rule (3.5). First, if we let $t > 0$, then the equation implies that $\delta_i^*(t) = 1$ for all $i \in S^-$; hence all hypotheses in S^- are rejected as desired. (This explains why the VCR is not used directly in (3.5), given that the VCR is not meaningful in S^-.) Second, a solution path can be generated as we vary t continuously from large to small. Along the path $\boldsymbol{\delta}^*(t)$ sequentially rejects the hypotheses in S^+ according to their VCRs. Denote by $H_{(1)}, \cdots, H_{(m)}$ the hypotheses sequentially rejected by $\boldsymbol{\delta}^*$. (The actual ordering of the hypotheses within S^- does not matter in the decision process since all are always rejected.)

The next task is to choose a cutoff along the ranking to achieve exact wFDR control. The difficulty is that the maximum capacity may not be attained by a sequential rejection procedure. To exhaust the wFDR level, we permit a randomized decision rule. Denote the Lfdr values and the weights corresponding to $H_{(i)}$ by $\text{Lfdr}_{(i)}$, $a_{(i)}$, and $b_{(i)}$. Let

$$C(j) = \sum_{i=1}^{j} a_{(i)}(\text{Lfdr}_{(i)} - \alpha) \tag{3.6}$$

denote the capacity up to jth rejection. According to the constraint in equation (3.2), we choose $k = \max\{j : C(j) \leq 0\}$ so that the capacity is not yet reached when $H_{(k)}$ is rejected but would be exceeded if $H_{(k+1)}$ is rejected. The idea is to split the decision point at $H_{(k+1)}$ by randomization.

Let U be a Uniform $(0,1)$ variable that is independent of the truth, the observations, and the weights. Define

$$t^* = \frac{b_{(k+1)}\left(1 - \mathrm{Lfdr}_{(k+1)}\right)}{a_{(k+1)}\left(\mathrm{Lfdr}_{(k+1)} - \alpha\right)} \quad \text{and} \quad p^* = -\frac{C(k)}{C(k+1) - C(k)}.$$

Let $\mathcal{I}_A$ be an indicator, which takes value 1 if event A occurs and 0 otherwise. We propose the *oracle decision rule* $\boldsymbol{\delta}_{OR} = \{\delta^i_{OR} : i = 1, \cdots, m\}$, where

$$\delta^i_{OR} = \begin{cases} 1 & \text{if } b_i(1 - \mathrm{Lfdr}_i) > t^* a_i(\mathrm{Lfdr}_i - \alpha), \\ 0 & \text{if } b_i(1 - \mathrm{Lfdr}_i) < t^* a_i(\mathrm{Lfdr}_i - \alpha), \\ \mathcal{I}_{U<p^*} & \text{if } b_i(1 - \mathrm{Lfdr}_i) = t^* a_i(\mathrm{Lfdr}_i - \alpha). \end{cases} \tag{3.7}$$

Remark 2 The randomization step is only employed for theoretical considerations to enforce the wFDR to be exactly α. Thus the optimal power can be effectively characterized. Moreover, only a single decision point at $H_{(k+1)}$ is randomized, which has a negligible effect in large-scale testing problems. We do not pursue randomized rules for the data-driven procedures developed in later sections.

Let $\mathrm{wFDR}(\boldsymbol{\delta})$ and $\mathrm{ETP}(\boldsymbol{\delta})$ denote the wFDR and ETP of a decision rule $\boldsymbol{\delta}$, respectively. Theorem 1 shows that the oracle procedure (3.7) is valid and optimal for wFDR control.

Theorem 1 *Consider model (2.1) and oracle procedure $\boldsymbol{\delta}_{OR}$ defined in (3.7). Let $\mathcal{D}_\alpha$ be the collection of decision rules such that for any $\boldsymbol{\delta} \in \mathcal{D}_\alpha$, $wFDR(\boldsymbol{\delta}) \le \alpha$. Then we have*

(i). *$wFDR(\boldsymbol{\delta}_{OR}) = \alpha$.*

(ii). *$ETP(\boldsymbol{\delta}_{OR}) \ge ETP(\boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \mathcal{D}_\alpha$.*

3.2 Comparison with the optimality results in Spjøtvoll (1972) and Benjamini and Hochberg (1997)

Spjøtvoll (1972) showed that the likelihood ratio (LR) statistic

$$T_{LR}^{i} = \frac{f_{0i}(x_i)}{f_{1i}(x_i)} \tag{3.8}$$

is optimal for the following multiple testing problem

$$\text{maximize } E_{\cap H_{1i}}\left(\sum_{i=1}^{m}\delta_i\right) \text{ subject to } E_{\cap H_{0i}}\left\{\sum_{i=1}^{m}\delta_i\right\} \leq \alpha, \tag{3.9}$$

where $\cap H_{0i}$ and $\cap H_{1i}$ denote the intersections of the nulls and non-nulls, respectively. The error criterion $E_{\cap H_{0i}}\{\sum_i a_i\delta_i\}$ is referred to as the intersection tests error rate (ITER). A weighted version of problem (3.9) was considered by Benjamini and Hochberg (1997), where the goal is to

$$\text{maximize } E_{\cap H_{1i}}\left(\sum_{i=1}^{m}b_i\delta_i\right) \text{ subject to } E_{\cap H_{0i}}\left\{\sum_{i=1}^{m}a_i\delta_i\right\} \leq \alpha. \tag{3.10}$$

The optimal solution to (3.10) is given by the next proposition.

Proposition 1 *(Benjamini and Hochberg, 1997). Define the weighted likelihood ratio (WLR)*

$$T_{IT}^{i} = \frac{a_i f_{0i}(x_i)}{b_i f_{1i}(x_i)}. \tag{3.11}$$

Then the optimal solution to (3.10) is a thresholding rule of the form $\delta_{IT}^{i} = (T_{IT}^{i} < t_\alpha)$, where t_α is the largest threshold that controls the weighted ITER at level α.

The ITER is very restrictive in the sense that the expectation is taken under the conjunction of the null hypotheses. The ITER is inappropriate for mixture model (2.1) where a mixture of null and non-null hypotheses are tested simultaneously. To extend intersection

tests to multiple tests, define the per family error rate (PFER) as

$$\text{PFER}(\boldsymbol{\delta}) = E\left\{\sum_{i=1}^{m} a_i(1-\theta_i)\delta_i\right\}. \tag{3.12}$$

The power function should be modified correspondingly. Therefore the goal is to

$$\text{maximize } E\left(\sum_{i=1}^{m} b_i\theta_i\delta_i\right) \text{ subject to } E\left\{\sum_{i=1}^{m} a_i(1-\theta_i)\delta_i\right\} \leq \alpha. \tag{3.13}$$

The key difference between the ITER and PFER is that the expectation in (3.12) is now taken over all possible combinations of the null and non-null hypotheses. The optimal PFER procedure is given by the next proposition.

Proposition 2 *Consider model (2.1) and assume continuity of the LR statistic. Let* $\mathcal{D}_{PF}^{\alpha}$ *be the collection of decision rules such that for every* $\boldsymbol{\delta} \in \mathcal{D}_{PF}^{\alpha}$, *PFER*$(\boldsymbol{\delta}) \leq \alpha$. *Define the weighted posterior odds (WPO)*

$$T_{PF}^{i} = \frac{a_i(1-p_i)f_{0i}(x_i)}{b_i p_i f_{1i}(x_i)}. \tag{3.14}$$

Denote by $Q_{PF}(t)$ *the PFER of* $\delta_{PF}^{i} = I(T_{PF}^{i} < t)$. *Then the oracle PFER procedure is* $\boldsymbol{\delta}_{PF} = (\delta_{PF}^{i} : i = 1, \cdots, m)$, *where* $\delta_{PF}^{i} = I(T_{PF}^{i} < t_{PF})$ *and* $t_{PF} = \sup\{t : Q_{PF}(t) \leq \alpha\}$. *This oracle rule satisfies:*

(i). *ETP*$(\boldsymbol{\delta}_{PF}) = \alpha$.

(ii). *ETP*$(\boldsymbol{\delta}_{PF}) \geq$ *ETP*$(\boldsymbol{\delta})$ *for all* $\boldsymbol{\delta} \in \mathcal{D}_{PF}^{\alpha}$.

Our formulation (2.6) modifies the conventional formulations in (3.10) and (3.13) to the multiple testing situation with an FDR type criterion. These modifications lead to methods that are more suitable for large-scale scientific studies. The oracle procedure (3.7) uses the VCR (3.4) to rank the hypotheses. The VCR, which optimally combines the

decision weights, significance measure (Lfdr) and test level α, produces a more powerful ranking than the WPO (3.14) in the wFDR problem; this is explained in detail next.

3.3 Optimal ranking: VCR vs. WPO

Although the WPO is optimal for PFER control, it is suboptimal for wFDR control. This section discusses a toy example to provide some insights on why the WPO ranking is dominated by the VCR ranking. We simulate 1000 z-values from a mixture model $(1-p)N(0,1)+pN(2,1)$ with $p = 0.2$. The weights a_i are fixed at 1 for all i, and b_i are generated from log-normal distribution with location parameter $\ln 3$ and scale parameter 1. At wFDR level $\alpha = 0.10$, we can reject 68 hypotheses along the WPO ranking, with the number of true positives being 60; in contrast, we can reject 81 hypotheses along the VCR ranking, with the number of true positives being 73. This shows that the VCR ranking enables us to "pack more objects" under the capacity wFDR = 0.1 compared to the WPO ranking. Detailed simulation results are presented in Section 5.

Next we give some intuitions on why the VCR ranking is more efficient in the wFDR problem. The test level α, which can be viewed as the initial capacity for the error rate, plays an important role in the ranking process. Under the wFDR criterion, the capacity may either increase or decrease when a new rejection is made; the quantity that affects the current capacity is the *excessive error rate* $(\text{Lfdr}_i - \alpha)$. A different α would yield a different excessive error rate and hence a different ranking. (This is very different from the PFER criterion, under which the capacity always decreases when a new rejection is made and α is not useful in ranking.) The next example shows that, although the WPO ranking always remains the same the VCR ranking can be altered by the choice of α.

Example 1 Consider two units A and B with observed values and weights $x_A = 2.73$, $x_B = 3.11$, $b_A = 83.32$, and $b_B = 11.95$. The Lfdr values are $\text{Lfdr}_A = 0.112$ and $\text{Lfdr}_B =$

0.055, ranking B ahead of A. Taking into account of the decision weights, the WPO values are $\text{WPO}_A = 0.0015$ and $\text{WPO}_B = 0.0049$, ranking A ahead of B, and this ranking remains the same at all wFDR levels. At $\alpha = 0.01$, we have $\text{VCR}_A = 725.4$ and $\text{VCR}_B = 250.9$, yielding the same ranking as the WPO. However, at $\alpha = 0.05$, we have $\text{VCR}_A = 1193.5$ and $\text{VCR}_B = 2258.6$, reversing the previous ranking. This reversed ranking is due to the small excessive error rate ($\text{Lfdr}_B - \alpha$) at $\alpha = 0.05$, which makes the rejection of B, rather than A, more "profitable."

4 Data-Driven Procedures and Asymptotics

The oracle procedure (3.7) cannot be implemented in practice since it relies on unknown quantities such as Lfdr_i and t^*. This section develops a data-driven procedure to mimic the oracle. We first propose a test statistic to rank the hypotheses and discuss related estimation issues. A step-wise procedure is then derived to determine the best cutoff along the ranking. Finally, asymptotic results on the validity and optimality of the proposed procedure are presented.

4.1 Proposed test statistic and its estimation

The oracle procedure utilizes the ranking based on the VCR (3.4). However, the VCR is only meaningful for the tests in S^+ and becomes problematic when both S^- and S^+ are considered. Moreover, the VCR could be unbounded, which would lead to difficulties in both numerical implementations and technical derivations. We propose to rank the hypotheses using the following statistic (in increasing values)

$$R_i = \frac{a_i(\text{Lfdr}_i - \alpha)}{b_i(1 - \text{Lfdr}_i) + a_i|\text{Lfdr}_i - \alpha|}. \tag{4.1}$$

As shown in the next proposition, R_i always ranks hypotheses in S^- higher than hypotheses in S^+ (as desired), and yields the same ranking as that by the VCR (3.4) for hypotheses in S^+. The other drawbacks of VCR can also be overcome by R_i: R_i is always bounded in the interval $[-1, 1]$ and is a continuous function of the Lfdr_i.

Proposition 3 *(i) The rankings generated by the decreasing values of VCR (3.4) and increasing values of R_i (4.1) are the same in both S^- and S^+. (ii) The ranking based on increasing values of R_i always puts hypotheses in S^- ahead of hypotheses in S^+.*

Next we discuss how to estimate R_i; this involves the estimation of the Lfdr statistic (3.1), which has been studied extensively in the multiple testing literature. We give a review of related methodologies. If all observations follow a common mixture distribution (2.2), then we can first estimate the non-null proportion p and the null density f_0 using the methods in Jin and Cai (2007), and then estimate the mixture density f using a standard kernel density estimator (e.g. Silverman, 1986). If all observations follow a multi-group model (2.3), then we can apply the above estimation methods to separate groups to obtain corresponding estimates $\hat{p}_k$, $\hat{f}_{0k}$, and $\hat{f}_{\cdot k}$, $k = 1, \cdots, K$. The theoretical properties of these estimators have been established in Sun and Cai (2007) and Cai and Sun (2009). In practice, estimation problems may arise from more complicated models. Related theories and methodologies have been studied in Storey (2007), Ferkingstad et al. (2008), and Efron (2008, 2010); theoretical supports for these estimators are yet to be developed.

The estimated Lfdr value for H_i is denoted by $\widehat{\text{Lfdr}}_i$. By convention, we take $\widehat{\text{Lfdr}}_i = 1$ if $\widehat{\text{Lfdr}}_i > 1$. This modification only facilitates the development of theory and has no practical effect on the testing results (since rejections are essentially only made for small $\widehat{\text{Lfdr}}_i$'s). The ranking statistic R_i can therefore be estimated as

$$\widehat{R}_i = \frac{a_i(\widehat{\text{Lfdr}}_i - \alpha)}{b_i(1 - \widehat{\text{Lfdr}}_i) + a_i|\widehat{\text{Lfdr}}_i - \alpha|}. \tag{4.2}$$

The performance of the data driven procedure relies on the accuracy of the estimate $\widehat{\text{Lfdr}}_i$; some technical conditions are discussed in the next subsection.

4.2 Proposed testing procedure and its asymptotic properties

Consider $\widehat{R}_i$ defined in (4.2). Denote by $\widehat{N}_i = a_i(\widehat{\text{Lfdr}}_i - \alpha)$ the estimate of excessive error rate when H_i is rejected. Let $\widehat{R}_{(1)}, \cdots, \widehat{R}_{(m)}$ be the ordered test statistics (in increasing values). The hypothesis and estimated excessive error rate corresponding to $\widehat{R}_{(i)}$ are denoted by $H_{(i)}$ and $\widehat{N}_{(i)}$. The idea is to choose the largest cutoff along the ranking based on $\widehat{R}_i$ so that the maximum capacity is reached. Motivated by the constraint in (3.2), we propose the following step-wise procedure.

Procedure 1 *(wFDR control with general weights). Rank hypotheses according to $\widehat{R}_i$ in increasing values. Let $k = \max\left\{ j : \sum_{i=1}^{j} \widehat{N}_{(i)} \le 0 \right\}$. Reject $H_{(i)}$, for $i = 1, \ldots, k$.*

It is important to note that in Procedure 1, $\widehat{R}_i$ is used in the ranking step whereas $\widehat{N}_i$ (or a weighted transformation of $\widehat{\text{Lfdr}}_i$) is used in the thresholding step. The ranking by $\widehat{\text{Lfdr}}_i$ is in general different from that by $\widehat{R}_i$. In some applications where the weights are proportional, i.e. $\boldsymbol{a} = c \cdot \boldsymbol{b}$ for some constant $c > 0$, then the rankings by $\widehat{R}_i$ and $\widehat{\text{Lfdr}}_i$ are identical. Specifically $\widehat{R}_i$ is then monotone in $\widehat{\text{Lfdr}}_i$. Further, choosing the cutoff based on $\widehat{N}_i$ is equivalent to that of choosing by a weighted $\widehat{\text{Lfdr}}_i$. This leads to an Lfdr based procedure (Sun et al., 2015), which can be viewed as a special case of Procedure 1.

Procedure 2 *(wFDR control with proportional weights). Rank hypotheses according to $\widehat{Lfdr}_i$ in increasing values. Denote the hypotheses and weights corresponding to $\widehat{Lfdr}_{(i)}$ by $H_{(i)}$ and $a_{(i)}$. Let*

$$k = \max\left\{ j : \left(\sum_{i=1}^{j} a_{(i)} \right)^{-1} \sum_{i=1}^{j} a_{(i)} \widehat{Lfdr}_{(i)} \le \alpha \right\}.$$

Reject $H_{(i)}$*, for* $i = 1, \ldots, k$*.*

Next we investigate the asymptotic performance of Procedure 1. We first give some regularity conditions for the weights. Our theoretical framework requires that the decision weights must be obtained from external sources such as prior data, biological insights, or economical considerations. In particular, *the observed data* $\{X_i : i = 1, \cdots, m\}$ *cannot be used to derive the weights*. The assumption is not only crucial in theoretical developments, but also desirable in practice (to avoid using data twice). Therefore given the domain knowledge, the decision weights do not depend on observed values. Moreover, a model with random (known) weights is employed for technical convenience, as done in Genovese et al. (2006) and Roquain and van de Wiel (2009). We assume that the weights are independent with each other across testing units. Formally, denote e_i the external domain knowledge for hypothesis i, we require the following condition.

Condition 1 *(i)* $(a_i, b_i|X_i, \theta_i, e_i) \overset{d}{\sim} (a_i, b_i|e_i)$ *for* $1 \leq i \leq m$*. (ii)* (a_i, b_i) *and* (a_j, b_j) *are independent for* $i \neq j$*.*

In weighted multiple testing problems, the analysis is always carried out in light of the external information e_i implicitly. The notation of conditional distribution on e_i will be suppressed when there is no ambiguity. In practice, the weights a_i and b_i are usually bounded. We need a weaker condition in our theoretical analysis.

Condition 2 *(Regularity conditions on the weights). Let* C *and* c *be two positive constants.* $E(\sup_i a_i) = o(m), E(\sup_i b_i) = o(m), E(a_i^4) \leq C$*, and* $\min\{E(a_i), E(b_i)\} \geq c$*.*

A consistent Lfdr estimate is needed to ensure the large-sample performance of the data-driven procedure. Formally, we need the following condition.

Condition 3 *It holds that* $\widehat{Lfdr}_i - Lfdr_i = o_P(1)$. *Also,* $\widehat{Lfdr}_i \xrightarrow{d} Lfdr$, *where Lfdr is an independent copy of* $Lfdr_i$.

Remark 3 Condition 3 is a reasonable assumption in many applications. We give a few important scenarios where Condition 3 holds. For the simple random mixture model (2.2), it can be shown that the estimators proposed in Jin and Cai (2007) satisfy $\hat{p} \xrightarrow{p} p$ and $E\|\hat{f}_0 - f_0\|^2 \to 0$. In addition, it is known that the kernel density estimator satisfies $E\|\hat{f} - f\|^2 \to 0$. It follows from Sun and Cai (2007) that Condition 3 holds when the above estimators are used. For the multi-group model (2.3), let $\hat{p}_k$, $\hat{f}_{k0}$, and $\hat{f}_k$ be 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 $\widehat{\text{Lfdr}}_i = (1 - \hat{p}_k)\hat{f}_{0k}(x_i)/\hat{f}_k(x_i)$ if $i \in \mathcal{G}_k$. It follows from Cai and Sun (2009) that Condition 3 holds when we apply Jin and Cai's estimators to the groups separately.

The oracle procedure (3.7) provides an optimal benchmark for all wFDR procedures. The next theorem establishes the asymptotic validity and optimality of the data-driven procedure by showing that the wFDR and ETP levels of the data-driven procedure converge to the oracle levels as $m \to \infty$.

Theorem 2 *Assume Conditions 1-3 hold. Denote by* $wFDR_{DD}$ *the wFDR level of the data-driven procedure (Procedure 1). Let* ETP_{OR} *and* ETP_{DD} *be the ETP levels of the oracle procedure* (3.7) *and data-driven procedure, respectively. Then we have*

(i). $\text{wFDR}_{DD} = \alpha + o(1)$.

(ii). $\text{ETP}_{DD}/\text{ETP}_{OR} = 1 + o(1)$.

5 Simulation Studies

In all simulation studies, we consider a two-point normal mixture model

$$X_i \sim (1-p)N(0,1) + pN(\mu, \sigma^2), i = 1, \cdots, m.$$

The nominal wFDR is fixed at $\alpha = 0.10$. Section 5.1 considers the comparison of different methods under the scenario where there are two groups of hypotheses and within each group the weights are proportional. Section 5.2 compares our methods with existing methods using general weights a_i and b_i that are generated from probability distributions. The proposed method (Procedure 1 in Section 4.2) is denoted by $\triangle$ DD. Other methods include:

1. The wFDR method proposed by Benjamini and Hochberg (1997); denoted by $\square$ BH97. In simulations where $a_i = 1$ for all i, BH97 reduces to the well-known step-up procedure in Benjamini and Hochberg (1995), denoted by BH95.

2. A stepwise wFDR procedure, which rejects hypotheses along the WPO (3.14) ranking sequentially and stops at $k = \max\left\{j : \sum_{i=1}^{j} \widehat{N}_{(i)} \leq 0\right\}$, with $\widehat{N}_{(i)}$ defined in Section 4.2. The method is denoted by $\circ$ WPO. Following similar arguments in the proof of Theorem 2, we can show that the WPO method controls the wFDR at the nominal level asymptotically. This is also verified by our simulation. Meanwhile, we expect that the WPO method will be outperformed by the proposed method ($\triangle$ DD), which operates along the VCR ranking.

3. The adaptive z-value method in Sun and Cai (2007), denoted by + AZ. AZ is valid and optimal in the unweighted case but suboptimal in the weighted case.

5.1 Group-wise weights

This section considers group-wise weights. Our setting is motivated by the applications to GWAS, where the hypotheses can be divided into two groups: those in preselected regions and those in other regions. It is desirable to assign varied weights to separate groups to reflect that the discoveries in preselected regions are more biologically meaningful.

The first simulation study investigates the effect of weights. Consider two groups of hypotheses with group sizes $m_1 = 3000$ and $m_2 = 1500$. In both groups, the non-null proportion is $p = 0.2$ and the non-null distribution is $N(1.9, 1)$. We fix $a_i = 1$ for all i. Hence BH97 reduces to the method proposed in Benjamini and Hochberg (1995), denoted by BH95. The wFDR reduces to the regular FDR, and all methods being considered are valid for FDR control. For hypotheses in group 1, we let $c_1 = a_i/b_i$. For hypotheses in group 2, we let $c_2 = a_i/b_i$. We choose $c_1 = 3$ and vary c_2. Hence the weights are proportional with respective groups and vary across groups.

In each simulation setting, we apply the four methods to the simulated data set and obtain the wFDR and ETP levels by averaging the multiple testing results over 200 replications. In Figure 1, we plot the wFDR and ETP levels of different methods as functions of c_2, which is varied over $[0.1, 0.8]$. Panel (a) shows that all methods control the wFDR under the nominal level, and the BH97 method is conservative. Panel (b) shows that the proposed method dominates all existing methods. The proposed method is followed by the WPO method, which outperforms all unweighted methods (AZ and BH95) since b_i, the weights in the power function, are incorporated in the testing procedure. The BH97 (or BH95) has the smallest ETP. As c_2 approaches 1 or the weights a_i and b_i equalizes, the relative difference of the various methods (other than BH95) becomes less.

In the second simulation study, we investigate the effect of the signal strength μ. Similar as before, consider two groups of hypotheses with group sizes $m_1 = 3000$ and $m_2 = 1500$.

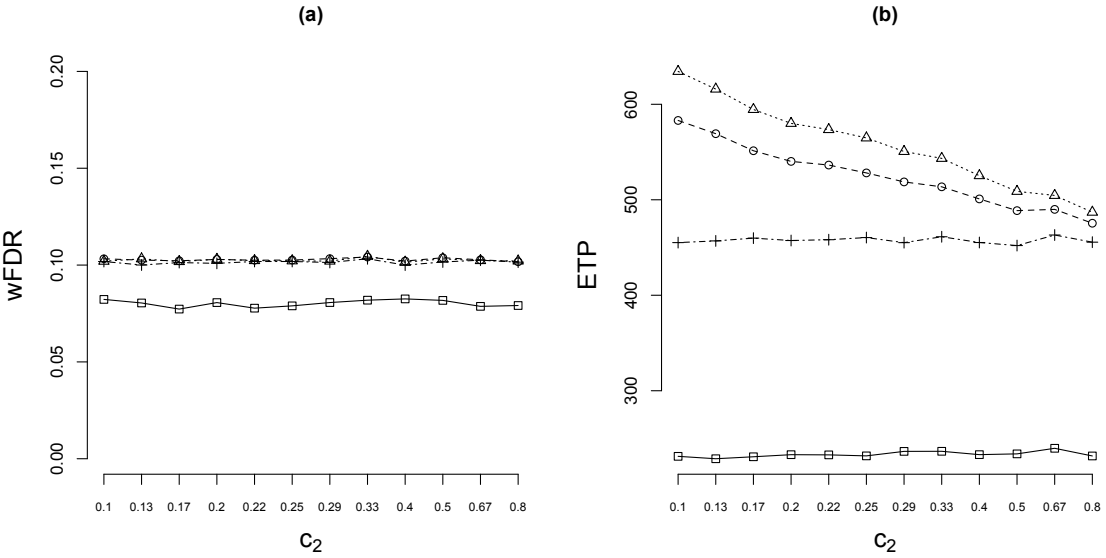

Figure 1: Comparison under group-wise weights: □ BH97 (or BH95), ∘ WPO, △ DD (proposed), and + AZ. The efficiency gain of the proposed method increases as c_1 and c_2 become more distinct.

Under this setting c_1 and c_2 are fixed at 3 and 0.33, respectively. The non-null proportion is $p = 0.2$ and the signal strength μ is varied from 1.75 to 2.5. We apply different methods to the simulated data sets and obtain the wFDR and ETP levels as functions of μ by averaging results over 200 replications. The simulation results are summarized in Figure 2. We can see from Panel (a) that all methods control the wFDR at the nominal level 0.1 approximately (the BH95 method is very conservative and the result is not displayed). Panel (b) shows that the proposed methods dominates other competing methods; and the gain in power is more pronounced when the signals are weak. (The ETP increases rapidly with increased signal strength. For better visualization of results, we present the graph in a logarithmic scale. See Table 1 for results of the BH95 method, as well as the ETP levels in original scales.)

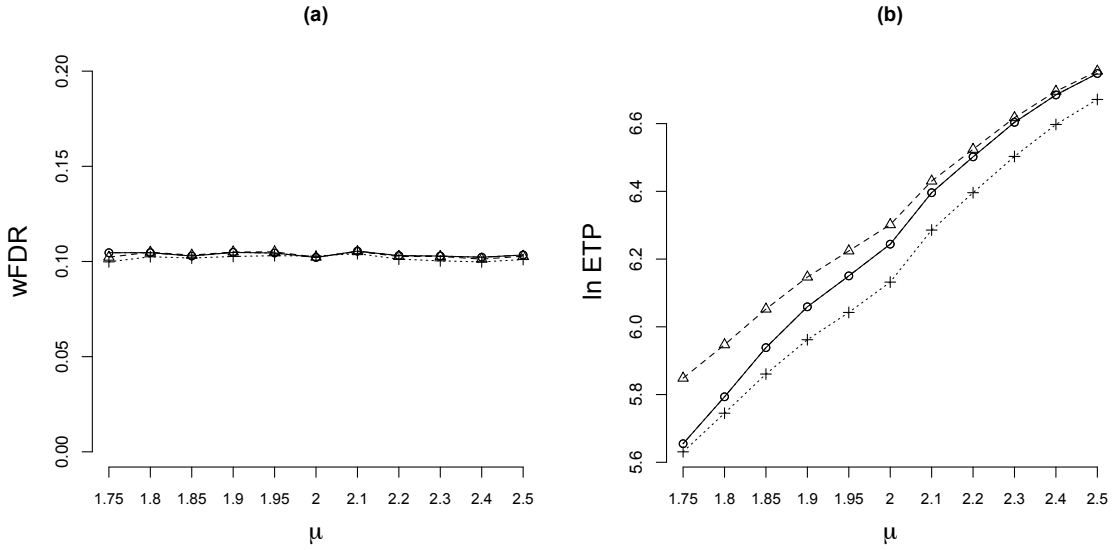

Figure 2: Comparison under group-wise weights: ∘ WPO, △ DD (proposed), and + AZ. The efficiency gain of the proposed method is more pronounced when signals are weak.

5.2 General weights

In applications where domain knowledge is precise (e.g. spatial cluster analysis), dividing the hypotheses into groups and assigning group-wise weights would not be satisfying. This section investigates the performance of our method when random weights (a_i, b_i) are generated from a bivariate distribution.

In the third simulation study, we test $m = 3000$ hypotheses with a_i, the weights associated with the wFDR control, fixed at 1. We generate b_i, the weights associated with the power (or ETP), from log-normal distribution with location parameter $\ln 3$ and scale parameter 1. The location parameter is chosen in a way such that the median weight is 3, similar to those in previous settings. We apply different methods with 200 replications.

The simulation results are summarized in Figure 3. The first row fixes $\alpha = 0.10$ and $p = 0.2$, and plots the wFDR and ETP as functions of μ. The second row fixes $\alpha = 0.10$ and $\mu = 1.9$, and plots the wFDR and ETP as functions of p. The last row fixes $p = 0.2$ and

$\mu = 1.9$, and plots the wFDR and ETP as functions of α. In the plots, we omit the BH95 method (which is very conservative) and present the ETP in a logarithmic scale (for better visualization of results). The following observations can be made: (i) all methods control the wFDR at the nominal level approximately; (ii) by exploiting the weights b_i, the WPO method outperforms the unweighted AZ method; (iii) the proposed method outperforms all competing methods; (iv) Panel (f) shows that gains in power of the proposed method over the WPO method vary at different nominal levels α; (v) similar to the observations in previous simulation studies, the difference between the WPO method and the proposed method decreases with increased signal strength, the efficiency gain of the proposed method is larger as signals become more sparse.

Table 1: ETP values (in original scale) of various methods corresponding to Figure 2

$\mu =$	1.75	1.80	1.85	1.90	1.95	2.0	2.1	2.2	2.3	2.4	2.5
BH95	102.5	125.8	150.6	179.6	204.6	237.0	301.7	361.5	431.2	501.0	567.4
AZ	278.9	312.6	350.9	388.3	420.9	460.4	536.8	599.4	667.2	733.2	789.4
WPO	285.7	328.1	379.4	428.0	468.9	514.9	599.4	666.4	737.8	800.1	852.3
DD (proposed)	346.7	382.6	425.1	467.3	504.8	545.4	620.4	681.3	748.1	808.7	858.2

In the last simulation study, a_i's are assigned to two groups of hypotheses with group sizes $m_1 = 3000$ and $m_2 = 1500$. In groups 1 and 2, we fix $a_i = 1$ and $a_i = 3$, respectively. Conventional FDR methods are only guaranteed to work when all a_i are fixed at 1. Under this setting, we expect that the unweighted AZ may fail to control the wFDR. We then generate random weights b_i from log-normal distribution with location $\ln 6$ and scale 1. The non-null proportion for group 1 is 0.2, and that for group 2 is 0.1. The mean of the the non-null distribution for group 1 or μ_1 is varied between $[-3.75, -2]$ while that for group 2 is fixed at 2. The simulation results are shown in Figure 4. We can see that the unweighted AZ method fails to control the wFDR at the nominal level, which verifies our conjecture. The observations regarding the ETP are similar to those in the previous simulation study. Overall, all numerical studies together substantiate our theoretical results and affirm the

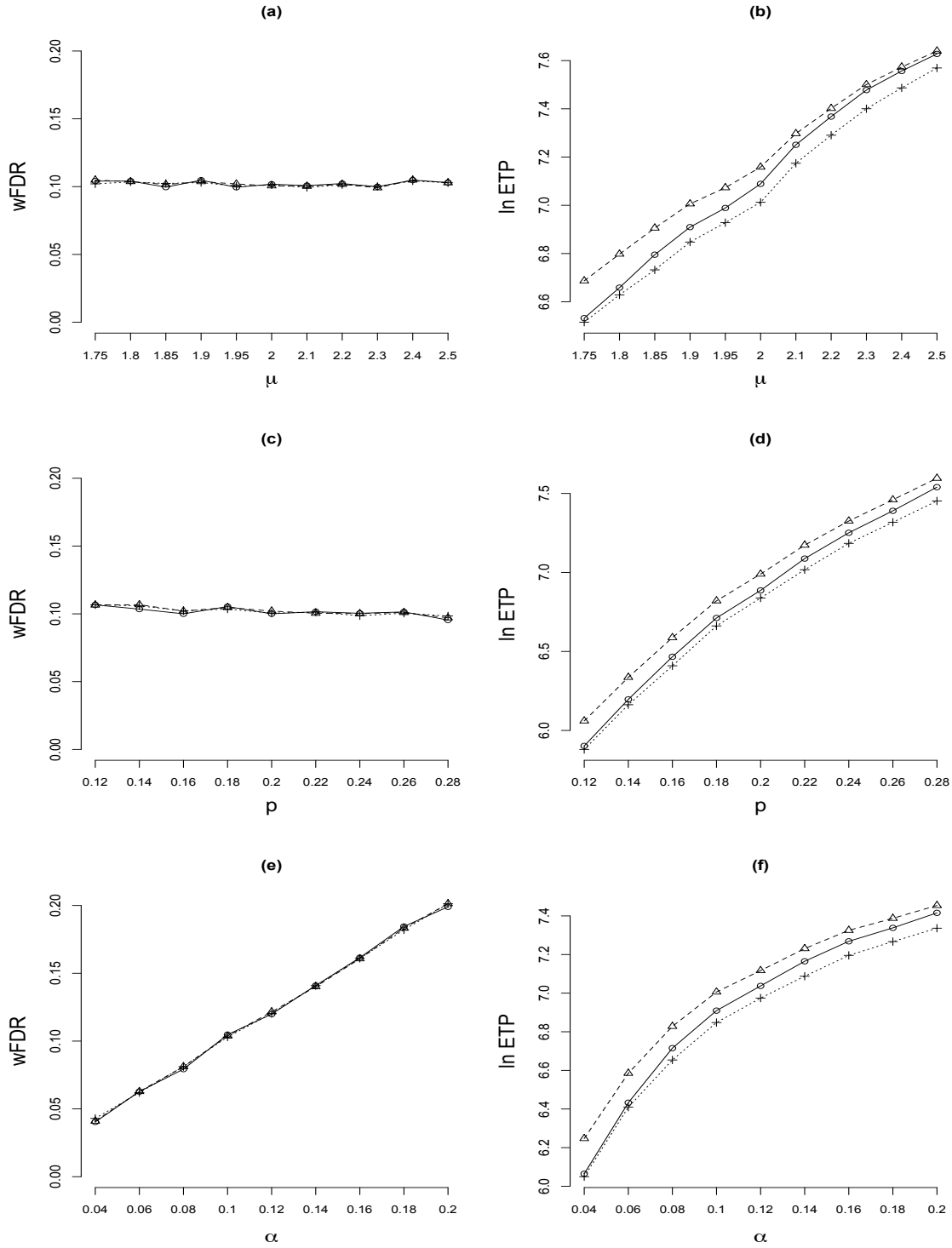

Figure 3: Comparison with general weights: ∘ WPO, △ DD (proposed), and + AZ. All methods control the wFDR approximately at the nominal level. The efficiency gains of the proposed method become more pronounced when (i) the signal strength decreases, (ii) the signals become more sparse, or (iii) the test level α decreases.

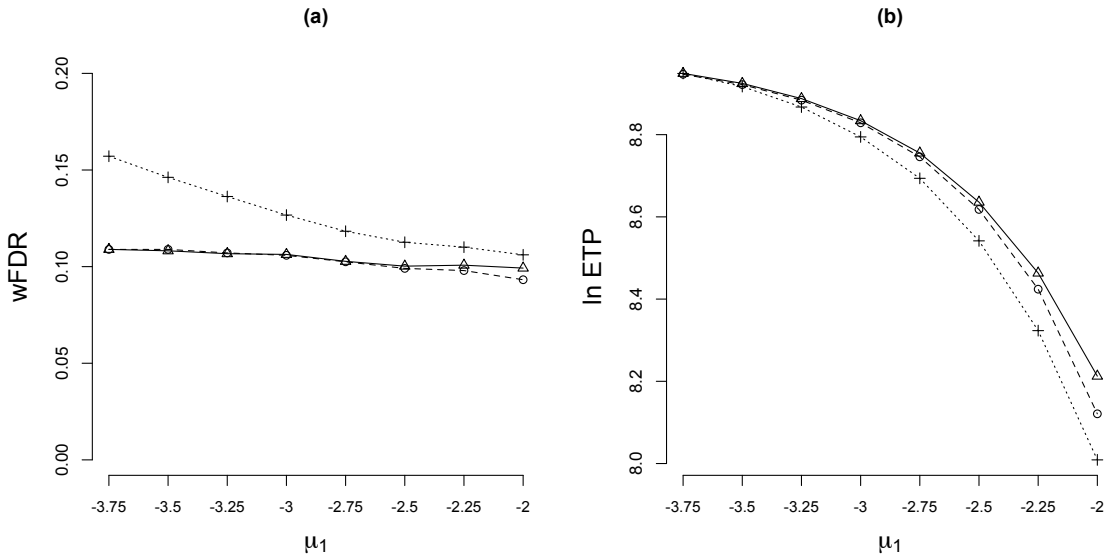

Figure 4: Comparison with general weights: ∘ WPO, △ DD (proposed), and + AZ. The unweighted AZ method fails to control the wFDR at the nominal level. The efficiency gain of the proposed method increases as signals become weaker.

use of the methodology in various settings.

6 Application to GWAS

Weighted FDR procedures have been widely used in GWAS to prioritize the discoveries in pre-selected genomic regions. This section applies the proposed method for analyzing a data set from Framingham Heart Study (Fox et al., 2007; Jaquish, 2007). A brief description of the study, the implementation of our methodology, and the results are discussed in turn.

6.1 Framingham Heart Study

The goal of the study is to decipher the genetic architecture behind the cardiovascular disorders for the Caucasians. Started in 1948 with nearly 5,000 healthy subjects, the study is currently in its third generation of the participants. The biomarkers responsible for

the cardiovascular diseases, for e.g., body mass index (BMI), weight, blood pressure, and cholesterol level, were measured longitudinally.

We analyze a subset of the original data set containing 977 subjects with 418 males and 559 females, whose BMIs are measured over time. Subjects are mostly from 29 years to 85 years old. The current data set also contains genetic information or genotype group of each participant over 5 million single nucleotide polymorphism (SNPs) on different chromosomes. Following the traditional GWAS, we exclude the rare SNPs, that is, SNPs with minor allele frequency less than 0.10, from our analyses. Male and Female populations are analyzed separately. For purpose of illustration, we only report the results from the Male population.

6.2 Multiple testing and wFDR control

We consider the BMI as the response variable and develop a dynamic model to detect the SNPs associated with the BMI. Let $Y_i(t_{ij})$ denote the response (BMI) from the i-th subject at time t_{ij}, $j = 1, \ldots, T_i$. Consider the following model for longitudinal traits:

$$Y_i(t_{ij}) = f(t_{ij}) + \beta_k G_{ik} + \gamma_{i0} + \gamma_{i1} t_{ij} + \epsilon_i(t_{ij}), \tag{6.1}$$

where $f(\cdot)$ is the general effect of time that is modeled by a polynomial function of suitable order, β_k is the effect of the k-th SNP on the response and G_{ik} denotes the genotype of the i-th subject for the k-th SNP. We also consider the random intercepts and random slopes, denoted γ_{0i} and γ_{1i}, respectively, for explaining the subject-specific response trajectories. A bivariate normal distribution for $\gamma_i = (\gamma_{0i}, \gamma_{1i})$ is assumed. Moreover, we assume that the residual errors are normally distributed with zero mean, and covariance matrix Σ_i with an order-one auto-regressive structure.

We fit model (6.1) for each SNP and obtain the estimate of the genetic effect $\widehat{\beta}_k$. If we reject the null hypothesis $H_0 : \beta_k = 0$ vs. $H_1 : \beta_k \neq 0$, then we conclude that the k-th SNP

has a significant association with the BMI. Since we have nearly 5 million SNPs, the false discovery rate needs to be controlled for making scientifically meaningful inference. For each k, we take standardized $\widehat{\beta}_k$ as our z-scores and obtain the estimated ranking statistic $\widehat{R}_k$ as described in (4.2). For selecting the weights, we use the following three methods, with $a_k = 1$ in all the three cases:

Method I: Here we take $b_k = 1$; this is just the unweighted case.

Method II: We first perform a baseline association test for each SNP. We consider only the baseline BMI for each subject and group the response values as High (BMI higher than 25), Low (BMI lower than 18.5), and Medium. For each SNP, we have three genotypes and thus we get a 3×3 table for each SNP and perform a chi-square association test. The p-values are recorded. Now we partition the SNPs into three groups based on these p-values (lower than 0.01, higher than 0.10, and in between). For each group, we compute the average of the inverse of the p-values and take this average as b_k's for all the SNPs belonging to this particular group.

Method III: We consider the dynamic model (6.1) and derive the p-values for testing $H_0 : \beta_k = 0$ vs. $H_1 : \beta_k \neq 0$ for the Female population. We partition the SNPs into three groups based on these p-values: lower than 0.01, higher than 0.10, and in between. For each group, we compute the average of the inverse of the p-values and take this average as the b_k's for all the SNPs belonging to this particular group while analyzing the data from the Male population. Similar methodology of deriving weights from a reference population has been previously explored in Xing et al. (2010).

6.3 Results

In Table 2, we present the number of selected SNPs from three different methods at different α levels. Now we study in detail the SNPs identified at $\alpha = 10^{-6}$; this choice of α is typical

in GWA studies. In Table 3, we list some important SNPs (previously reported in the literature) detected by all three methods. For example, SNPs on chromosomes 3, 5, and 17 were reported in Das et al. (2011, Human Heredity) as significant SNPs associated with high blood pressure and related cardio-vascular diseases. SNPs on chromosome 6 and 8 were reported in Das et al. (2011, Human Genetics). Li et al. (2011, Bioinformatics) reported the SNP on chromosome 10 to be associated with BMI.

Table 2: Number of selected SNPs at different α levels for the Male population

α	Method I	Method II	Method III
10^{-3}	1384	988	1093
10^{-4}	832	447	518
10^{-5}	271	69	91
10^{-6}	86	12	33

Table 3: Some previously reported SNPs detected by all three methods

Chromosome	SNP	Position	Trait/Disease (associated with)
3	ss66149495	16,140,422	Blood Pressure
5	ss66501706	147,356,971	Blood Pressure
6	ss66068448	131,562,687	BMI
8	ss66359352	11093585	BMI
10	ss66311679	32,719,838	BMI
17	ss66154967	29,846,491	Blood Pressure

In Table 4, we list previously reported SNPs which were detected only by Methods II and III. Das et al. (2011) reported the SNP on chromosome 12. Li et al. (2011) reported the SNPs on chromosomes 1, 10, 20, and 22.

Table 4: Some previously reported SNPs detected by Methods II and III only

Chromosome	SNP	Position	Trait/Disease (associated with)
1	ss66185476	8,445,140	BMI
10	ss66293192	32,903,593	BMI
12	ss66379521	130,748,789	Blood Pressure
20	ss66171460	22,580,931	BMI
22	ss66055592	23,420,006	BMI

In Table 5, we list some previously reported SNPs which were detected only by Method III. Das et al. (2011) reported the SNP on chromosome 19. Li et al. (2011) reported the SNPs on chromosomes 1, 10, and 22.

Table 5: Some previously reported SNPs detected by Method III only.

Chromosome	SNP	Position	Trait/Disease (associated with)
1	ss66364251	198321700	BMI
10	ss66303064	32,995,111	BMI
19	ss66092412	56,060,316	Blood Pressure
22	ss66164329	23,420,370	BMI

Note that 11 out of 12 SNPs identified by method II have been, as tabulated in Tables 3 and 4, previously identified in different studies. The SNP ss66077670 on Chromosome 9 is the only identified SNP that has not been previously reported, to our knowledge, and may be further explored by domain experts.

7 Discussion

In the multiple testing literature, procedural, decision, and class weights are often viewed as distinct weighting schemes and have been mostly investigated separately. Although this paper focuses on the decision weights approach, the decision-theoretic framework enables a unified investigation of other weighting schemes. For example, a comparison of the LR (3.8) and WLR (3.11) demonstrates how the LR statistic may be adjusted optimally to account for the decision gains/losses. This shows that procedural weights may be derived in the decision weights framework. Moreover, the difference between the WLR (3.11) and WPO (3.14) shows the important role that p_i plays in multiple testing. In particular the WPO (3.14) provides important insights on how prior beliefs may be incorporated in a decision weights approach to derive appropriate class weights. To see this, consider the multi-class model (2.3). Following the arguments in Cai and Sun (2009), we can conclude that in order to maximize the power, different FDR levels should be assigned to different classes. Similar suggestions for varied class weights have been made in Westfall and Young (1993, pages 169 and 186). These examples demonstrate that the decision weights approach provides a powerful framework to derive both procedural weights and class weights.

We have assumed that the decision weights are pre-specified by the investigators. It is of interest to extend the work to the setting where the weights are unknown. Due to the variability in the quality of external information, subjectivity of investigators, and complexity in modeling and analysis, a systematic study of the issue is beyond the scope of the current paper. Notable progresses have been made, for example, in Roeder and Wasserman (2009) and Roquain and van de Wiel (2009). However, these methods are mainly focused on the weighted p-value approach under the unweighted FDR criterion, hence do not apply to the framework in Benjamini and Hochberg (1997). Moreover, the optimal decision rule in the wFDR problem in general is not a thresholding rule based on the adjusted p-values. Much work is still needed to derive decision weights that would optimally incorporate domain knowledge in large-scale studies.

A Appendix: Proofs

We prove all the technical results in this Appendix.

A.1 Proof of Theorem 1

Proof of Part (i). To show that $\text{wFDR}(\boldsymbol{\delta}_{OR}) = \alpha$, we only need to establish that

$$E_{U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}\left\{\sum_{i=1}^{m} a_i \delta_{OR}^i(\text{Lfdr}_i - \alpha)\right\} = 0,$$

where the notation $E_{U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}$ denotes that the expectation is taken over $U, \boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{X}$. According to the definitions of the capacity function $C(\cdot)$ and threshold t^*, we have

$$\sum_{i=1}^{m} a_i \delta_{OR}^i(\text{Lfdr}_i - \alpha) = C(k) + I(U < p^*)\{C(k+1) - C(k)\}.$$

It follows from the definition of p^* that

$$E_{U|\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}\left\{\sum_{i=1}^{m} a_i\delta_{OR}^i(\text{Lfdr}_i-\alpha)\right\}=C(k)+\{C(k+1)-C(k)\}p^*=0,$$

where the notation $E_{U|\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}$ indicates that the expectation is taken over U while holding $(\boldsymbol{a},\boldsymbol{b},\boldsymbol{X})$ fixed. Therefore

$$E_{U,\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}\left\{\sum_{i=1}^{m} a_i\delta_{OR}^i(\text{Lfdr}_i-\alpha)\right\}=0, \tag{A.1}$$

and the desired result follows.

Proof of Part (ii). Let $\boldsymbol{\delta}^*$ be an arbitrary decision rule such that $wFDR(\boldsymbol{\delta}^*)\leq\alpha$. It follows that

$$E_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}\left\{\sum_{i=1}^{m} a_iE(\delta_i^*|\boldsymbol{a},\boldsymbol{b},\boldsymbol{x})(\text{Lfdr}_i-\alpha)\right\}\leq 0. \tag{A.2}$$

The notation $E(\delta_i^*|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})$ means that the expectation is taken to average over potential randomization conditional on the observations and weights.

Let $\mathcal{I}^+=\{i:\delta_{OR}^i-E(\delta_i^*|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})>0\}$ and $\mathcal{I}^-=\{i:\delta_{OR}^i-E(\delta_i^*|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})<0\}$. For $i\in\mathcal{I}^+$, we have $\delta_{OR}^i=1$ and hence $b_i(1-\text{Lfdr}_i)\geq t^*a_i(\text{Lfdr}_i-\alpha)$. Similarly for $i\in\mathcal{I}^-$, we have $\delta_{OR}^i=0$ and so $b_i(1-\text{Lfdr}_i)\leq t^*a_i(\text{Lfdr}_i-\alpha)$. Thus

$$\sum_{i\in\mathcal{I}^+\cup\mathcal{I}^-}\left\{\delta_{OR}^i-E(\delta_i^*|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})\right\}\left\{b_i(1-\text{Lfdr}_i)-t^*a_i(\text{Lfdr}_i-\alpha)\right\}\geq 0.$$

Note that δ_{OR}^i is perfectly determined by $\boldsymbol{X}$ except for $(k+1)$th decision. Meanwhile,

$$b_{(k+1)}\left(1-\text{Lfdr}_{(k+1)}\right)-t^*a_{(k+1)}\left(\text{Lfdr}_{(k+1)}-\alpha\right)=0$$

by our choice of t^*. It follows that

$$E_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{X}}\left[\sum_{i=1}^{m}\left\{E(\delta_{OR}^{i}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})-E(\delta_{i}^{*}|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})\right\}\left\{b_i(1-\mathrm{Lfdr}_i)-t^{*}a_i(\mathrm{Lfdr}_i-\alpha)\right\}\right]\geq 0. \quad \text{(A.3)}$$

Recall that the power function is given by

$$\mathrm{ETP}(\boldsymbol{\delta})=E\left\{\sum_{i=1}^{m}E(\delta_i|\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})b_i(1-\mathrm{Lfdr}_i)\right\}$$

for any decision rule $\boldsymbol{\delta}$. Combining equations (A.1) – (A.3) and noting that $t^* > 0$, we claim that $\mathrm{ETP}(\boldsymbol{\delta}_{OR}) \geq \mathrm{ETP}(\boldsymbol{\delta}^*)$ and the desired result follows. ∎

A.2 Proof of Theorem 2

A.2.1 Notations

We first recall and define a few useful notations. Let $\mathcal{I}_A$ be an indicator function, which equals 1 if event A occurs and 0 otherwise. Let

$$\begin{aligned}
&N_i=a_i(\mathrm{Lfdr}_i-\alpha),\ \widehat{N}_i=a_i(\widehat{\mathrm{Lfdr}}_i-\alpha),\\
&R_i=\frac{a_i(\mathrm{Lfdr}_i-\alpha)}{b_i(1-\mathrm{Lfdr}_i)+a_i|\mathrm{Lfdr}_i-\alpha|},\ \widehat{R}_i=\frac{a_i(\widehat{\mathrm{Lfdr}}_i-\alpha)}{b_i(1-\widehat{\mathrm{Lfdr}}_i)+a_i|\widehat{\mathrm{Lfdr}}_i-\alpha|},\\
&Q(t)=\frac{1}{m}\sum_{i=1}^{m}N_i\mathcal{I}_{R_i\leq t}\text{ and }\widehat{Q}(t)=\frac{1}{m}\sum_{i=1}^{m}\widehat{N}_i\mathcal{I}_{\widehat{R}_i\leq t}\text{ for }t\in[0,1].
\end{aligned}$$

Note that $Q(t)$ and $\widehat{Q}(t)$, the estimates for oracle and data driven capacities, are non-decreasing and right-continuous. We can further define

$$\lambda_{OR}=\inf\{t\in[0,1]:Q(t)\leq 0\}\text{ and }\widehat{\lambda}=\inf\{t\in[0,1]:\widehat{Q}(t)\leq 0\}. \quad \text{(A.4)}$$

Next we construct a continuous version of $Q(\cdot)$ for later technical developments. Specifically, for $0 \leq R_{(k)} < t \leq R_{(k+1)}$, let

$$Q^c(t) = \{1 - r(t)\}Q\left(R_{(k)}\right) + r(t)Q\left(R_{(k+1)}\right),$$

where c indicates "continuous" and $r(t) = (t - R_{(k)})/(R_{(k+1)} - R_{(k)})$. Let $R_{(m+1)} = 1$ and $N_{(m+1)} = 1$. Similarly we can define a continuous version of $\widehat{Q}(t)$. For $0 \leq \widehat{R}_{(k)} < t \leq \widehat{R}_{(k+1)}$, let

$$\widehat{Q}^c(t) = [1 - \widehat{r}(t)]\widehat{Q}(\widehat{R}_{(k)}) + \widehat{r}(t)\widehat{Q}(\widehat{R}_{(k+1)}),$$

with $\widehat{r}(t) = (t - \widehat{R}_{(k)})/(\widehat{R}_{(k+1)} - \widehat{R}_{(k)})$. Now the inverses of $Q^c(t)$ and $\widehat{Q}^c(t)$ are well defined; denote these inverses by $Q^{c,-1}(t)$ and $\widehat{Q}^{c,-1}(t)$, respectively. By construction, it is easy to see that

$$\mathcal{I}_{R_i \leq \lambda_{OR}} = \mathcal{I}_{R_i \leq Q^{c,-1}(0)} \text{ and } \mathcal{I}_{\widehat{R}_i \leq \widehat{\lambda}} = \mathcal{I}_{\widehat{R}_i \leq \widehat{Q}^{c,-1}(0)}.$$

A.2.2 A useful lemma

We first state and prove a lemma that contains some key facts to prove the theorem.

Lemma 1 *Assume that Conditions 1-3 hold. For any $t \in [0, 1]$, we have*

(i) $E\left(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \leq t]} - N_i \mathcal{I}_{[R_i \leq t]}\right)^2 = o(1)$,

(ii) $E\left\{\left(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \leq t]} - N_i \mathcal{I}_{[R_i \leq t]}\right)\left(\widehat{N}_j \mathcal{I}_{[\widehat{R}_j \leq t]} - N_j \mathcal{I}_{[R_j \leq t]}\right)\right\} = o(1)$, *and*

(iii) $\widehat{Q}^{c,-1}(0) - Q^{c,-1}(0) \xrightarrow{p} 0$.

Proof of Part (i). We first decompose $E\left(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \leq t]} - N_i \mathcal{I}_{[R_i \leq t]}\right)^2$ into three terms:

$$\begin{aligned} & E\left(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \leq t]} - N_i \mathcal{I}_{[R_i \leq t]}\right)^2 \\ = \; & E[(\widehat{N}_i - N_i)^2 \mathcal{I}_{\widehat{R}_i \leq t, R_i \leq t}] + E[\widehat{N}_i^2 \mathcal{I}_{\widehat{R}_i \leq t, R_i > t}] + E[N_i^2 \mathcal{I}_{\widehat{R}_i > t, R_i \leq t}]. \end{aligned} \tag{A.5}$$

Next we argue below that all three terms are of $o(1)$.

First, it follows from the definitions of $\widehat{N}_i$ and N_i that

$$E\left\{(\widehat{N}_i - N_i)^2 \mathcal{I}_{\widehat{R}_i \le t, R_i \le t}\right\} = E\left\{a_i^2 \left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^2 \mathcal{I}_{\widehat{R}_i \le t, R_i \le t}\right\}$$
$$\le E\left\{a_i^2 \left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^2\right\}.$$

By an application of Cauchy-Schwarz inequality, we have

$$E\left\{a_i^2 \left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^2\right\} \le \left\{E(a_i^4)\right\}^{1/2} \left\{E\left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^4\right\}^{1/2}.$$

It follows from Condition 2 that $E(a_i^4) = O(1)$. To show $E\left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^4 = o(1)$, note that both Lfdr_i and $\widehat{\mathrm{Lfdr}}_i$ are in $[0,1]$. Hence $E\left(\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i\right)^4 \le E|\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i|$. Using the fact that $\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i = o_P(1)$, the uniform integrability for bounded random variables, and the Vitali convergence theorem, we conclude that $E|\mathrm{Lfdr}_i - \widehat{\mathrm{Lfdr}}_i| = o(1)$. Therefore, the first term in (A.5) is of $o(1)$.

Next we show that $E\left(\widehat{N}_i^2 \mathcal{I}_{\widehat{R}_i \le t, R_i > t}\right) = o(1)$. Applying Cauchy-Schwarz inequality again, we have

$$E\left(\widehat{N}_i^2 \mathcal{I}_{\widehat{R}_i \le t, R_i > t}\right) \le (1-\alpha)^2 \left\{E(a_i^4)\right\}^{1/2} \left\{P\left(\widehat{R}_i \le t, R_i > t\right)\right\}^{1/2}.$$

Condition 2 implies that $E(a_i^4) = O(1)$; hence we only need to show that $P(\widehat{R}_i \le t, R_i > t) = o(1)$. Let $\eta > 0$ be a small constant. Then

$$P(\widehat{R}_i \le t, R_i > t) = P\left(\widehat{R}_i \le t, R_i \in (t, t+\eta]\right) + P\left(\widehat{R}_i \le t, R_i > t + \eta\right)$$
$$\le P(R_i \in (t, t+\eta]) + P(|\widehat{R}_i - R_i| > \eta).$$

Since R_i is a continuous random variable, we can find $\eta_t > 0$ such that $P(R_i \in (t, t+\eta]) < \varepsilon/2$ for a given ε. For this fixed $\eta_t > 0$, we can show that $P(|\widehat{R}_i - R_i| > \eta_t) < \varepsilon/2$ for sufficiently large n. This follows from $\text{Lfdr}_i - \widehat{\text{Lfdr}}_i = o_P(1)$ and the continuous mapping theorem. Similar argument can be used to prove that $E[N_i^2 \mathcal{I}_{\widehat{R}_i > t, R_i \le t}] = o(1)$, hence completing the proof of part (i).

Proof of Part (ii). As X_i and X_j are identically distributed and our estimates are invariant to permutation, we have

$$E\left\{(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \le t]} - N_i \mathcal{I}_{[R_i \le t]})(\widehat{N}_j \mathcal{I}_{[\widehat{R}_j \le t]} - N_j \mathcal{I}_{[R_j \le t]})\right\} \le E\left(\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \le t]} - N_i \mathcal{I}_{[R_i \le t]}\right)^2.$$

The desired result follows from part (i).

Proof of Part (iii). Define $Q_\infty(t) = E(N_i \mathcal{I}_{R_i \le t})$, where the expectation is taken over $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{X}, \boldsymbol{\theta})$. Let

$$\lambda_\infty = \inf\{t \in [0,1] : Q_\infty(t) \le 0\}.$$

We will show that (i) $Q^{c,-1}(0) \xrightarrow{p} \lambda_\infty$ and (ii) $\widehat{Q}^{c,-1}(0) \xrightarrow{p} \lambda_\infty$. Then the desired result $\widehat{Q}^{c,-1}(0) - Q^{c,-1}(0) \xrightarrow{p} 0$ follows from (i) and (ii).

Fix $t \in [0,1]$. By Condition 2 and WLLN, we have that $Q(t) \xrightarrow{p} Q_\infty(t)$. Since $Q^{c,-1}(\cdot)$ is continuous, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $|Q^{c,-1}(Q_\infty(\lambda_\infty)) - Q^{c,-1}(Q^c(\lambda_\infty))| < \varepsilon$ whenever $|Q_\infty(\lambda_\infty) - Q^c(\lambda_\infty)| < \delta$. It follows that

$$\begin{aligned} & P\left\{|Q_\infty(\lambda_\infty) - Q^c(\lambda_\infty)| > \delta\right\} && \text{(A.6)} \\ \ge\ & P\left\{|Q^{c,-1}(Q_\infty(\lambda_\infty)) - Q^{c,-1}(Q^c(\lambda_\infty))| > \varepsilon\right\} \\ =\ & P\left\{|Q^{c,-1}(0) - \lambda_\infty| > \varepsilon\right\}. && \text{(A.7)} \end{aligned}$$

Equation (A.7) holds since $Q_\infty(\lambda_\infty) = 0$ by the continuity of R_i, and $Q^{c,-1}(Q^c(\lambda_\infty)) = \lambda_\infty$ by the definition of inverse. Therefore we only need to show that for any $t \in [0,1], Q^c(t) \xrightarrow{p}$

$Q_\infty(t)$. Note that

$$E|Q(t) - Q^c(t)| \leq \frac{E(\sup_i a_i)}{m} \to 0,$$

by Condition 2. Using Markov's inequality, $Q(t) - Q^c(t) \xrightarrow{p} 0$. Following from $Q(t) \xrightarrow{p} Q_\infty(t)$, we have $Q^c(t) \xrightarrow{p} Q_\infty(t)$. Therefore (A.6) and hence (A.7) goes to 0 as $m \to \infty$, establishing the desired result (i) $Q^{c,-1}(0) \xrightarrow{p} \lambda_\infty$.

To show result (ii) $\widehat{Q}^{c,-1}(0) \xrightarrow{p} \lambda_\infty$, we can repeat the same steps. In showing $Q^{c,-1}(0) \xrightarrow{p} \lambda_\infty$, we only used the facts that (a) $Q(t) \xrightarrow{p} Q_\infty(t)$, (b) $Q^{c,-1}(\cdot)$ is continuous, and (c) $Q(t) - Q^c(t) \xrightarrow{p} 0$. Therefore to prove $\widehat{Q}^{c,-1}(0) \xrightarrow{p} \lambda_\infty$, we only need to check whether the same conditions (a) $\widehat{Q}(t) \xrightarrow{p} Q_\infty(t)$, (b) $\widehat{Q}^{c,-1}(\cdot)$ is continuous, and (c) $\widehat{Q}(t) - \widehat{Q}^c(t) \xrightarrow{p} 0$ still hold. It is easy to see that (b) holds by definition, and (c) holds by noting that

$$E|\widehat{Q}(t) - \widehat{Q}^c(t)| \leq \frac{E(\sup_i a_i)}{m} \to 0.$$

The only additional result we need to establish is (a).

Previously, we have shown that $Q(t) \xrightarrow{p} Q_\infty(t)$. Therefore the only additional fact that we need to establish is that $|\widehat{Q}(t) - Q(t)| \xrightarrow{p} 0$. Now consider the following quantity:

$$\Delta Q = \{\widehat{Q}(t) - Q(t)\} - [E\{\widehat{Q}(t)\} - E\{Q(t)\}]. \tag{A.8}$$

By repeating the steps of part (i) we can show that

$$|E\{\widehat{Q}(t)\} - E\{Q(t)\}| = |E(N_i \mathcal{I}_{R_i \leq t}) - E(\widehat{N}_i \mathcal{I}_{\widehat{R}_i \leq t})| \to 0. \tag{A.9}$$

By definition, $\Delta Q = m^{-1} \sum_{i=1}^m \{\widehat{N}_i \mathcal{I}_{[\widehat{R}_i \leq t]} - N_i \mathcal{I}_{[R_i \leq t]}\} - [E(\widehat{N}_i \mathcal{I}_{\widehat{R}_i \leq t}) - E(N_i \mathcal{I}_{R_i \leq t})]$. For an application of WLLN for triangular arrays (see, for e.g., Theorem 6.2 of Billingsley, 1991),

we need to show that

$$\text{var}(\sum_{i=1}^{m}\{\widehat{N}_i\mathcal{I}_{[\widehat{R}_i\leq t]} - N_i\mathcal{I}_{[R_i\leq t]}\})/m^2 \to 0.$$

Using the result in Part (i) we deduce that,

$$\begin{aligned} & m^{-2}\text{Var}\left\{\sum_{i=1}^{m}\left(\widehat{N}_i\mathcal{I}_{[\widehat{R}_i\leq t]} - N_i\mathcal{I}_{[R_i\leq t]}\right)\right\} \leq m^{-2}E\left\{\sum_{i=1}^{m}\left(\widehat{N}_i\mathcal{I}_{[\widehat{R}_i\leq t]} - N_i\mathcal{I}_{[R_i\leq t]}\right)\right\}^2 \\ = & \left(1-\frac{1}{m}\right)E\left\{\left(\widehat{N}_i\mathcal{I}_{[\widehat{R}_i\leq t]} - N_i\mathcal{I}_{[R_i\leq t]}\right)\left(\widehat{N}_j\mathcal{I}_{[\widehat{R}_j\leq t]} - N_j\mathcal{I}_{[R_j\leq t]}\right)\right\} \\ & +\frac{1}{m}E\left(\widehat{N}_i\mathcal{I}_{[\widehat{R}_i\leq t]} - N_i\mathcal{I}_{[R_i\leq t]}\right)^2 = o(1). \end{aligned}$$

It follows from the WLLN for triangular arrays that $|\Delta Q| \xrightarrow{p} 0$. Combining (A.8) and (A.9), we conclude that $|\widehat{Q}(t) - Q(t)| \xrightarrow{p} 0$, which completes the proof. ∎

A.2.3 Proof of Theorem 2

Proof of Part (i). Consider the oracle and data driven thresholds λ_{OR} and $\widehat{\lambda}$ defined in Equation (A.4) in Appendix A.2.1. The wFDRs of the oracle and data-driven procedures are

$$\begin{aligned} \text{wFDR}_{OR} &= \frac{E\left\{\sum_i a_i(1-\theta_i)\delta^i_{OR}\right\}}{E\left(\sum_i a_i\delta^i_{OR}\right)}, \\ \text{wFDR}_{DD} &= \frac{E\left\{\sum_i a_i(1-\theta_i)\mathcal{I}_{\widehat{R}_i\leq\widehat{\lambda}}\right\}}{E\left(\sum_i a_i\mathcal{I}_{\widehat{R}_i\leq\widehat{\lambda}}\right)}. \end{aligned}$$

Making the randomization explicit, the wFDR of the oracle procedure is

$$\mathrm{wFDR}_{OR} = \frac{E\left\{m^{-1}\sum_i a_i(1-\theta_i)\mathcal{I}_{R_i\le\lambda_{OR}} + m^{-1}a_{i^*}(1-\theta_{i^*})\delta_{OR}^{i^*}\right\}}{E\left(m^{-1}\sum_i a_i\mathcal{I}_{R_i\le\lambda_{OR}} + m^{-1}a_{i^*}\delta_{OR}^{i^*}\right)},$$

where i^* indicates the randomization point in a realization. Note that both $E\{a_{i^*}(1-\theta_{i^*})\delta_{OR}^{i^*}/m\}$ and $E\{a_{i^*}\delta_{OR}^{i^*}/m\}$ are bounded by $E(a_{i^*}/m)$. Hence by Condition 2 both quantities are of $o(1)$.

From the discussion in Appendix A.2.1, $\mathcal{I}_{R_i\le\lambda_{OR}} = \mathcal{I}_{R_i\le Q^{c,-1}(0)}$ and $\mathcal{I}_{\widehat{R}_i\le\widehat{\lambda}} = \mathcal{I}_{\widehat{R}_i\le\widehat{Q}^{c,-1}(0)}$. According to Part (iii) of Lemma 1, we have $\{\widehat{R}_i - \widehat{Q}^{c,-1}(0)\} - \{R_i - Q^{c,-1}(0)\} = o_P(1)$. Following the proof of Lemma 1 that

$$E\left\{a_i(1-\theta_i)\mathcal{I}_{\widehat{R}_i-\widehat{Q}^{c,-1}(0)\le 0}\right\} = E\left\{a_i(1-\theta_i)\mathcal{I}_{R_i-Q^{c,-1}(0)\le 0}\right\} + o(1). \tag{A.10}$$

It follows that $m^{-1}E\{\sum_i a_i(1-\theta_i)\mathcal{I}_{\widehat{R}_i\le\widehat{\lambda}}\} \to m^{-1}E\{\sum_i a_i(1-\theta_i)\mathcal{I}_{R_i\le\lambda_{OR}}\}$. Similarly, we can show that

$$E\left(a_i\mathcal{I}_{\widehat{R}_i-\widehat{Q}^{c,-1}(0)\le 0}\right) = E\left(a_i\mathcal{I}_{R_i-Q^{c,-1}(0)\le 0}\right) + o(1). \tag{A.11}$$

Further from Condition 2 the quantity $m^{-1}E(\sum_i a_i\mathcal{I}_{R_i\le\lambda_{OR}})$ is bounded away from zero. To see this, note that Condition 1 implies that a_i is independent of Lfdr_i. It follows that

$$m^{-1}E\left(\sum_{i=1}^m a_i\mathcal{I}_{R_i\le\lambda_{OR}}\right) = E\left(a_i\mathcal{I}_{R_i\le\lambda_{OR}}\right) \ge c\widetilde{p}_\alpha > 0, \tag{A.12}$$

where $P(\mathrm{Lfdr}(X) \le \alpha) \ge \widetilde{p}_\alpha$ for some $\widetilde{p}_\alpha \in (0,1]$ for the choice of the nominal level $\alpha \in (0,1)$ and X, an i.i.d copy of X_i. This holds for any non-vanishing α. (Note that all hypotheses with $\mathrm{Lfdr}_i < \alpha$ will be rejected automatically). Therefore we conclude that

$$\mathrm{wFDR}_{DD} = \mathrm{wFDR}_{OR} + o(1) = \alpha + o(1).$$

Proof of Part (ii). The quantity $m^{-1}ETP_{DD}$ is defined as $m^{-1}E\left(b_i\theta_i\mathcal{I}_{[\widehat{R}_i\leq\widehat{\lambda}]}\right)$. Making the randomization explicit, we have

$$m^{-1}ETP_{OR} = E\left(\frac{1}{m}\sum_i b_i\theta_i\mathcal{I}_{[R_i\leq\lambda_{OR}]} + \frac{1}{m}b_{i^*}\theta_{i^*}\delta_{OR}^{i^*}\right),$$

where i^* indicates the randomized point. By Condition 2

$$0 \leq m^{-1}E\left(b_{i^*}\theta_{i^*}\delta_{OR}^{i^*}\right) \leq \frac{Eb_{i^*}}{m} \leq \frac{E\sup_i b_i}{m} = o(1).$$

From the discussion in Appendix A.2.1, $\mathcal{I}_{R_i\leq\lambda_{OR}} = \mathcal{I}_{R_i\leq Q^{c,-1}(0)}$ and $\mathcal{I}_{\widehat{R}_i\leq\widehat{\lambda}} = \mathcal{I}_{\widehat{R}_i\leq\widehat{Q}^{c,-1}(0)}$. Repeating the steps in proving the wFDR, we can show that

$$E\left(b_i\theta_i\mathcal{I}_{[\hat{R}_i\leq\hat{\lambda}]}\right) = E\left(b_i\theta_i\mathcal{I}_{[R_i\leq\lambda_{OR}]}\right) + o(1).$$

Finally, it is easy to show that $E\left(b_i\theta_i\mathcal{I}_{[R_i\leq\lambda_{OR}]}\right) \geq c(1-\alpha)\widetilde{p}_\alpha$, which is bounded below by a nonzero constant. Here the positive constant c is as defined in Condition 2 and $\widetilde{p}_\alpha$ is defined immediately after (A.12). We conclude that $ETP_{DD}/ETP_{OR} = 1 + o(1)$ and the proof is complete. ∎

A.3 Proofs of Propositions

A.3.1 Proof of Proposition 2

Let $\lambda > 0$ be the relative cost of a false positive to a false negative. Consider the following weighted classification problem with loss function:

$$L_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{\theta},\boldsymbol{\delta}) = \sum_{i=1}^{m}\{\lambda a_i(1-\theta_i)\delta_i + b_i\theta_i(1-\delta_i)\}. \tag{A.13}$$

We aim to find $\boldsymbol{\delta}$ that minimizes the posterior loss $E_{\boldsymbol{\theta}|\boldsymbol{X}}\{L_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{\theta},\boldsymbol{\delta})\}$

$$\begin{aligned}&\operatorname*{argmin}_{\boldsymbol{\delta}} \sum_{i=1}^{m} \{\lambda a_i P(\theta_i = 0|X_i)\delta_i + b_i P(\theta_i = 1|X_i)(1-\delta_i)\} \\ =&\operatorname*{argmin}_{\boldsymbol{\delta}} \sum_{i=1}^{m} \{\lambda a_i P(\theta_i = 0|X_i) - b_i P(\theta_i = 1|X_i)\}\,\delta_i.\end{aligned}$$

Therefore the optimal decision rule $\boldsymbol{\delta}_{PF} = (\delta^i_{PF} : i = 1, \cdots, m)$ is of the form

$$\delta^i_{PF} = I\left[\frac{a_i P(\theta_i = 0|X_i)}{b_i P(\theta_i = 1|X_i)} < \frac{1}{\lambda}\right], \tag{A.14}$$

which reduces to the test statistic defined in (3.14).

Next note that $Q_{PF}(t)$ is a continuous and increasing function of t. Therefore we can find t_{PF} such that $Q_{PF}(t_{PF}) = \alpha$. For an arbitrary decision rule $\boldsymbol{\delta}^* \in \mathcal{D}_\alpha$, we must have $ETP(\boldsymbol{\delta}^*) \leq ETP(\boldsymbol{\delta}_{PF})$. Suppose not, then there exists $\boldsymbol{\delta}^* \in \mathcal{D}_\alpha$ such that $\mathrm{PFER}(\boldsymbol{\delta}^*) \leq \alpha = \mathrm{PFER}(\boldsymbol{\delta}_{PF})$ and $-\mathrm{ETP}(\boldsymbol{\delta}^*) < -\mathrm{ETP}(\boldsymbol{\delta}_{PF})$. Consider a weighted classification problem with $\lambda = 1/t_{PF}$. Then we can show that $\boldsymbol{\delta}^*$ has a smaller classification risk compared to $\boldsymbol{\delta}_{PF}$. This is a contradiction. Therefore we must have $\mathrm{ETP}(\boldsymbol{\delta}^*) \leq \mathrm{ETP}(\boldsymbol{\delta}_{PF})$. ∎

A.3.2 Proof of Proposition 3

Proof of Part (i). For convenience of notation, define $S_i = 1/\mathrm{VCR}_i$. We show that rankings by increasing values of R_i and S_i are the same. If $i \in S^+$, then all values are positive. Sorting by increasing S_i is the same as sorting by decreasing $(1/S_i) + 1$ and hence by increasing $1/(1/S_i + 1)$, which is precisely sorting by increasing R_i. If $i \in S^-$, then all values are negative. Sorting by increasing S_i is the same as sorting by decreasing $(1/S_i) - 1$ and hence by increasing $1/(1/S_i - 1)$, which is again the same as sorting by increasing R_i. The desired result follows.

Proof of Part (ii). The result follows directly from the facts that (a) R_i is negative when $i \in S^-$ and (b) R_i is positive if $i \in S^+$. ∎

References

[1] Benjamini, Y., and Heller, R. (2007), "False Discovery Rates for Spatial Signals," *Journal of the American Statistical Association*, 102, 1272–1281.

[2] Benjamini, Y., and 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.

[3] Benjamini, Y., and Hochberg, Y. (1997), "Multiple Hypotheses Testing with Weights," *Scandinavian Journal of Statistics*, 24, 407–418.

[4] Billingsley, P. (1991), *Probability and Measure* (2nd ed.), New York: John Wiley & Sons.

[5] Box, G. E. P. (1950), "Problems in the Analysis of Growth and Wear Curves," *Biometrics*, 6, 362–389.

[6] Cai, T. T., and Sun, W. (2009), "Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks," *Journal of the American Statistical Association*, 104, 1467–1481.

[7] Das, K., Li, J., Fu, G., Wang, Z., and Wu, R. (2011), "Genome-Wide Association Studies for Bivariate Sparse Longitudinal Data," *Human Heredity*, 72, 110–120.

[8] Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., Xu, M., Ahn, K., Mauger, D., Li, R., and Wu, R. (2011), "A Dynamic Model for Genome-Wide Association Studies," *Human Genetics*, 129, 629–639.

[9] Efron, B. (2008), "Simultaneous Inference: When Should Hypothesis Testing Problems be Combined?" *The Annals of Applied Statistics*, 2, 197–223.

[10] Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, New York: Cambridge University Press.

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

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

[13] Fox, C. S., Heard-Costa, N., Cupples, L. A., Dupuis, J., Vasan, R. S., and Atwood, L. D. (2007), "Genome-Wide Association to Body Mass Index and Waist Circumference: the Framingham Heart Study 100K Project," *BMC Medical Genetics*, 8, S18.

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

[15] Hu, J. X., Zhao, H., and Zhou, H. H. (2010), "False Discovery Rate Control With Groups," *Journal of the American Statistical Association*, 105, 1215–1227.

[16] Jaquish, C. E. (2007), "The Framingham Heart Study, on its Way to Becoming the Gold Standard for Cardiovascular Genetic Epidemiology?" *BMC Medical Genetics*, 8, 63.

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

[18] Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011), "The Bayesian Lasso for Genome-Wide Association Studies," *Bioinformatics*, 27, 516–523.

[19] Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X., and Liang, L. (2013), "Using eQTL Weights to Improve Power for Genome-Wide Association Studies: a Genetic Study of Childhood Asthma," *Frontiers in Genetics*, 4, Article 103.

[20] Lin, W.-Y., and Lee, W.-C. (2012), "Improving Power of Genome-Wide Association Studies with Weighted False Discovery Rate Control and Prioritized Subset Analysis," *PLoS ONE*, 7, e33716.

[21] Lin, C.-Y., Xing, G., Ku, H.-C., Elston, R. C., and Xing, C. (2014), "Enhancing the Power to Detect Low-Frequency Variants in Genome-Wide Screens," *Genetics*, 196, 1293–1302.

[22] Pẽna, E. A., Habiger, J. D., and Wu, W. (2011), "Power-Enhanced Multiple Decision Functions Controlling Family-Wise Error and False Discovery Rates," *The Annals of Statistics*, 39, 556–583.

[23] Roeder, K., Bacanu, S.-A., Wasserman, L., and Devlin, B. (2006), "Using Linkage Genome Scans to Improve Power on Association in Genome Scans," *The American Journal of Human Genetics*, 78, 243–252.

[24] Roeder, K., and Wasserman, L. (2009), "Genome-Wide Significance Levels and Weighted Hypothesis Testing," *Statistical Science*, 24, 398–413.

[25] Rogosa, D. (2003), "Student Progress in California Charter Schools," 1999–2002. Available at http://statweb.stanford.edu/ rag/draft/charter9902.pdf.

[26] Roquain, E., and van de Wiel, M. A. (2009), "Optimal Weighting for False Discovery Rate Control," *Electronic Journal of Statistics*, 3, 678–711.

[27] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

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

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

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

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

[32] Sun, W., Reich, B., Cai, T. T., Guindani, M., and Schwartzman, A. (2015), "False Discovery Control in Large-Scale Spatial Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 77, 59–83.

[33] Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: John Wiley and Sons.

[34] Xing, C., Cohen, J. C., and Boerwinkle, E. (2010), "A Weighted False Discovery Rate Control Procedure Reveals Alleles at FOXA2 that Influence Fasting Glucose Levels," *The American Journal of Human Genetics*, 86, 440–446.