Signal Classification for the Integrative Analysis of Multiple Sequences of Multiple Tests

Dongdong Xiang
University of Pennsylvania, Philadelphia, USA
Sihai Dave Zhao
University of Illinois at Urbana-Champaign, Champaign, USA
T. Tony Cai
University of Pennsylvania, Philadelphia, USA

Summary. The integrative analysis of multiple datasets is becoming increasingly important in many fields of research. When the same features are studied in several independent experiments, a common integrative approach is to jointly analyze the multiple sequences of multiple tests that result. It is frequently necessary to classify each feature into one of several categories, depending on the null and non-null configuration of its corresponding test statistics. This paper studies this signal classification problem, motivated by a range of applications in large-scale genomics. Two new types of misclassification rates are introduced, and both oracle and data-driven procedures are developed to control each of these types while also achieving the largest expected number of correct classifications. The proposed data-driven procedures are proved to be asymptotically valid and optimal under mild conditions, and are shown in numerical experiments to be nearly as powerful as oracle procedures, with substantial gains in power over their competitors in many settings. In an application to psychiatric genetics, the proposed procedures are used to discover genetic variants that may affect both bipolar disorder and schizophrenia, as well as variants that may help distinguish between these conditions.

Keywords: Integrative Analysis, Multiple Testing, Set-specific Marginal False Discovery Rate, Signal Classification, Total Marginal False Discovery Rate

1. Introduction

1.1. Overview

Most statistical methods for multiple testing are intended for analyzing a single sequence of multiple tests, arising from a single study. In recent years, however, summary test statistics and p-values from multiple studies have become readily publicly accessible. Researchers have realized that a great deal of information is contained in the comparison of these studies, and that much can be learned by discovering their similarities and differences through an integrative analysis. Thus an emerging statistical problem is to develop powerful and efficient methods for the joint analysis of multiple sequences of multiple tests, where the same features are tested in each sequence.

These types of joint analyses are especially prevalent in modern large-scale genomics studies, for example the effort to understand the genetic regulation of gene expression

Table 1. Signal classes and labels for two sequences of multiple tests

Class label	θ_{1i}	θ_{2i}
0	0	0
1	0	1
2	1	0
3	1	1

in humans. The Genotype Tissue Expression Project (Lonsdale et al., 2013) collected genotype as well as gene expression data from 53 tissue types from hundreds of donors. A major task is to determine which genetic variants regulate the expression levels of which genes. This is accomplished by significance testing, for each gene in each tissue, of the association between the expression level and each typed genetic variant. But because some regulatory variants may only be active in certain tissue types, an important problem is to classify each variant in terms of the tissues in which their associated test statistics are or are not significant (Flutre et al., 2013; Torres et al., 2014; GTEx Consortium, 2015). This requires the simultaneous consideration of an enormous number of sequences of multiple tests.

Similar joint analyses arise in psychiatric genetics. Some disorders, such as schizophrenia and bioplar disorder, share many symptoms and can be difficult to differentiate in clinical diagnoses (Andreassen et al., 2013). Several large genome-wide association studies have now made it possible to compare the genetics of these two diseases (Ruderfer et al., 2014; Gratten et al., 2014; Cross-Disorder Group of Psychiatric Genomics Consortium, 2013b). Identifying genetic variants that are significantly associated with one disease but not the other can pave the way for a molecular diagnostic procedure that can more accurately distinguish the two conditions, while identifying variants that are associated with both conditions can shed light on their common biological basis. Classifying variants in this way requires the joint analysis of two sets of summary statistics, one from each disorder.

These types of integrative analysis abound across genetics and genomics research, and can frequently be formulated in terms of grouping genomic features into different classes based on their corresponding test statistics across the multiple sequences of tests. To fix ideas, let X_{ji} be the z-score for the ith genomic feature in the jth study ($i = 1, \ldots, m$; $j = 1, \ldots, J$); for example, X_{ji} can denote the test statistic, in the jth tissue, for the association between the ith genetic variant and the expression level of a given gene. This paper will only consider $J = 2$, but extensions to more than two studies are straightforward. Let $\theta_{jk} \in \{0, 1\}$ indicate whether X_{ji} contains signal or not, so that $\theta_{ji} = 1$ if X_{ji} is truly non-null and $\theta_{ji} = 0$ otherwise. The four possible configurations of $(\theta_{1i}, \theta_{2i})$ determine four classes to which each genomic feature can belong. Table 1 lists and labels these classes. If X_{ji} corresponds to the ith expression quantitative trait locus in the jth tissue, for example, identifying cross-tissue versus tissue-specific loci becomes equivalent to classifying the tests either into class 3, or into classes 1 or 2.

This paper studies this signal classification problem, where the goal is to correctly assign as many genomic features into these signal classes as possible while controlling some measure of misclassification error. Signal classification can be viewed as a generalization

Table 2. Example confusion matrix after applying a signal classification procedure

	True class				
Predicted class	0	1	2	3	Total
0	C_{00}	C_{01}	C_{02}	C_{03}	R_0
1	C_{10}	C_{11}	C_{12}	C_{13}	R_1
2	C_{20}	C_{21}	C_{22}	C_{23}	R_2
3	C_{30}	C_{31}	C_{32}	C_{33}	R_3
Total	m_0	m_1	m_2	m_3	m

of the standard multiple testing problem, which only seeks to determine whether each feature is null or non-null and is therefore equivalent to binary classification. In contrast, signal classification is more similar to multi-class classification, where the results of applying a classification procedure to two sequences of multiple tests can be displayed in the form of a confusion matrix. An example is shown in Table 2.

This paper proposes novel methods for signal classification. New concepts for measuring misclassification error are first defined. In the usual multiple testing framework, where signals are either null or non-null, misclassification error is frequently measured using the false discovery rate (Benjamini and Hochberg, 1995). However, when signals can fall into more than two classes, there are multiple possible types of false discovery rates, each of which measures different combinations of the off-diagonal entries of the confusion matrix in Table 2. Two types in particular are considered in this paper. New asymptotically optimal methods are then developed under the framework of Lagrangian multiplier optimization to control each of these types of misclassification error while achieving the largest possible number of correct classifications. Related theoretical results that determine the optimal thresholds for the proposed procedures and reveal relationships between the multi-class and binary classification approaches, are also provided. Though signal classification is discussed here in the context of the joint analysis of multiple sequences of test statistics, the framework and methods proposed in this paper can be readily extended to other situations where classification into multiple signal classes is necessary.

1.2. Related work

Studying multiple sequences of tests has become relevant as interest in areas such as integrative genomics (Hawkins et al., 2010; Kristensen et al., 2014; Li, 2013; Ritchie et al., 2015) has grown. However, research in the multiple sequence setting has still focused on binary classification, typically on the problem of determining whether or not signals belong to class 3 of Table 1. This is of great interest because class 3 signals are more likely to constitute replicable scientific findings (Benjamini et al., 2009; Bogomolov and Heller, 2013; Heller et al., 2014).

A common framework is to posit a four-group mixture model for the (X_{1i}, X_{2i}), where each mixture component corresponds to one of the signal classes in Table 1. Many authors have shown that the optimal multiple testing procedure is based on the local false discovery rate for being in class 3, which requires the unknown null and alternative distributions of the test statistics in each sequence. One approach is to approximate the local false discovery rate in some way (Chi, 2008; Du and Zhang, 2014). An alternative is to estimate the unknown distributions and obtain a data-driven version of the optimal testing procedure (Chung et al., 2014; Heller and Yekutieli, 2014). Recent technical reports by Urbut et al. (2017) and Li et al. (2013, 2017) extend this type of approach to three or more sequences of test statistics.

All of these methods are still limited to only two possible decisions for each tested feature: whether that feature belongs to a given set of classes of interest, or not. For example, Heller and Yekutieli (2014) defines the set of interest to contain only class 3, in order to discover features that are significant in both sequences. Alternatively, the set of interest could be defined to contain both classes 1 and 2, in order to identify signals that are unique to only one of the two sequences, and a modified version of the method of Heller and Yekutieli (2014) could be applied.

However, there appear to be no existing methods for signal classification with multiple sequences of tests that allow for two or more sets of signal classes of interest. A common approach is to identify null and non-null genomic features in each sequence separately, controlling sequence-specific false discovery rates. These separate discoveries are then used to determine the signal class of each feature. For example, a feature called as a non-discovery in sequence 1, at a false discovery rate of level α_1, and a discovery in sequence 2 at level α_2, would be assigned to class 2 of Table 1. However, it is unclear how the separate error levels α_1 and α_2 contribute to the overall misclassification error.

1.3. Organization of the paper

Section 2 proposes two definitions of misclassification error in this multi-class setting and then formalizes the related signal classification problems. Section 3 develops new oracle and data-driven methods to achieve optimal classification under error control, and establishes related theoretical results. Simulation results demonstrating the performance of the proposed methods are given in Section 4. In Section 5, the proposed procedures are applied to study the genetic architectures of bipolar disorder and schizophrenia. A discussion on possible extensions is given in Section 6. Proofs and additional results are contained in Section 7 and the Appendices.

2. Problem formulation

2.1. Definitions

As illustrated in Table 2, two sequences of test statistics X_{1i} and X_{2i} give rise to four possible signal classes $0, \ldots, 3$. However, in most applications not all signal classes are equally interesting. Frequently, the four possible classes are partitioned into $K + 1$ disjoint subsets, where K may equal 1, 2, or 3. Let $\mathcal{S}_0 \subset \{0, \ldots, 3\}$ denote the set of classes that are not of interest, and let $\mathcal{S}_k \subset \{0, \ldots, 3\} \backslash \mathcal{S}_0$ for $k = 1, \ldots, K$ denote disjoint subsets of the remaining important classes, such that $\cup_{k=0}^{K} \mathcal{S}_k = \{0, \ldots, 3\}$. For a concrete example, suppose that X_{ji} is the differential expression z-score of the ith gene in brain region j. In some analyses the goal may be to classify each gene as being active only in region 1, active only in region 2, or active in both regions. In this case, $K = 3$ and $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, and $\mathcal{S}_3 = \{3\}$. In other applications the goal may only be to distinguish genes that are region-specific, regardless of region, from those that are not. In this case $K = 2$, $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1, 2\}$, and $\mathcal{S}_2 = \{3\}$.

A signal classification procedure is represented by a decision rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$, where $\delta_i \in \{0, \ldots, K\}$ that indicates the set $\mathcal{S}_k$ to which the ith genomic feature is assigned. Usual notions of power (Sarkar, 2002; Genovese and Wasserman, 2002; Taylor et al., 2005; Basu et al., 2017; Cai and Sun, 2017) and false discovery rate (Storey, 2002; Benjamini and Hochberg, 1995; Genovese and Wasserman, 2002) need to be generalized in order to accommodate multiple sets of signal classes of interest. To measure the power of $\boldsymbol{\delta}$, define the total expected number of true positives to be

$$\text{tETP}(\boldsymbol{\delta}) = \mathbb{E}(\sum_{k=1}^{K} \sum_{\ell \in \mathcal{S}_k} C_{\ell\ell}), \tag{1}$$

where the $C_{\ell\ell}$ are diagonal entries of the confusion matrix in Table 2. This measure equals the total number of tests correctly classified by $\boldsymbol{\delta}$ into any of the sets $\mathcal{S}_k$ of interest.

There are multiple ways to measure the misclassification error incurred by $\boldsymbol{\delta}$. One possibility is the total marginal false discovery rate, defined to be

$$\text{tmFDR}(\boldsymbol{\delta}) = \frac{\mathbb{E}(\sum_{k=1}^{K} \sum_{\ell \in \mathcal{S}_k} \sum_{\ell' \neq \ell} C_{\ell\ell'})}{\mathbb{E}(\sum_{k=1}^{K} \sum_{\ell \in \mathcal{S}_k} R_\ell)}, \tag{2}$$

where the $C_{\ell\ell'}$ are the off-diagonal entries of Table 2. The numerator of (2) is the average number of features incorrectly classified into any of the $\mathcal{S}_k$, and the denominator equals the expected value of the total number of features classified into any of the $\mathcal{S}_k$. The quantity (2) reduces to the standard marginal false discovery rate in the binary classification problem of distinguishing between $\mathcal{S}_0$ and $\cup_{k=1}^{K} \mathcal{S}_k$. Alternatively, define the set-specific marginal false discovery rate for set k to be

$$\text{smFDR}_k(\boldsymbol{\delta}) = \frac{\mathbb{E}(\sum_{\ell \in \mathcal{S}_k} \sum_{\ell' \neq \ell} C_{\ell\ell'})}{\mathbb{E}(\sum_{\ell \in \mathcal{S}_k} R_\ell)}, \quad k = 1, \ldots, K, \tag{3}$$

which measures the proportion of misclassifications only for the kth set of interest. This reduces to the standard marginal false discovery rate when distinguishing between only $\cup_{\ell \neq k} \mathcal{S}_\ell$ and $\mathcal{S}_k$.

2.2. Signal classification problems

The two measures of false discovery lead to two different signal classification problems.

Definition 1 (Total). *The signal classification problem under total error is to find the $\boldsymbol{\delta}$ that*

$$\text{maximizes tETP}(\boldsymbol{\delta}) \text{ subject to } \text{tmFDR}(\boldsymbol{\delta}) \leq \alpha \tag{4}$$

for a given error level $0 < \alpha < 1$.

DEFINITION 2 (SET-SPECIFIC). *The signal classification problem under set-specific error is to find the $\boldsymbol{\delta}$ that*

$$\text{maximizes } \text{TETP}(\boldsymbol{\delta}) \text{ subject to } \text{SMFDR}_k(\boldsymbol{\delta}) \leq \alpha_k \tag{5}$$

for given error levels $0 < \alpha_1, \ldots, \alpha_K < 1$.

When $K = 1$, i.e., $\mathcal{S}_1$ is the only set of signal classes of interest, problems (4) and (5) coincide. In this case, signal classification reduces to the usual multiple testing framework, albeit with non-standard null and alternative distributions, and some special cases have been previously studied (Andreassen et al., 2013; Chung et al., 2014; Heller and Yekutieli, 2014). In general, however, these two problems can give different classification rules.

The advantage of problem (5) is that the different α_k allow for fine error control over the different types of misclassifications. For example, if the X_{1i} come from a study with a very large sample size while the X_{2i} come from a much smaller study, it may be desirable to choose a more stringent α_k when classifying features into class 1 of Table 2, as compared to class 2. However, it may not always be clear how the α_k should be chosen, so problem (4) offers total error control at a single error level. It is straightforward to show that the optimal rule of problem (5) is also a feasible solution to problem (4) at level $\alpha = \max_k \alpha_k$, though it may not maximize the total expected number of true positives in problem (4).

3. Proposed methods

3.1. Oracle procedures

Similar to the two-groups model for a single sequence of multiple tests (Sun and Cai, 2007), let the signal indicators $(\theta_{1i}, \theta_{2i})$ be independent and identically distributed across features i. Since in many applications the test statistics X_{1i} and X_{2i} arise from independent datasets, assume that they are independent conditional on θ_{1i} and θ_{2i}. Finally, let $F_{j0}(x)$ and $F_{j1}(x)$ denote the distribution functions of X_{ji} conditional on $\theta_{ji} = 0$ and $\theta_{ji} = 1$, respectively, where F_{j0} is known. Then the test statistics are distributed according to the four-group model

$$(X_{1i}, X_{2i}) \sim \sum_{\ell=0}^{3} \pi_\ell F_{1\ell_1} F_{2\ell_2}, \tag{6}$$

where for $\ell \in \{0, \ldots, 3\}$, ℓ_1 equals the value of θ_{1i} for signals in class ℓ and ℓ_2 equals the value of θ_{2i}. For example, from Table 1, $\ell = 2$ implies $\ell_1 = 1$ and $\ell_2 = 0$. Finally, $\pi_\ell = \mathbb{P}(\theta_{1i} = \ell_1, \theta_{2i} = \ell_2)$.

It is easy to check that the total error control problem (4) is equivalent to maximizing

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{m} I(\delta_i = k)\{1 - T_k^{OR}(X_{1i}, X_{2i})\}\right]$$

subject to

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{i=1}^{m} I(\delta_i = k)\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha\}\right] \leq 0,$$

where

$$T_k^{OR}(x_1, x_2) = \frac{\sum_{\ell \notin \mathcal{S}_k} \pi_\ell f_{1\ell_1}(x_1) f_{2\ell_2}(x_2)}{\sum_{\ell=0}^{3} \pi_\ell f_{1\ell_1}(x_1) f_{2\ell_2}(x_2)} \tag{7}$$

and f_{j0} and f_{j1} are the densities corresponding to F_{j0} and F_{j1}.

This optimization problem can be solved by minimizing the Lagrangian

$$\begin{aligned} L_T(\lambda, \boldsymbol{\delta}) &= \textstyle\sum_{k=1}^{K} \sum_{i=1}^{m} I(\delta_i \neq k)\{1 - T_k^{OR}(X_{1i}, X_{2i})\} \\ &+ \textstyle\sum_{k=1}^{K} \sum_{i=1}^{m} \lambda I(\delta_i = k)\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha\}, \end{aligned}$$

since any $\boldsymbol{\delta}$ that minimizes the $L(\lambda, \boldsymbol{\delta})$ conditional on the observed test statistics will also minimize $\mathbb{E}\{L_T(\lambda, \boldsymbol{\delta})\}$. Here the penalty parameter $\lambda > 0$ can be viewed as the relative cost of a misclassification into the wrong class of interest, compared to the cost of a misclassification into the null class $\mathcal{S}_0$. The λ and the error level α are related by the following result. This can be regarded as a generalization of the compound decision theoretic treatment of false discovery rate, proposed by Sun and Cai (2007), to signal classification.

PROPOSITION 1. *For any $\lambda > 0$, define the classification rule $\boldsymbol{\delta}_T^\lambda = (\delta_{T1}^\lambda, \ldots \delta_{Tm}^\lambda)$ where*

$$\delta_{Ti}^\lambda = \underset{k \in \{0,\ldots,K\}}{\arg\min} \sum_{k' \in \{1,\ldots,K\}, k' \neq k} \{1 - T_{k'}^{OR}(X_{1i}, X_{2i})\} + \lambda\{T_{k'}^{OR}(X_{1i}, X_{2i}) - \alpha\}. \tag{8}$$

(i) $\boldsymbol{\delta}_T^\lambda$ minimizes $\mathbb{E}\{L_T(\lambda, \boldsymbol{\delta})\}$;

(ii) Let $N_T^{OR}(\lambda) = \mathbb{E}\left[\sum_{k=1}^{K} I(\delta_{Ti}^\lambda = k)\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha\}\right]$ and define

$$\lambda^\star = \inf\{\lambda : N_T^{OR}(\lambda) \leq 0\}.$$

If $N_T^{OR}(0) \geq 0$ holds, then $N_T^{OR}(\lambda^\star) = 0$.

REMARK 1. Intuitively, an optimal decision rule should make the most of the misclassification error that it is allowed, in order to maximize the number of discoveries it makes. In other words, $\boldsymbol{\delta}_T^\lambda$ should achieve TMFDR $= \alpha$. The quantity $N_T^{OR}(\lambda)$ in Proposition 1 derives from the constraint on TMFDR and can be interpreted as a measure of how much of the allotted misclassification error has not been used up by $\boldsymbol{\delta}_T^\lambda$. The assumption $N_T^{OR}(0) \geq 0$ is indispensable, otherwise it would be possible for all features to be classified into one of the $\mathcal{S}_k$ but the total marginal false discovery rate to still be less than α. It can be shown that $N_T^{OR}(\lambda)$ is non-increasing in λ, so $N_T^{OR}(0) < 0$ would imply that for some values of α there may not exist any λ such that TMFDR$(\boldsymbol{\delta}_T^\lambda)$ exactly attains α. In this sense, the assumption $N_T^{OR}(0) \geq 0$ ensures that the nominal level α can be achieved exactly by some λ.

The oracle procedure $\boldsymbol{\delta}_T^\star = (\delta_{T1}^\star, \ldots, \delta_{Tm}^\star)$ for the total error control problem (4) can now be defined. Theorem 1 shows that $\delta_T^\star$ achieves the largest total expected number of true positives among all rules that can control the total marginal false discovery rate.

THEOREM 1. *With δ^{λ}_{Ti} and $\lambda^\star$ defined in Proposition 1, define*

$$\boldsymbol{\delta}^\star_T = (\delta^{\lambda^\star}_{T1}, \ldots, \delta^{\lambda^\star}_{Tm}).$$

If α satisfies $N^{OR}_T(0) \geq 0$ from Proposition 1(ii), then,

(i) $\text{TMFDR}(\boldsymbol{\delta}^\star_T) = \alpha$*;*

(ii) For any other classification rule $\boldsymbol{\delta}$ that satisfies $\text{TMFDR}(\boldsymbol{\delta}) \leq \alpha$*,*

$$\text{TETP}(\delta^\star_T) \geq \text{TETP}(\boldsymbol{\delta}).$$

Similarly, the constraints in the set-specific error control problem (5) can be equivalently expressed as

$$\mathbb{E}\left[\sum_{i=1}^{m} I(\delta_i = k)\{T^{OR}_k(X_{1i}, X_{2i}) - \alpha_k\}\right] \leq 0 \text{ for } k = 1, \ldots, K,$$

so problem (5) can be solved by minimizing the Lagrangian

$$\begin{aligned} L_S(\boldsymbol{\lambda}, \boldsymbol{\delta}) \ = \ & \textstyle\sum_{k=1}^{K}\sum_{i=1}^{m} I(\delta_i \neq k)\{1 - T^{OR}_k(X_{1i}, X_{2i})\} \\ + \ & \textstyle\sum_{k=1}^{K}\sum_{i=1}^{m} \lambda_k I(\delta_i = k)\{T^{OR}_k(X_{1i}, X_{2i}) - \alpha_k\}. \end{aligned}$$

Proposition 2 makes the connection between the set-specific error control problem (5) and a minimization problem with the objective function $\mathbb{E}\{L_S(\boldsymbol{\lambda}, \boldsymbol{\delta})\}$.

PROPOSITION 2. *For any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ with $\lambda_k > 0$, define the classification rule $\boldsymbol{\delta}^{\boldsymbol{\lambda}}_S = (\delta^{\boldsymbol{\lambda}}_{S1}, \ldots \delta^{\boldsymbol{\lambda}}_{Sm})$ where*

$$\delta^{\boldsymbol{\lambda}}_{Si} = \underset{k\in\{0,\ldots,K\}}{\arg\min} \sum_{k'\in\{1,\ldots,K\}, k'\neq k} \{1 - T^{OR}_{k'}(X_{1i}, X_{2i})\} + \lambda_k\{T^{OR}_k(X_{1i}, X_{2i}) - \alpha_k\}. \tag{9}$$

(i) $\boldsymbol{\delta}^{\boldsymbol{\lambda}}_S$ minimizes $\mathbb{E}\{L_S(\boldsymbol{\lambda}, \boldsymbol{\delta})\}$;

(ii) Let $N^{OR}_k(\boldsymbol{\lambda}) = \mathbb{E}\left[I(\delta^{\boldsymbol{\lambda}}_{Si} = k)\{T^{OR}_k(X_{1i}, X_{2i}) - \alpha_k\}\right]$ and define

$$\check{\lambda}_{k,n} = \inf\{\lambda_k \leq \check{\lambda}_{k,n-1} : N^{OR}_k(\check{\boldsymbol{\lambda}}_{k,n-1}) \leq 0\}, k = 1, \ldots K,$$

where $n \geq 1$, $\check{\lambda}_{k,0} = \infty$, $\check{\boldsymbol{\lambda}}_{k,n-1}$ is the $\boldsymbol{\lambda}$ with $\lambda_{k'} = \check{\lambda}_{k',n-1}$, $k' \neq k$. Suppose that $\alpha_k + \alpha_{k'} \leq 1$ holds for any $k \neq k' \in \{1, \ldots, K\}$ and $\mathbf{0} \in \{(N^{OR}_1(\boldsymbol{\lambda}), \ldots, N^{OR}_K(\boldsymbol{\lambda})) : \boldsymbol{\lambda} \in \{\Re_+ \cup \{0\}\}^K\}$. Then, sequence $\{\check{\lambda}_{k,n}, n \geq 1\}$ is convergent and $N^{OR}_k(\boldsymbol{\lambda}^\star) = 0$ for all $k = 1, \ldots, K$ where $\boldsymbol{\lambda}^\star = (\lambda^\star_1, \ldots, \lambda^\star_K)$ and $\lambda^\star_k = \lim_{n\to\infty} \check{\lambda}_{k,n}$.

REMARK 2. Proposition 2(ii) plays the same role as the condition on $N^{OR}_T(0)$ in Proposition 1(ii). In the set-specific error control problem (5), not all error levels $\alpha_1, \ldots, \alpha_K$ correspond to a $\boldsymbol{\lambda}$ such that $\text{SMFDR}_k(\boldsymbol{\delta}^{\boldsymbol{\lambda}}_S)$ attains α_k for all $k = 1, \ldots, K$. The restriction that $\alpha_k + \alpha_{k'} < 1$ for any $k \neq k' \in \{1, \ldots, K\}$ is mild since it includes a wide range of choices for $\boldsymbol{\alpha}$. For example, it allows $0 \leq \alpha_k \leq 1/2, k = 1, \ldots, K$, which is adequate for many applications.

The oracle procedure $\boldsymbol{\delta}^\star_S = (\delta^\star_{S1}, \ldots, \delta^\star_{Sm})$ for the set-specific error control problem (5) can now be defined. Theorem 2 shows that $\boldsymbol{\delta}^\star_S$ achieves the largest total expected number of true positives among all rules that can control the set-specific marginal false discovery rates.

THEOREM 2. *With $\boldsymbol{\lambda}^\star$ and $\boldsymbol{\lambda}^\star$ defined in Proposition 2, define*

$$\boldsymbol{\delta}_S^\star = (\delta_{S1}^{\boldsymbol{\lambda}^\star}, \ldots, \delta_{Sm}^{\boldsymbol{\lambda}^\star}).$$

If $\alpha_1, \ldots, \alpha_K$ satisfy the conditions in Proposition 2(ii), then

(i) $\mathrm{SMFDR}_k(\boldsymbol{\delta}_S^\star) = \alpha_k$ *for* $k = 1, \ldots, K$*;*

(ii) For any other classification rule $\boldsymbol{\delta}$ that satisfies $\mathrm{SMFDR}_k(\boldsymbol{\delta}) \le \alpha_k$, $k = 1, \ldots, K$.

$$\mathrm{TETP}(\boldsymbol{\delta}_S^\star) \ge \mathrm{TETP}(\boldsymbol{\delta}).$$

When class 3 is the only class of interest, the two oracle methods described in Theorems 1 and 2 are identical to the oracle method proposed by Heller and Yekutieli (2014). Otherwise they are different in general, and this will be further explored in simulations in Section 4.

3.2. Data-driven procedures

The oracle procedures described in the previous section cannot be implemented in practice because they are functions of $T_k^{OR}(X_{1i}, X_{2i})$ defined in (7), which depends on the unknown mixture proportions π_ℓ and non-null densities f_{j1}. However, the T_k^{OR} can be estimated by first defining the marginal proportions $\pi_{j\ell_j} = \mathbb{P}(\theta_{ji} = \ell_j)$ and the marginal densities $f_j(x) = \pi_{j0} f_{j0}(x) + \pi_{j1} f_{j1}(x)$ and rewriting

$$T_k^{OR}(x_1, x_2) = \frac{\sum_{\ell \notin \mathcal{S}_k} \pi_\ell / (\pi_{1\ell_1}\pi_{2\ell_2}) \{\pi_{1\ell_1} f_{1\ell_1}(x_1)/f_1(x_1)\}\{\pi_{2\ell_2} f_{2\ell_2}(x_2)/f_2(x_2)\}}{\sum_{\ell=0}^{3} \pi_\ell / (\pi_{1\ell_1}\pi_{2\ell_2}) \{\pi_{1\ell_1} f_{1\ell_1}(x_1)/f_1(x_1)\}\{\pi_{2\ell_2} f_{2\ell_2}(x_2)/f_2(x_2)\}}.$$

Next, estimates $\hat{\pi}_{j\ell_j}$ for the marginal proportions can be obtained by applying the method of Jin and Cai (2007) to the statistics $\Phi^{-1}\{F_{j0}(X_{ji})\}$, and estimates $\hat{f}_j(x)$ of the marginal densities can be obtained using standard kernel-based methods (Silverman, 1986). The likelihood ratios in $T_k^{OR}(x_1, x_2)$ can then be estimated because

$$\hat{\pi}_{j0} f_{j0}(x)/\hat{f}_j(x) \xrightarrow{p} \pi_{j0} f_{j0}(x)/f_j(x), \; 1 - \hat{\pi}_{j0} f_{j0}(x)/\hat{f}_j(x) \xrightarrow{p} \pi_{j1} f_{j1}(x)/f_j(x).$$

In practice, each estimated likelihood ratio is set equal to 1 if its calculated value exceeds 1. Finally, an estimate $\hat{\pi}_0$ of $\mathbb{P}(\theta_{1i} = 0, \theta_{2i} = 0)$ can be obtained by applying the method of Jin and Cai (2007) to the statistics

$$\Phi^{-1}\left(G_{\chi^2,2}\left[\Phi^{-1}\{F_{10}(X_{1i})\}^2 + \Phi^{-1}\{F_{20}(X_{2i})\}^2\right]\right),$$

where $G_{\chi^2,2}$ is the distribution function of a chi-square random variable with two degrees of freedom, and estimates of the other π_ℓ can be calculating using $\hat{\pi}_0$ and the $\hat{\pi}_{j\ell_j}$. The above estimates can then be inserted into T_k^{OR} to give the plug-in statistic $\hat{T}_k$, which is set equal to 1 if its calculated value exceeds 1.

The data-driven procedure that solves the total error control problem (4) can be constructed as follows. First define $\hat{\delta}_{Ti}^\lambda$ to be the solution to the total error minimization problem (8) with $\hat{T}_k$ in place of T_k^{OR}. Next, define $\hat{N}_T(\lambda) = m^{-1} \sum_{i=1}^m \sum_{k=1}^K I(\hat{\delta}_{Ti}^\lambda =$

Table 3. Data-driven Algorithm For Total Error Control

Let $\hat{T}_{\min}(x_1, x_2) = \min_k \hat{T}_k(x_1, x_2)$, and let $\hat{T}_{\min}^{(i)}$ be the ordered statistics $\hat{T}_{\min}(x_{1i}, x_{2i})$ and $\hat{\delta}^{\star}_{T(i)}$, $T_k^{(i)}$ be the corresponding decision functions and testing statistics. Define $r = \max\{j : 1/j \sum_{i=1}^{j} \hat{T}_{\min}^{(i)} \leq \alpha\}$. Then, $$\hat{\delta}^{\star}_{T(i)} = \begin{cases} k, & i \leq r \text{ and } \hat{T}_{\min}^{(i)} = \hat{T}_k^{(i)} \\ 0, & i > r \end{cases}$$

$k)\{\hat{T}_k(X_{1i}, X_{2i}) - \alpha\}$. This expression can be simplified because it can be seen from the definition of the oracle total error control rule in (8) that for $k = 1, \ldots, K$

$$I(\hat{\delta}^{\lambda}_{Ti} = k) = I\left\{\hat{T}_k \leq \alpha + \frac{1-\alpha}{\lambda+1}, \quad \hat{T}_{k'} < \min_{k' \neq k} \hat{T}_{k'}\right\}.$$

Thus the $I(\hat{\delta}^{\lambda}_{Ti} = k)$ in $\hat{N}_T(\lambda)$ can be replaced with the right-hand side of the above equation. Finally define

$$\hat{\lambda}^{\star} = \inf\{\lambda : \hat{N}_T(\lambda) \leq 0\}. \tag{10}$$

Then the data-driven classification rule that solves problem (4) is defined to be

$$\hat{\boldsymbol{\delta}}^{\star}_T = (\hat{\delta}^{\hat{\lambda}^{\star}}_{T1}, \ldots, \hat{\delta}^{\hat{\lambda}^{\star}}_{Tm}),$$

and a simple algorithm for its calculation is presented in Table 3. This algorithm is similar to multiple testing procedures that use local false discovery rates (Sun and Cai, 2007).

Theorem 3 shows that the data-driven $\hat{\boldsymbol{\delta}}^{\star}_T$ is asymptotically valid and optimal and thus can be very useful in practice.

THEOREM 3. *If*

(C1) $\hat{\pi}_{j\ell} \xrightarrow{p} \pi_{j\ell}$, $j = 1, 2$ *and* $\ell = 0, 1$;

(C2) $\hat{\pi}_{j0} f_{j0}(x_{ji})/\hat{f}_j(x_{ji}) \xrightarrow{p} \pi_{j0} f_{j0}(x_{ji})/f_j(x_{ji})$, *uniformly for all* i, $j = 1, 2$

hold, then (i) $\text{TMFDR}(\hat{\boldsymbol{\delta}}^{\star}_T) = \alpha + o(1)$, *and (ii)* $\text{TETP}(\hat{\boldsymbol{\delta}}^{\star}_T)/\text{TETP}(\boldsymbol{\delta}^{\star}_T) = 1 + o(1)$.

The data-driven rule that solves the set-specific error control problem (5) can be similarly developed. Let $\hat{\delta}^{\boldsymbol{\lambda}}_{Si}$ be the solution to the set-specific error minimization problem (9) with $\hat{T}_k$ in place of T_k^{OR}, $\hat{N}_k(\boldsymbol{\lambda}) = 1/m \sum_{i=1}^{m} I(\hat{\delta}^{\boldsymbol{\lambda}}_{Si} = k)[\hat{T}_k(x_{1i}, x_{2i}) - \alpha_k]$ and construct a sequence $\{\hat{\lambda}_{k,n}, n \geq 1\}$ that satisfies

$$\hat{\lambda}_{k,n} = \inf\{\lambda \leq \hat{\lambda}_{k,n-1} : \hat{N}_k(\hat{\boldsymbol{\lambda}}_{k,n-1}) \leq 0\}, \tag{11}$$

where $\hat{\lambda}_{k,0} = \infty$, $\hat{\boldsymbol{\lambda}}_{k,n-1}$ is the $\boldsymbol{\lambda}$ with $\lambda_k = \lambda$ and $\lambda_{k'} = \hat{\lambda}_{k',n-1}$ for $k' \neq k$. Like sequence $\{\check{\lambda}_{k,n}, n \geq 1\}$ defined in Proposition 2, the convergence of the sequence $\{\hat{\lambda}_{k,n}, n \geq 1\}$ can be proved similarly. Let $\hat{\lambda}^{\star}_k$ be the value to which $\{\hat{\lambda}_{k,n}, n \geq 1\}$ converges. Then the data-driven procedure that solves problem (5) can be defined to be

$$\hat{\boldsymbol{\delta}}^{\star}_S = (\hat{\delta}^{\hat{\boldsymbol{\lambda}}^{\star}}_{S1}, \ldots, \hat{\delta}^{\hat{\boldsymbol{\lambda}}^{\star}}_{Sm}),$$

Table 4. Data-driven Algorithm For Set-Specific Error Control

• Step 1. Let $\hat{T}_k^{(i)}$ be the ordered statistics $\hat{T}_k(x_{1i}, x_{2i})$ and determine the initial threshold vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, where for each $k \in \{1, \ldots, K\}$, $\lambda_k = (1-\alpha_k)/(\hat{T}_k^{(r_k)} - \alpha_k) - 1$, and $r_k = \max\{j : 1/j \sum_{i=1}^{j} \hat{T}_k^{(i)} \le \alpha_k\}$.
• Step 2. For each k, calculate $\hat{N}_k(\boldsymbol{\lambda})$ and $\hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,r_k+1})$ where $\tilde{\boldsymbol{\lambda}}_{k,j} = (\tilde{\lambda}_{1,j}, ..., \tilde{\lambda}_{K,j})$ with $\tilde{\lambda}_{k,j} = (1-\alpha_k)/(\hat{T}_k^{(j)} - \alpha_k) - 1$ and $\tilde{\lambda}_{k',j} = \lambda_{k'}$, $k' \ne k$. If $\hat{N}_k(\boldsymbol{\lambda}) \le 0$ and $\hat{N}_k(\tilde{\boldsymbol{\lambda}}_{r_k+1}) > 0$ hold for all k, $\boldsymbol{\lambda}$ is the desired threshold vector. Otherwise go to Step 3.
• Step 3. Let $\tilde{r}_k = \max\{j \ge r_k : \hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,j}) \le 0\}$ and reset $r_k = \tilde{r}_k$. Then, update the $\boldsymbol{\lambda}$ in Step 1 and repeat Steps 2 and 3 till this loop is terminated. The $\boldsymbol{\lambda}$ in the last iteration is the desired $\hat{\boldsymbol{\lambda}}^\star$.
• Step 4. Apply $\hat{\boldsymbol{\lambda}}^\star$ to equation (9) with $\hat{T}_k$ in place of T_k^{OR} to obtain the classification rule $\hat{\boldsymbol{\delta}}_S^\star = (\hat{\delta}_{S1}^\star, \ldots, \hat{\delta}_{Sm}^\star)$.

where $\hat{\boldsymbol{\lambda}}^\star = (\hat{\lambda}_1^\star, \ldots, \hat{\lambda}_K^\star)$.

A fast algorithm for calculating $\hat{\boldsymbol{\delta}}_S^\star$ is provided in Table 4, which shows that the algorithm can be regarded as a stage-wise multiple testing procedure for identifying set-specific signals. That is, in each stage, or each iteration of steps 2 and 3, a two class multiple testing procedure is performed for each of the K sets of interest in turn. This process terminates when the estimated threshold sequences converge.

Theorem 4 shows that the data-driven $\hat{\boldsymbol{\delta}}_S^\star$ is asymptotically valid and optimal.

THEOREM 4. *If Conditions (C1) and (C2) stated in Theorem 3 hold, then for all* $k \in \{1, \ldots, K\}$, *(i)* $\text{SMFDR}_k(\hat{\boldsymbol{\delta}}_S^\star) = \alpha_k + o(1)$, *and (ii)* $\text{TETP}(\hat{\boldsymbol{\delta}}_S^\star)/\text{TETP}(\boldsymbol{\delta}_S^\star) = 1 + o(1)$

3.3. Adjusted separate discovery procedure

Based on the proposed data-driven test statistics $\hat{T}_k^{OR}(x_{1i}, x_{2i})$, the separate discovery procedure for signal classification, described in Section 1.2, can be adjusted so that it provides valid control of the two types of misclassification errors introduced in this paper. Without loss of generality, this section considers the case $K = 3$ for illustration.

Let P_{ji} be the p-value of the ith feature in sequence j. The separate discovery procedure would set $\delta_i = 1$ if $P_{1i} > c_1$ and $P_{2i} \le c_2$, $\delta_i = 2$ if $P_{1i} \le c_1$ and $P_{2i} > c_2$, and $\delta_i = 3$ if $P_{1i} \le c_1$ and $P_{2i} \le c_2$, for some cutoffs c_j such that the marginal false discovery rate for sequence j attains α_j. As mentioned in Section 1.2, the separate discovery procedure cannot control the TMFDR and the SMFDR at desired nominal levels. This cannot be remedied by merely choosing different values for α_j. Instead, the key difficulty is that the non-discovery classifications in each sequence are unreliable.

This limitation can be overcome by employing two different cutoffs for each study. Specifically, set $\delta_i = 1$ if $P_{1i} > c_{11}$ and $P_{2i} \le c_{12}$, $\delta_i = 2$ if $P_{1i} \le c_{21}$ and $P_{2i} > c_{22}$, and $\delta_i = 3$ if $P_{1i} \le c_{31}$ and $P_{2i} \le c_{32}$, where the c_{kj} can all be unequal. Then the separate discovery procedure can be adjusted by finding the c_{kj} such that $\sum_{i=1}^m I(\delta_i = k)[\hat{T}_k^{OR}(x_{1i}, x_{2i}) - \alpha_k] \approx 0$ for the set-specific error control problem, and $\sum_{i=1}^m \sum_{k=1}^K I(\delta_i = k)[\hat{T}_k^{OR}(x_{1i}, x_{2i}) - \alpha] \approx 0$ for the total error control problem. This new procedure may lead to some features being classified into more than one set of interest. Appendix C in the supplementary file provides an algorithm to find the cutoffs, as well as details for resolving overlapping classifications.

This adjusted separate discovery procedure can approximately control the different misclassification errors. However, unlike the other procedures proposed in this paper, it is computationally intensive and that its cutoffs are not optimal in the sense having the largest TETP values.

4. Simulations

This section investigates the numerical performance of the proposed oracle and data-driven procedures. Pairs of test statistics (X_{1i}, X_{2i}) for $i = 1, \ldots, m$ were generated for $m = 20,000$ features according to the four-group model (6), with class labels defined as in Table 1. Specifically, the null and alternative density functions were

$$f_{j0}(x) = \phi(x), \quad f_{j1}(x) = \phi\left(\frac{x - \mu_j}{\sigma_j}\right)$$

for sequences $j = 1, 2$, where $\phi(x)$ is the standard normal density. The signal standard deviation σ_j was set to $4/10^{1/2}$ throughout while the mean signal strength μ_j, signal proportions, and nominal total or set-specific marginal false discovery rates were varied across simulation settings. All settings were simulated 200 times.

The following procedures were compared:

(a) The oracle and data-driven procedures for the total and set-specific error control problems proposed in this paper.

(b) The method of Heller and Yekutieli (2014). Though originally developed to classify features into either $\mathcal{S}_0 = \{0, 1, 2\}$ or $\mathcal{S}_1 = \{3\}$, it can easily be modified to accommodate any set $\mathcal{S}_1$. However, it cannot be extended to the general classification problem when there is more than one set of classes of interest.

(c) The separate discovery approach based on p-values, using the procedure proposed by Genovese and Wasserman (2004). For the total error control problem, the error levels in each individual sequence are all set to equal the desired nominal total marginal false discovery rate (2). For the set-specific error control problem, the error levels in each individual sequence are all set to equal the average of the desired nominal set-specific marginal false discovery rates.

(d) The adjusted separate discovery approach based on p-values, given in Section 3.3.

Four sets of simulations were conducted. The first setting considered total marginal error control for the binary classification problem of identifying features that are only significant in one of the two studies, in order words classifying features into either $\mathcal{S}_0 = \{0, 3\}$ or $\mathcal{S}_1 = \{1, 2\}$. The signal strengths μ_1 and μ_2 were varied between 2.8 to 3.7, signal proportions were set as $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.8 - h, h, h, 0.2 - h)$ for h varying between 0.05 and 0.1, and nominal total marginal false discovery rates (2) were varied between 0.05 and 0.2.

Results in Figure 1 show that the oracle, data-driven, and adjusted separate discovery methods were all able to control the total marginal false discovery rate at the desired

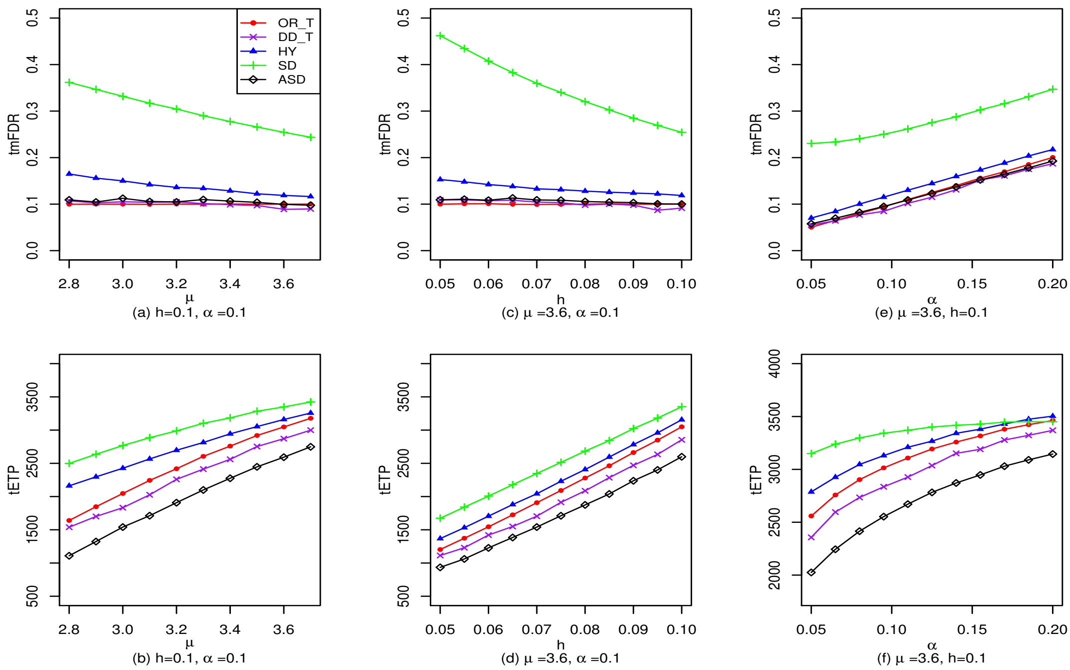

Fig. 1. Total marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0, 3\}$ or $\mathcal{S}_1 = \{1, 2\}$. tmFDR = empirical total marginal false discovery rate (2); tETP = empirical total number of true positives (1); OR_T: oracle total control error procedure from Theorem 1; DD_T = data-driven total error control procedure from Theorem 3; HY = method of Heller and Yekutieli (2014); SD = separate discovery procedure; ASD = adjusted separate discovery procedure from Section 3.3.

nominal level. Among these, the oracle procedure had the most power, as expected, but the data-driven procedure performed almost as well. The method of Heller and Yekutieli (2014) was slightly too liberal in controlling the false discovery rate when the signals were weak and there were few signals in $\mathcal{S}_1$, but otherwise performed as well as the proposed data-driven procedure in most situations. With stronger signals, more signals in $\mathcal{S}_1$, and higher nominal total marginal false discovery rates, all methods increased in power, and the difference between the oracle and data-driven procedures decreased.

The second simulation setting also considered total marginal error control, but for classifying signals into either $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, or $\mathcal{S}_3 = \{3\}$. All parameters were set as in the previous simulation setting except with $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (1 - h, h/3, h/3, h/3)$ for h varying between 0.21 and 0.36. The method of Heller and Yekutieli (2014) cannot be applied to this multiclass problem, but the other methods followed the same trends as before, as shown in Figure 2.

The next set of simulations studied the set-specific error control for this multiclass classification problem. Signal strengths were varied between 2.8 and 3.7, signal proportions were set to $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (1 - h, h/3, h/3, h/3)$ with h varying between 0.21 and 0.3, and nominal set-specific marginal false discovery rates α_k (3) were varied

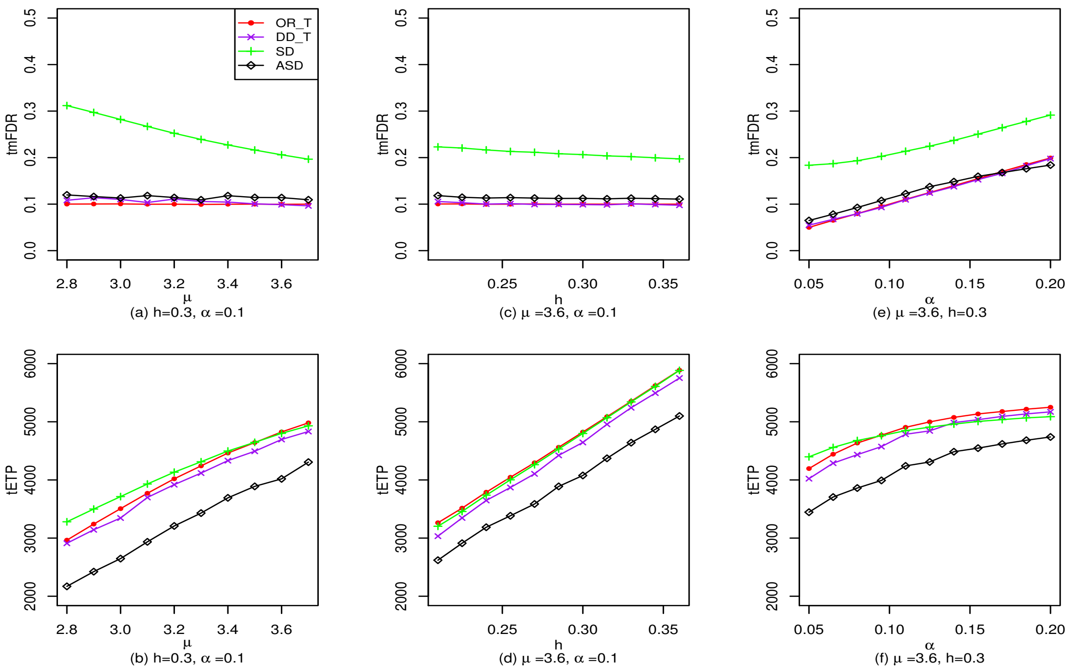

Fig. 2. Total marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, or $\mathcal{S}_3 = \{3\}$. tmFDR = empirical total marginal false discovery rate (2); tETP = empirical total number of true positives (1); OR_T: oracle total error control procedure from Theorem 1; DD_T = data-driven total error control procedure from Theorem 3; SD = separate discovery procedure; ASD = adjusted separate discovery procedure from Section 3.3.

between 0.05 and 0.2. For simplicity, all α_k were set to be equal for $k = 1, \dots, 3$. Results are plotted in Figure 3. The oracle and data-driven procedures again had the nearly same performance, and uniformly dominated the adjusted separate discovery procedure.

Finally, simulations were conducted to explore the relationship between the total (4) and set-specific (5) error control problems for the multiclass problem with sets $\mathcal{S}_k = \{k\}$ for $k = 0, \dots, 3$. Signal strengths were either set equal to μ in both sequences, or to $\mu - 1$ in sequence 1 and $\mu + 2$ in sequence 2, with μ varying between 4 and 5. Signal proportions among the sets $\mathcal{S}_k$ of interest were either uniform, with $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.7, 0.1, 0.1, 0.1)$, or non-uniform, equal to $(0.7, 0.05, 0.05, 0.2)$.

The first row of Figure 4 reports the empirical set-specific false discovery rates acheived by the oracle total error control procedure, where the nominal total error was set to 0.1. The realized set-specific errors differed across the sets of interest. It was always lowest for $\mathcal{S}_3$ and highest for the set $\mathcal{S}_2$ that contained features that were significant only in test statistic sequence 1, which in these simulations had the higher signal strength. The differences between the realized set-specific errors increased as the signal strengths and the non-uniformity of the signal proportions increased.

The second row of Figure 4 reports the empirical total false discovery rates of the oracle set-specific error control procedure, where the nominal set-specific errors were set

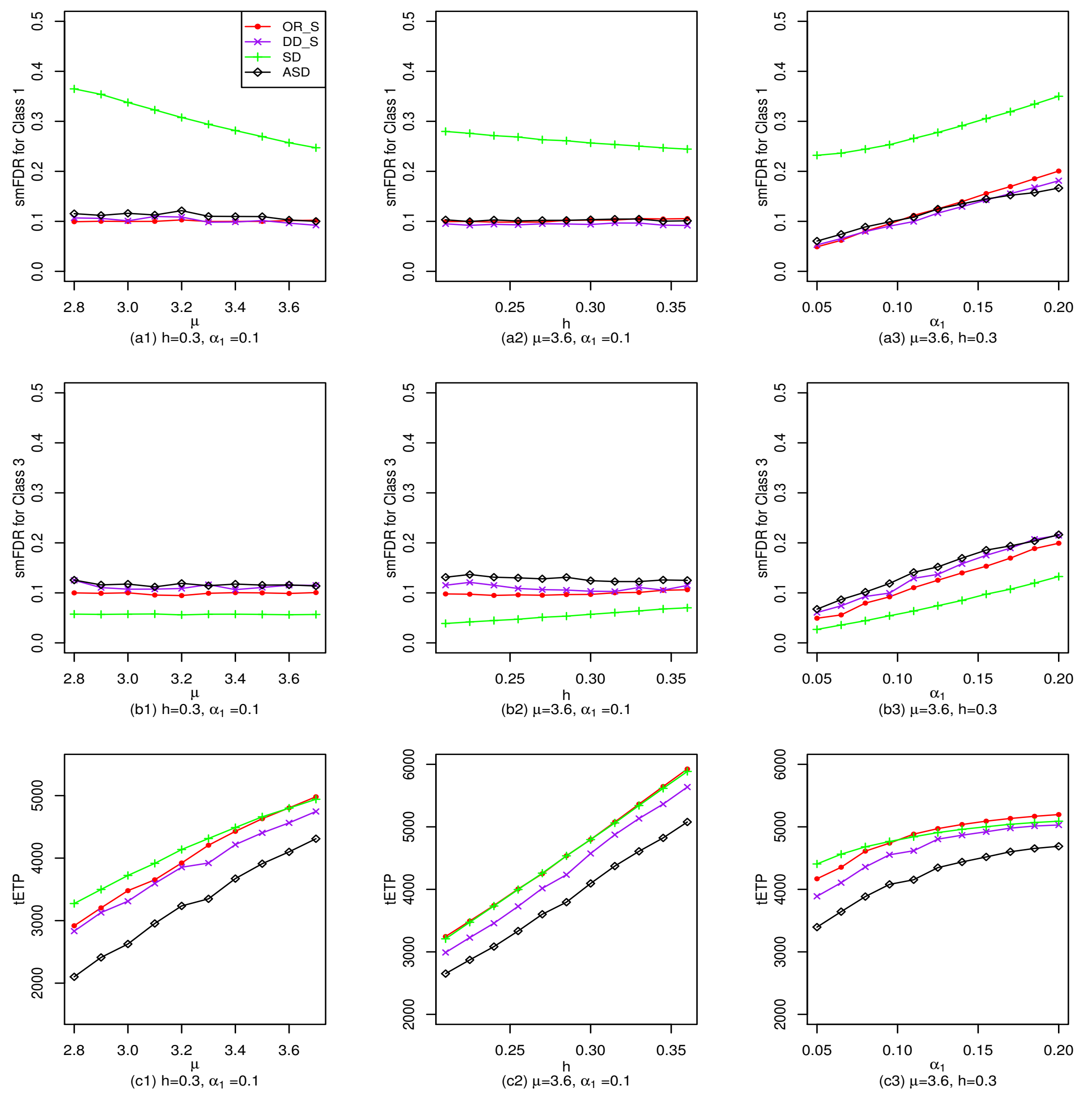

Fig. 3. Set-specific marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, or $\mathcal{S}_3 = \{3\}$. Due to symmetry ($\alpha_1 = \alpha_2 = \alpha_3$), plots of the set-specific marginal false discovery rate for classification into $\mathcal{S}_2$ are identical to those for classification into $\mathcal{S}_1$ and therefore are omitted. smFDR = empirical set-specific marginal false discovery rate (3); tETP = empirical total number of true positives (1); OR_S: oracle set-specific error control procedure from Theorem 2; DD_S = data-driven set-specific error control procedure from Theorem 4; SD = separate discovery procedure; ASD = adjusted separate discovery procedure from Section 3.3.

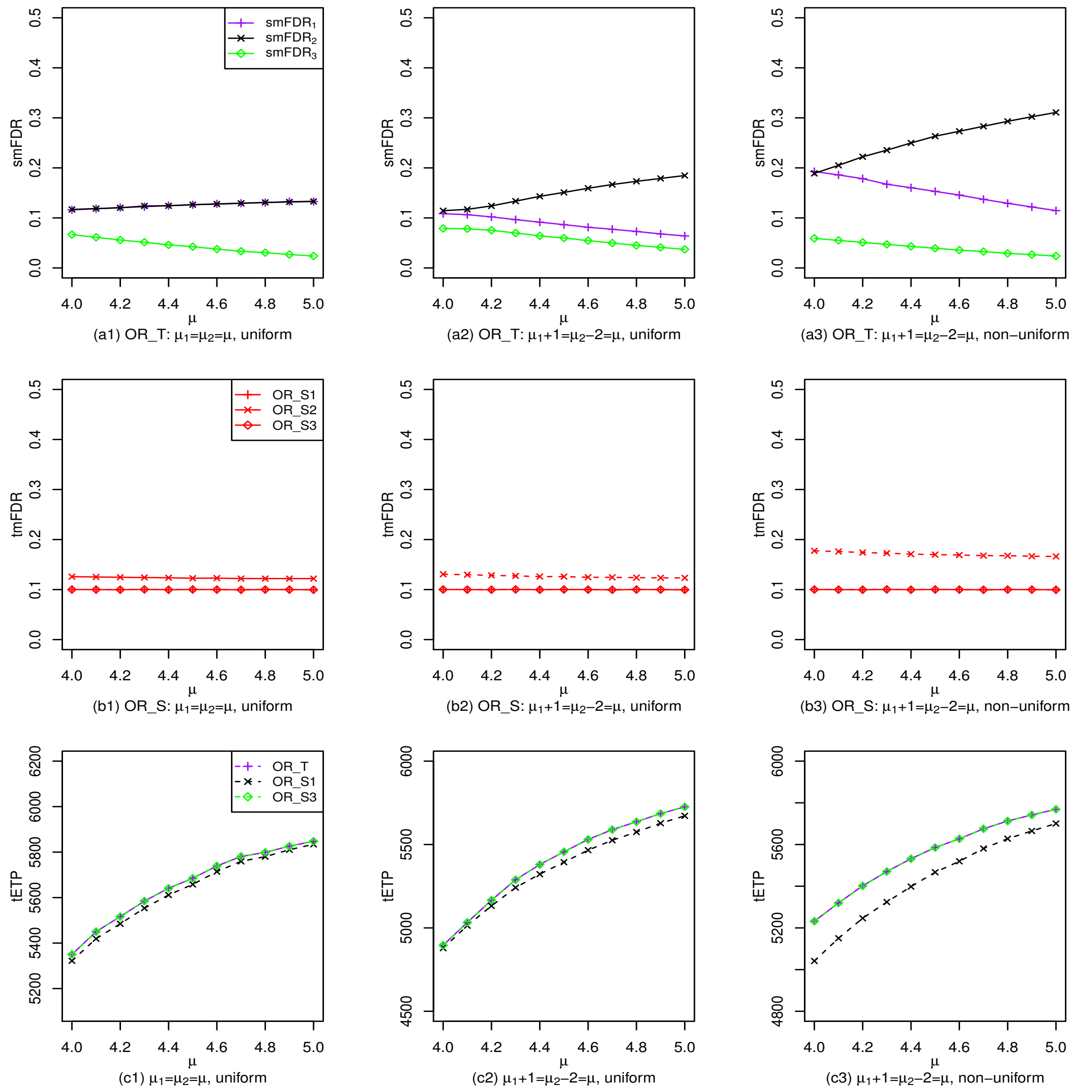

Fig. 4. Comparison of total and set-specific error control problems for classifying signals into $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, or $\mathcal{S}_3 = \{3\}$. smFDR = empirical set-specific marginal false discovery rate (3); smFDR_k = the smFDR for class $k = 1, 2, 3$; tmFDR = empirical total marginal false discovery rate (2); OR_T: oracle procedure from Theorem 1 with $\alpha = 0.1$; OR_S: oracle set-specific error control procedure from Theorem 2; OR_S1: the OR_S procedure with $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$; OR_S2: the OR_S procedure with $\alpha_1 = 2\alpha_2 = \alpha_3/2 = 0.1$; OR_S3: the OR_S procedure with the nominal set-specific errors induced by the OR_T procedure.

to $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$, $\alpha_1 = 2\alpha_2 = \alpha_3/2 = 0.1$, or the set-specific errors that were induced by the oracle total control procedure with $\alpha = 0.1$. These plots show that the oracle set-specific procedure was also able to control the total error when all α_k equaled the desired total nominal level, or when they equaled the induced error levels. With uniform signal proportions, the oracle set-specific procedure controlled the total error at roughly the average of the nominal set-specific errors.

The third row of Figure 4 reports the realized average number of true positives for both oracle procedures. For the oracle total error control procedure, the nominal total error was set to 0.1. To conduct a fair comparison, the nominal set-specific errors for the oracle set-specific error control procedure were set as either $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$, or the set-specific errors induced by running the oracle total control procedure at $\alpha = 0.1$. The plots show that the oracle set-specific error control procedure with induced error levels was as powerful as the oracle total error procedure, and more powerful than when $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$. This trend was more pronounced with larger signal strengths and non-uniform signal proportions.

5. Application to psychiatric genetics

The proposed methods were applied to study the genetic architectures of bipolar disorder and schizophrenia. A better understanding of the genetic differences and similarities between these diseases could lead to more effective diagnosis and treatment. To explore this question, Ruderfer et al. (2014) performed two large genome-wide association studies, one of bipolar disorder, with 10,410 cases and 10,700 controls, and the other of schizophrenia, with 9,369 cases and 8,723 controls. These studies were comprised of completely independent samples, in particular, they did not share any control subjects. Summary Z-scores are available from the website of the Psychiatric Genomics Consortium. Before analyses reported below, the SNPs were first pruned at a linkage disequilibrium r^2 threshold of 0.5, using genotype data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015) as a reference panel. There were 439,040 variants remaining after pruning.

The data-driven total error control procedure was first applied to classify these SNPs into sets $\mathcal{S}_0$, containing SNPs that were not significant in either study, $\mathcal{S}_1$, containing SNPs associated only with schizophrenia, $\mathcal{S}_2$, containing SNPs associated only with bipolar disorder, and $\mathcal{S}_3$, containing SNPs significant in both studies. The nominal total marginal false discovery rate was set to 0.05. The first row of Table 5 reports the number of SNPs classified to each of the three classes. The majority of the discovered SNPs were classified into $\mathcal{S}_3$, consistent with previous work showing that bipolar disorder and schizophrenia have closely related genetic etiologies (Huang et al., 2010; Cross-Disorder Group of Psychiatric Genomics Consortium, 2013a,b).

In some cases, however, SNPs in $\mathcal{S}_3$ may not be of primary interest. For example, SNPs in $\mathcal{S}_1$ or $\mathcal{S}_2$ are more useful than SNPs in $\mathcal{S}_3$ for developing more accurate diagnostic procedures to differentiate patients with bipolar disorder from those with schizophrenia. Currently this differential diagnosis is difficult to perform, especially in the early stages of these disorders (Ruderfer et al., 2014). For the purpose of addressing this problem, capturing SNPS that belong to $\mathcal{S}_1$ and $\mathcal{S}_2$ is more important than finding SNPS in $\mathcal{S}_3$,

Table 5. Number of SNPs from Ruderfer et al. (2014) classified into different sets of interest. $\mathcal{S}_1$: SNPs associated only with schizophrenia; $\mathcal{S}_2$: SNPs associated only with bipolar disorder.

Method	Marginal false discovery rate	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$
Total	$\alpha = 0.05$	2	1	54
Set-specific	$\alpha_1 = 0.1, \alpha_2 = 0.1, \alpha_3 = 0.01$	4	2	8

Table 6. SNPs from Ruderfer et al. (2014) classified as being disease-specific, using the set-specific error control procedure with $\alpha_1 = \alpha_2 = 0.1$ and $\alpha_3 = 0.01$. $\mathcal{S}_1$: SNPs significantly associated only with schizophrenia; $\mathcal{S}_2$: SNPs significantly associated only with bipolar disorder; BIP: Z-score for bipolar disorder; SCZ: Z-score for schizophrenia.

Class $\mathcal{S}_1$			Class $\mathcal{S}_2$		
SNP	BIP	SCZ	SNP	BIP	SCZ
rs9273012	0.5391	5.2989	rs13166360	4.9054	-0.4711
rs1977	1.5329	5.0143	rs9788865	5.3086	1.0716
rs6932590	1.9759	5.2021			
rs1150753	1.3222	4.9154			

though all three classes remain of interest.

The proposed set-specific error control procedure can be applied to this type of setting. To illustrate, the method was applied with nominal set-specific marginal false discovery rates set to 0.10 for $\mathcal{S}_1$ and $\mathcal{S}_2$ and to 0.01 for $\mathcal{S}_3$. The more liberal thresholds for $\mathcal{S}_1$ and $\mathcal{S}_2$ allow for the discovery of more SNPs that are potentially diagnostically useful, and are offset by the more stringent threshold for $\mathcal{S}_3$. The second row of Table 5 shows that more disease-specific SNPs were indeed detected.

These SNPs, along with their Z-scores for the two diseases, are reported in Table 6. The four SNPs specific to schizophrenia are all located on chromosome 6 inside the major histocompatibility region, which indicates that the immune system might be differentially involved in schizophrenia; this is consistent with conclusions of Ruderfer et al. (2014). In contrast, SNPs rs13166360 and rs9788865, found to be specific to bipolar disorder, are located on chromosomes 5 and 16, respectively. The former is a coding SNP in the adenylyl cyclase type 2 gene (Mühleisen et al., 2014) and thus indicates that cyclic AMP signaling may differ between the two diseases. The latter appears to regulate levels of a long non-coding RNA (Lonsdale et al., 2013), which may point to a new mechanism of action in bipolar disorder.

6. Discussion

This paper studies signal classification for multiple sequences of test statistics. It introduces two new criteria for measuring misclassification errors and proposes powerful procedures for controlling these errors using a generalized compound decision-theoretic framework. It is shown that the proposed methods are asymptotically optimal.

It is straightforward to extend the proposed procedures to more than two sequences of test statistics. For example, considering three studies would allow for eight possible signal classes, which can be accommodated by extending model (6) to have eight components instead of four. The proposed oracle and data-driven procedures can then be modified accordingingly. However, the current implementation of these methods can grow unwieldy as the number of possible signal classes increases. In addition, the proposed procedures are developed under the assumption that the test statistics are independent across features. They may be robust to certain types of dependence, but could lose power if the features are highly correlated. These issues will be further studied in future work.

When domain knowledge, such as biological theory or prior experimental results, are available, they can be used as prior information to weight the observed test statistics, which can further improve the power of the proposed procedures. For the single data sequence, notable progress on weighting methods has been made (Roeder and Wasserman, 2009; Roquain and van de Wiel, 2009; Basu et al., 2017). However, it is unclear how these methods can be applied to multiple sequences of tests. This is an interesting problem for future study.

7. Proofs of Some Theoretical Results

This section only proves the theoretical results (Proposition 2, Theorem 2 and Theorem 4) for the set-specific error control problem (5). The proofs on the total error control problem (4) (Proposition 1, Theorem 1 and Theorem 3) are provided in Appendix A of the supplementary file.

7.1. Proof of Proposition 2

(i) To derive the oracle procedure that minimizes $L_S(\boldsymbol{\lambda}, \boldsymbol{\delta})$ it suffices to minimize each of the terms

$$\sum_{k=1}^{K}[I(\delta_i \neq k)\{1 - T_k^{OR}(X_{1i}, X_{2i})\} + \lambda_k I(\delta_i = k)\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k\}]$$

for $i = 1, \ldots, m$, which is achieved by $\delta_{Si}^{\boldsymbol{\lambda}}$ defined in (9). Thus, for any $\boldsymbol{\delta} \in \{0,1\}^m$,

$$L_S(\boldsymbol{\lambda}, \boldsymbol{\delta}_T^{\boldsymbol{\lambda}}) \leq L_S(\boldsymbol{\lambda}, \boldsymbol{\delta}),$$

where $\boldsymbol{\delta}_T^{\boldsymbol{\lambda}} = (\delta_{S1}^{\boldsymbol{\lambda}}, \ldots, \delta_{Sm}^{\boldsymbol{\lambda}})$. Take the expectation to both sides, then $\mathbb{E}\{L_S(\boldsymbol{\lambda}, \boldsymbol{\delta}_T^{\boldsymbol{\lambda}})\} \leq \mathbb{E}\{L_S(\boldsymbol{\lambda}, \boldsymbol{\delta})\}$ holds for any $\boldsymbol{\delta} \in \{0,1\}^m$.

(ii) Before proving the result of this part, the following result need to be discussed first. That is, $N_k^{OR}(\boldsymbol{\lambda})$ is non-increasing in λ_k but non-decreasing in $\lambda_{k'}$, $k' \neq k$.

Let $A_{\lambda_k} = \{T_k^{OR} \leq \alpha_k + \frac{1-\alpha_k}{\lambda_k+1}\}$ and $B_{\lambda_k} = \{\lambda_k(T_k^{OR} - \alpha_k) + T_k^{OR} < \min_{k' \neq k} \lambda_{k'}(T_{k'}^{OR} - \alpha_k) + T_{k'}^{OR}\}$, then

$$\begin{aligned} N_k^{OR}(\boldsymbol{\lambda}) = \;& E\{I\{T_k^{OR} \leq \alpha_k + \tfrac{1-\alpha_k}{\lambda_k+1}, \lambda_k(T_k^{OR} - \alpha_k) + T_k^{OR} \\ & < \min_{k' \neq k}[\lambda_{k'}(T_{k'}^{OR} - \alpha_k) + T_{k'}^{OR}]\}(T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k)\} \\ = \;& \mathbb{E}\left[I_{A_{\lambda_k}} I_{B_{\lambda_k}}[T_k^{OR}(X_{1i}, X_{2i}) - \alpha]\right] \end{aligned}$$

Suppose that $\lambda_k^{(1)} > \lambda_k^{(2)} > 0$, it can be concluded that $A_{\lambda_k^{(1)}} \subseteq A_{\lambda_k^{(2)}}$ and $B_{\lambda_k^{(1)}} \subseteq B_{\lambda_k^{(2)}}$. The former can be easily derived because $\alpha_k+(1-\alpha_k)/(\lambda_k^{(1)}+1) < \alpha_k+(1-\alpha_k)/(\lambda_k^{(2)}+1)$ when $\lambda_k^{(1)} > \lambda_k^{(2)} > 0$. The latter can be proved as follows.

If $T_k^{OR} < \alpha_k$, then $1 - T_{k'}^{OR} + 1 - T_k^{OR} \leq 1$ and $T_{k'}^{OR} > 1 - \alpha_k \geq \alpha_{k'}$. Thus, $\lambda_k(T_k^{OR} - \alpha_k) + T_k^{OR} < \min_{k' \neq k} \lambda_{k'}(T_{k'}^{OR} - \alpha_k) + T_{k'}^{OR}$ will always hold for any λ_k. That is, $B_{\lambda_k^{(1)}} \cap \{T_k^{OR} < \alpha_k\} = B_{\lambda_k^{(2)}} \cap \{T_k^{OR} < \alpha_k\}$.

If $T_k^{OR} \geq \alpha_k$, then $\lambda_k^{(1)}(T_k^{OR} - \alpha_k) + T_k^{OR} \geq \lambda_k^{(2)}(T_k^{OR} - \alpha_k) + T_k^{OR}$ and thus $B_{\lambda_k^{(1)}} \cap \{T_k^{OR} \geq \alpha_k\} \subseteq B_{\lambda_k^{(2)}} \cap \{T_k^{OR} \geq \alpha_k\}$. Till now, $B_{\lambda_k^{(1)}} \subseteq B_{\lambda_k^{(2)}}$ is proved completely.

Applying the results that $A_{\lambda_k^{(1)}} \subseteq A_{\lambda_k^{(2)}}$ and $B_{\lambda_k^{(1)}} \subseteq B_{\lambda_k^{(2)}}$, then

$$N_k^{OR}(\boldsymbol{\lambda}_1) - N_k^{OR}(\boldsymbol{\lambda}_2) =$$

$$\mathbb{E}\left[(I_{A_{\lambda_k^{(1)}}} I_{B_{\lambda_k^{(1)}}} - I_{A_{\lambda_k^{(2)}}} I_{B_{\lambda_k^{(2)}}}) I\{T_k^{OR} \geq \alpha_k\}[T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k]\right] \leq 0.$$

where $\boldsymbol{\lambda}_j$, $j = 1, 2$, is the $\boldsymbol{\lambda}$ with $\lambda_k = \lambda_k^{(j)}$ and $\lambda_{k'} = \lambda_{k'}$, $k' \neq k$. That is, $N_k^{OR}(\boldsymbol{\lambda})$ is non-increasing in λ_k. Similarly, it can be shown that $N_k^{OR}(\boldsymbol{\lambda})$ is non-decreasing in $\lambda_{k'}$.

The result of part (ii) in Proposition 2 can now be proved. It follows from Lemma 1 that there exists a $\boldsymbol{\lambda}^{**} \in \Lambda^{**}$ such that our constructed K sequences $\{\check{\lambda}_{k,n}, n \geq 1\}$ satisfy the relationships $\check{\lambda}_{k,1} \geq \cdots \geq \check{\lambda}_{k,n} \geq \cdots \geq \lambda_k^{**}$. and that $N_k^{OR}(\check{\boldsymbol{\lambda}}'_{k,n}) = 0$ holds for $k = 1, \ldots, K$ and $n \geq 1$.

Following from the monotone convergence theorem, each sequence $\{\check{\lambda}_{k,n}, n \geq 1\}$ will converge to a number, denoted as $\lambda_k^\star$. Let $\check{\boldsymbol{\lambda}}_n = (\check{\lambda}_{1,n}, \ldots, \check{\lambda}_{K,n})$, then

$$N_k^{OR}(\boldsymbol{\lambda}^\star) = \lim_{n\to\infty} N_k^{OR}(\check{\boldsymbol{\lambda}}_n) = \lim_{n\to\infty} N_k^{OR}(\check{\boldsymbol{\lambda}}'_{k,n}) = 0.$$

7.2. Proof of Theorem 2

(i) Similar as the proof (i) of Theorem 1, the results $\text{smFDR}_k(\boldsymbol{\delta}_S^\star) = \alpha_k$ for $k = 1, \ldots, K$ are very straightforward.

(ii) For any $\boldsymbol{\delta}$, if $\text{smFDR}_k(\boldsymbol{\delta}) \leq \alpha_k$, for all $k = 1, \ldots, K$ then

$$\begin{aligned}
&\mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K [1 - T_k^{OR}(X_{1i}, X_{2i})] - I(\delta_{Ti}^\star = k)[1 - T_k^{OR}(X_{1i}, X_{2i})]\} \\
&= \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K I(\delta_{Ti}^\star \neq k)[1 - T_k^{OR}(X_{1i}, X_{2i})] + \lambda_k^\star I(\delta_{Ti}^\star = k)(T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k)\} \\
&= \mathbb{E}\{L_S(\boldsymbol{\lambda}^\star, \boldsymbol{\delta}_S^\star)\} \leq \mathbb{E}\{L_S(\boldsymbol{\lambda}^\star, \boldsymbol{\delta})\} \\
&= \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K \{I(\delta_i \neq k)[1 - T_k^{OR}(X_{1i}, X_{2i})] + \lambda_k^\star I(\delta_i = k)(T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k)\}\} \\
&\leq \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K [1 - T_k^{OR}(X_{1i}, X_{2i})] - I(\delta_i = k)[1 - T_k^{OR}(X_{1i}, X_{2i})]\}.
\end{aligned}$$

Thus, $\text{tETP}(\boldsymbol{\delta}_S^\star) \geq \text{tETP}(\boldsymbol{\delta})$ holds.

7.3. Proof of Theorem 4

For ease of presentation, in this proof the $T_k^{OR}(X_{1i}, X_{2i})$ and $\hat{T}_k(X_{1i}, X_{2i})$ will be denoted as $T_{k,i}^{OR}$ and $\hat{T}_{k,i}$, respectively. Let $\hat{N}_k^{OR}(\boldsymbol{\lambda}) = 1/m \sum_{i=1}^m I\{\delta_{Si}^{\boldsymbol{\lambda}} = k\}[T_{k,i}^{OR} - \alpha_k]$. According to the weak law of large numbers (WLLN), result (a) $\hat{N}_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} N_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1})$ holds where $\check{\boldsymbol{\lambda}}_{k,n-1}$ is defined in proof of Proposition 2.

For $k \in \{1, \ldots, K\}$, fix all $\lambda_{k'}$, $k' \neq k$. $\hat{N}_k(\boldsymbol{\lambda})$ is then a function of λ_k and its continuous version, denoted as $\hat{N}_k^C(\boldsymbol{\lambda}$, can be defined, which is similar to the definition of $\hat{N}_T^C(\lambda)$ in the proof of Theorem 3. It is easy to check that $\hat{N}_k^C(\boldsymbol{\lambda})$ is continuous in λ_k and monotone. Thus, its inverse function, denoted as $\hat{N}_k^{C,-1}(\boldsymbol{\lambda})$, is well defined, continuous and monotone. According to the construction of the $\hat{N}_k^C(\boldsymbol{\lambda})$, results (b) $\hat{N}_k(\hat{\boldsymbol{\lambda}}_{k,n-1}) - \hat{N}_k^C(\hat{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} 0$ and (c) $\hat{\lambda}_{k,n} - \hat{N}_k^{C,-1}(\hat{\boldsymbol{\lambda}}_{k0,n-1})) \xrightarrow{p} 0$ holds for all k.

Suppose that $\hat{\lambda}_{k',n-1} \xrightarrow{p} \check{\lambda}_{k',n-1}$ for all $k' \neq k$, results (d) and (e) can then be derived immediately: (d) $\hat{N}_k^C(\check{\boldsymbol{\lambda}}_{k,n-1}) - \hat{N}_k^C(\hat{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} 0$ and (e) $\hat{N}_k^{C,-1}(\hat{\boldsymbol{\lambda}}_{k0,n-1}) - \hat{N}_k^{C,-1}(\check{\boldsymbol{\lambda}}_{k0,n-1}) \xrightarrow{p} 0$ where $\hat{\boldsymbol{\lambda}}_{k0,n-1}$ and $\check{\boldsymbol{\lambda}}_{k0,n-1}$ are the $\boldsymbol{\lambda}$'s with kth component 0 and the rest same as the counterparts of $\hat{\boldsymbol{\lambda}}_{k,n-1}$ and $\check{\boldsymbol{\lambda}}_{k,n-1}$, respectively.

To prove Theorem 4, the following results shall be discussed in turn. Suppose that $\hat{\lambda}_{k',n-1} \xrightarrow{p} \check{\lambda}_{k',n-1}$ for all $k' \neq k$, then

(1) $\hat{N}_k(\hat{\boldsymbol{\lambda}}_{k,n-1}) - \hat{N}_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} 0$ holds for any $\lambda_k > 0$;

(2) $\hat{N}_k^{C,-1}(\check{\boldsymbol{\lambda}}_{k0,n-1}) \xrightarrow{p} \check{\lambda}_{k,n}$ and $\hat{\lambda}_{k,n} \xrightarrow{p} \check{\lambda}_{k,n}, n \geq 1$;

Proof of result (1): From the proof of result (1) in Proof of Theorem 3, it suffices to show

$$\mathbb{E}\{[\hat{T}_{k,i} - \alpha_k] I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k\} - [T_{k,i}^{OR} - \alpha_k] I\{\delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\}\}^2 = o(1)$$

because $\hat{N}_k(\hat{\boldsymbol{\lambda}}_{k,n-1}) - \hat{N}_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} 0$ can then be proved by repeating to use the above result. See the proof of result (1) in Proof of Theorem 3 for details.
Following from Lemma 2,

$$\begin{aligned}
& P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} \neq k\} \\
\leq \; & P\{\hat{T}_{k,i} \leq \alpha_k + (1-\alpha_k)/(\lambda_k+1), T_{k,i}^{OR} > \alpha_k + (1-\alpha_k)/(\lambda_k+1)\} \\
+ \; & P\{\hat{T}_{k,i} > \alpha_k + (1-\alpha_k)/(\lambda_k+1), T_{k,i}^{OR} \leq \alpha_k + (1-\alpha_k)/(\lambda_k+1)\} \\
+ \; & P\{\lambda_k(\hat{T}_{k,i} - \alpha_k) + \hat{T}_{k,i} \leq \min\nolimits_{k' \neq k} \hat{\lambda}_{k',n-1}(\hat{T}_{k',i} - \alpha_{k'}) + \hat{T}_{k',i}, \\
& \lambda_k(T_{k,i}^{OR} - \alpha_k) + \hat{T}_{k,i} > \min\nolimits_{k' \neq k} \check{\lambda}_{k',n-1}(T_{k',i}^{OR} - \alpha_{k'}) + T_{k,i}^{OR}\} \\
+ \; & P\{\lambda_k(\hat{T}_{k,i} - \alpha_k) + \hat{T}_{k,i} > \min\nolimits_{k' \neq k} \hat{\lambda}_{k',n-1}(\hat{T}_{k',i} - \alpha_{k'}) + \hat{T}_{k',i}, \\
& \lambda_k(T_{k,i}^{OR} - \alpha_k) + \hat{T}_{k,i} \leq \min\nolimits_{k' \neq k} \check{\lambda}_{k',n-1}(T_{k',i}^{OR} - \alpha_{k'}) + T_{k,i}^{OR}\} \\
= \; & o(1) + o(1) = o(1),
\end{aligned}$$

and similarly

$$P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} \neq k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\} = o(1).$$

Then,

$$\begin{aligned}
& \mathbb{E}\{[\hat{T}_{k,i} - \alpha_k] I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k\} - [T_{k,i}^{OR} - \alpha_k] I\{\delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\}\}^2 \\
& \leq \mathbb{E}\{[\hat{T}_{k,i} - T_{k,i}^{OR}]^2 I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\} \\
& + \mathbb{E}\{[\hat{T}_{k,i} - \alpha]^2 I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} \neq k\} + \mathbb{E}\{[T_{k,i}^{OR} - \alpha]^2 I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} \neq k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\} \\
& \leq \mathbb{E}\{[\hat{T}_{k,i} - T_{k,i}^{OR}]^2\} + P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} = k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} \neq k\} + P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}_{k,n-1}} \neq k, \delta_{Si}^{\check{\boldsymbol{\lambda}}_{k,n-1}} = k\} \\
& = o(1) + o(1) + o(1) = o(1).
\end{aligned}$$

where $\mathbb{E}\{[\hat{T}_{k,i} - T_{k,i}^{OR}]^2\} = o(1)$ follows from the results that $\hat{T}_{k,i} - T_{k,i}^{OR} \xrightarrow{p} 0$ and $|\hat{T}_{k,i} - T_{k,i}^{OR}| \leq 1$.

Proof of result (2): Similar to the proof of result (2) in the Proof of Theorem 3, it suffices to prove that

$$\hat{N}_k^C(\check{\boldsymbol{\lambda}}_{k,n-1}) \xrightarrow{p} \hat{N}_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1}),$$

which follows from the above results (1), (a), (b) and (d). Therefore, $\hat{N}_k^{C,-1}(\check{\boldsymbol{\lambda}}_{k0,n-1}) \xrightarrow{p} \check{\lambda}_{k,n}$. By this result, together with results (c) and (e), $\hat{\lambda}_{k,n} \xrightarrow{p} \check{\lambda}_{k,n}$ can be obtained.

The result of Theorem can now be proved. When $n = 1$, $\hat{\lambda}_{k'} = \check{\lambda}_{k'} = \infty$ holds for all $k' \neq k$, thus $\hat{\lambda}_{k,n} \xrightarrow{p} \check{\lambda}_{k,n}$. Repeat to apply the result (2), we have

$$\hat{\lambda}_{k,n} \xrightarrow{p} \check{\lambda}_{k,n}, n \geq 1.$$

Take the limitations on the both sides, it leads to $\hat{\lambda}_k^\star \xrightarrow{p} \lambda_k^*$.

Following from the Lemma 2, we have

$$\begin{aligned}
& P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k, \delta_{Si}^{\boldsymbol{\lambda}^\star} \neq k\} \\
\leq\ & P\{\hat{T}_{k,i} \leq \alpha_k + (1-\alpha_k)/(\hat{\lambda}_k^\star + 1), T_{k,i}^{OR} > \alpha_k + (1-\alpha_k)/(\lambda_k^\star + 1)\} \\
+\ & P\{\hat{T}_{k,i} > \alpha_k + (1-\alpha_k)/(\hat{\lambda}_k^\star + 1), T_{k,i}^{OR} \leq \alpha_k + (1-\alpha_k)/(\lambda_k^\star + 1)\} \\
+\ & P\{\hat{\lambda}_k^\star(\hat{T}_{k,i} - \alpha_k) + \hat{T}_{k,i} \leq \min_{k' \neq k} \hat{\lambda}_{k'}^\star(\hat{T}_{k',i} - \alpha_{k'}) + \hat{T}_{k',i}, \\
& \lambda_k^\star(T_{k,i}^{OR} - \alpha_k) + \hat{T}_{k,i} > \min_{k' \neq k} \lambda_{k'}^\star(T_{k',i}^{OR} - \alpha_{k'}) + T_{k,i}^{OR}\} \\
+\ & P\{\hat{\lambda}_k^\star(\hat{T}_{k,i} - \alpha_k) + \hat{T}_{k,i} > \min_{k' \neq k} \hat{\lambda}_{k'}^\star(\hat{T}_{k',i} - \alpha_{k'}) + \hat{T}_{k',i}, \\
& \lambda_k^\star(T_{k,i}^{OR} - \alpha_k) + \hat{T}_{k,i} \leq \min_{k' \neq k} \lambda_{k'}^\star(T_{k',i}^{OR} - \alpha_{k'}) + T_{k,i}^{OR}\} \\
=\ & o(1) + o(1) = o(1),
\end{aligned}$$

and similarly

$$P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} \neq k, \delta_{Si}^{\boldsymbol{\lambda}^\star} = k\} = o(1).$$

Then,

$$\begin{aligned}
& \mathbb{E}\{|I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\} - I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}|\} \leq P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k, \delta_{Si}^{\boldsymbol{\lambda}^\star} \neq k\} + P\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} \neq k, \delta_{Si}^{\boldsymbol{\lambda}^\star} = k\} \\
=\ & o(1) + o(1) = o(1).
\end{aligned}$$

By the above result, it is easy to show that

$$\begin{aligned}
& |\mathbb{E}\{1/m \textstyle\sum_{i=1}^m (T_{k,i}^{OR} - \alpha_k) I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\}\}| = |\mathbb{E}\{[T_{k,i}^{OR} - \alpha][I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\} - I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}]\}| \\
& \leq \mathbb{E}\{|I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\} - I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}|\} = o(1),
\end{aligned} \tag{12}$$

$$\begin{aligned}
& |\mathbb{E}\{1/m \textstyle\sum_{i=1}^m [1 - T_{k,i}^{OR}][I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\} - I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}]\}| \\
\leq\ & \mathbb{E}\{|I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\} - I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}|\} = o(1),
\end{aligned} \tag{13}$$

and

$$E\left\{1/m \sum_{i=1}^m I\{\delta_{Si}^{\hat{\boldsymbol{\lambda}}^\star} = k\}\right\} = E\left\{1/m \sum_{i=1}^m I\{\delta_{Si}^{\boldsymbol{\lambda}^\star} = k\}\right\} + o(1) > 0 \tag{14}$$

By (12) and (14), the result that $\text{sMFDR}_k(\hat{\boldsymbol{\delta}}_S^\star) = \alpha_k + o(1)$ can be derived. By (13), the result that $\text{tETP}(\hat{\boldsymbol{\delta}}_S^\star)/\text{tETP}(\boldsymbol{\delta}_S^\star) = 1 + o(1)$ can be derived. Then, the proof of Theorem 4 is completed.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Andreassen, O. A., W. K. Thompson, A. J. Schork, S. Ripke, M. Mattingsdal, J. R. Kelsoe, K. S. Kendler, M. C. O'Donovan, D. Rujescu, T. Werge, P. Sklar, The Psychiatric Genomics Consortium (PGC) Bipolar Disorder and Schizophrenia Working Groups, J. C. Roddey, C.-H. Chen, L. McEvoy, R. S. Desikan, S. Djurovic, and A. M. Dale (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics 9*(4), e1003455.

Basu, P., T. T. Cai, K. Das, and W. Sun (2017). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*. To appear.

Benjamini, Y., R. Heller, and D. Yekutieli (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 367*(1906), 4255–4271.

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

Bogomolov, M. and R. Heller (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association 108*(504), 1480–1492.

Cai, T. T. and W. Sun (2017). Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 197–223.

Chi, Z. (2008). False discovery rate control with multivariate p-values. *Electronic Journal of Statistics 2*, 368–411.

Chung, D., C. Yang, C. Li, J. Gelernter, and H. Zhao (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics 10*(11), e1004787.

Cross-Disorder Group of Psychiatric Genomics Consortium (2013a). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics 45*(9), 984–994.

Cross-Disorder Group of Psychiatric Genomics Consortium (2013b). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet 381*(9875), 1371–1379.

Du, L. and C. Zhang (2014). Single-index modulated multiple testing. *The Annals of Statistics 42*(4), 30–79.

Flutre, T., X. Wen, J. Pritchard, and M. Stephens (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics 9*(5), e1003486.

Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 499–517.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *The Annals of Statistics 32*(3), 1035–1061.

Gratten, J., N. R. Wray, M. C. Keller, and P. M. Visscher (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience 17*(6), 782–790.

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science 348*(6235), 648–660.

Hawkins, R. D., G. C. Hon, and B. Ren (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics 11*(7), 476–486.

Heller, R., M. Bogomolov, and Y. Benjamini (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences 111*(46), 16262–16267.

Heller, R. and D. Yekutieli (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics 8*(1), 481–498.

Huang, J., R. H. Perlis, P. H. Lee, A. J. Rush, M. Fava, G. S. Sachs, J. Lieberman, S. P. Hamilton, P. Sullivan, P. Sklar, et al. (2010). Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. *American Journal of Psychiatry 167*(10), 1254–1263.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association 102*(478), 495–506.

Kristensen, V. N., O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer 14*(5), 299–313.

Li, G., D. D. Jima, F. A. Wright, and A. B. Nobel (2017). HT-eQTL: Integrative eQTL Analysis in a Large Number of Human Tissues. *arXiv preprint arXiv:1701.05426*.

Li, G., A. A. Shabalin, I. Rusyn, F. A. Wright, and A. B. Nobel (2013). An empirical bayes approach for multiple tissue eqtl analysis. *arXiv preprint arXiv:1311.2948*.

Li, H. (2013). Systems biology approaches to epidemiological studies of complex diseases. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine 5*(6), 677–686.

Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics 45*(6), 580–585.

Mühleisen, T. W., M. Leber, T. G. Schulze, J. Strohmaier, F. Degenhardt, J. Treutlein, M. Mattheisen, A. J. Forstner, J. Schumacher, R. Breuer, et al. (2014). Genome-wide association study reveals two new risk loci for bipolar disorder. *Nature communications 5*, 3339.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics 16*(2), 85–97.

Roeder, K. and L. Wasserman (2009). Genome-wide significance levels and weighted hypothesis testing. *Journal of the American Statistical Association 24*, 398–413.

Roquain, E. and M. A. van de Wiel (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics 3*, 678–711.

Ruderfer, D. M., A. H. Fanous, S. Ripke, A. McQuillin, R. L. Amdur, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Cross-Disorder Working Group of the Psychiatric Genomics Consortium, P. V. Gejman, M. C. O'Donovan, O. A. Andreassen, S. Djurovic, C. M. Hultman, J. R. Kelsoe, S. Jamain, M. Landén, M. Leboyer, V. Nimgaonkar, J. Numberger, J. W. Smoller, N. Craddock, A. Corvin, P. F. Sullivan, P. Holmans, P. Sklar, and K. S. Kendler (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry 19*(9), 1017–1024.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics 30*, 239–257.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26. Boca Raton: CRC press.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 479–498.

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

Taylor, J., R. Tibshirani, and B. Efron (2005). The miss rate for the analysis of gene expression data. *Biostatistics 6*(1), 111–117.

Torres, J. M., E. R. Gamazon, E. J. Parra, J. E. Below, A. Valladares-Salgado, N. Wacher, M. Cruz, C. L. Hanis, and N. J. Cox (2014). Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *The American Journal of Human Genetics 95*(5), 521–534.

Urbut, S. M., G. Wang, and M. Stephens (2017). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *bioRxiv*.

A. Proofs of Proposition 1, Theorem 1 and Theorem 3

A.1. Proof of Proposition 1

(i) To derive the oracle procedure that minimizes $L_T(\lambda, \boldsymbol{\delta})$ it suffices to minimize each of the terms

$$\sum_{k=1}^{K}[I(\delta_i \neq k)\{1 - T_k^{OR}(X_{1i}, X_{2i})\} + \lambda I(\delta_i = k)\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha_k\}]$$

for $i = 1, \dots, m$. It is straightforward to check that the above subjective function is achieved by

$$\delta_{Ti}^{\lambda} = \arg\min_{k\in\{0,\dots,K\}} [\sum_{k'\in\{1,\dots,K\},k'\neq k} \{1 - T_{k'}^{OR}(X_{1i}, X_{2i})\} + \lambda\{T_k^{OR}(X_{1i}, X_{2i}) - \alpha\}] \quad (15)$$

for $i = 1, \dots, m$, where $\lambda_0 = 0$. Thus, for any $\boldsymbol{\delta} \in \{0, 1\}^m$,

$$L_T(\lambda, \boldsymbol{\delta}_T^{\lambda}) \leq L_T(\lambda, \boldsymbol{\delta}),$$

where $\boldsymbol{\delta}_T^{\lambda} = (\delta_{T1}^{\lambda}, \dots, \delta_{Tm}^{\lambda})$. Take the expectations on both sides, then $\mathbb{E}\{L_T(\lambda, \boldsymbol{\delta}_T^{\lambda})\} \leq \mathbb{E}\{L_T(\lambda, \boldsymbol{\delta})\}$ holds for any $\boldsymbol{\delta} \in \{0, 1\}^m$.

(ii) It is easy to see from (8) that for $k = 1, \dots, K$,

$$I(\delta_{Ti}^{\lambda} = k) = I\left\{T_k^{OR} \leq \alpha + \frac{1-\alpha}{\lambda+1},\ T_k^{OR} < \min_{k'\neq k} T_{k'}^{OR}\right\}.$$

Thus,

$$N_T^{OR}(\lambda) = \mathbb{E}\left[\sum_{k=1}^{K} I\left\{T_k^{OR} \leq \alpha + \tfrac{1-\alpha}{\lambda+1},\ T_k^{OR} < \min\nolimits_{k'\neq k} T_{k'}^{OR}\right\}\left(T_k^{OR}(X_{1i}, X_{2i}) - \alpha\right)\right]$$
$$= \mathbb{E}\left[I\left\{T_{\min}^{OR}(X_{1i}, X_{2i}) \leq \alpha + \tfrac{1-\alpha}{\lambda+1}\right\}\left(T_{\min}^{OR}(X_{1i}, X_{2i}) - \alpha\right)\right]$$

Let $\lambda_1 > \lambda_2 > 0$. Then,

$$N_T^{OR}(\lambda_2) - N_T^{OR}(\lambda_1) = \mathbb{E}\left[I\left\{\alpha + \tfrac{1-\alpha}{\lambda_1+1} < T_{\min}^{OR} \leq \alpha + \tfrac{1-\alpha}{\lambda_2+1}\right\}\left(T_{\min}^{OR}(X_{1i}, X_{2i}) - \alpha\right)\right] \geq 0.$$

That is, $N_T^{OR}(\lambda)$ is non-increasing in λ.

As $N_T^{OR}(0) \geq 0$ and $N_T^{OR}(\lambda)$ is continuous, there exists at least one λ such that $N_T^{OR}(\lambda) = 0$. Together with the above monotonicity of $N_T^{OR}(\lambda)$, $N_T^{OR}(\lambda^\star) = 0$ always holds. So far, Proposition 1 is completed.

A.2. Proof of Theorem 1

(i) Since $(X_{11}, X_{21}), \dots, (X_{1i}, X_{2i}), \dots, (X_{1m}, X_{2m})$ are independently and identically distributed, $N_T^{OR}(\lambda) = 0$ is equivalent to $\text{TMFDR}(\boldsymbol{\delta}_T^{\star}) = \alpha$. By the part (ii) of Proposition 1, $Q_T^{OR}(\lambda^\star) = \alpha$ holds.

(ii) For any $\boldsymbol{\delta}$, if $\text{TMFDR}(\boldsymbol{\delta}) \leq \alpha$, then

$$\begin{aligned}
& \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K [1 - T_k^{OR}(X_{1i}, X_{2i})] - I(\delta^\star_{Ti} = k)[1 - T_k^{OR}(X_{1i}, X_{2i})]\} \\
= \; & \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K I(\delta^\star_{Ti} \neq k)[1 - T_k^{OR}(X_{1i}, X_{2i})] + \lambda^\star I[\delta^\star_{Ti} = k)](T_k^{OR}(X_{1i}, X_{2i}) - \alpha)\} \\
= \; & \mathbb{E}\{L_T(\lambda^\star, \boldsymbol{\delta}^\star_T)\} \leq \mathbb{E}\{L_T(\lambda^\star, \boldsymbol{\delta})\} \\
= \; & \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K \{I(\delta_i \neq k)[1 - T_k^{OR}(X_{1i}, X_{2i})] + \lambda^\star I(\delta_i = k)(T_k^{OR}(X_{1i}, X_{2i}) - \alpha)\}\} \\
\leq \; & \mathbb{E}\{\textstyle\sum_{i=1}^m \sum_{k=1}^K [1 - T_k^{OR}(X_{1i}, X_{2i})] - I(\delta_i = k)[1 - T_k^{OR}(X_{1i}, X_{2i})]\}.
\end{aligned}$$

Thus, $\text{TETP}(\boldsymbol{\delta}^\star_T) \geq \text{TETP}(\boldsymbol{\delta})$ holds.

A.3. Proof of Theorem 3

For ease of presentation, in this proof, the $T^{OR}_{\min}(X_{1i}, X_{2i})$ and $\hat{T}_{\min}(X_{1i}, X_{2i})$ will be denoted as $T^{OR}_{\min,i}$ and $\hat{T}_{\min,i}$, respectively. Let $t = \alpha + (1-\alpha)/(\lambda+1)$ and define $\hat{N}^{OR}_T(\lambda) = 1/m \sum_{i=1}^m I\{T^{OR}_{\min,i} < t\}[T^{OR}_{\min,i} - \alpha]$. Then, by the weak law of large numbers (WLLN), $\hat{N}^{OR}_T(\lambda) \xrightarrow{p} N^{OR}_T(\lambda)$ holds for any $\lambda > 0$.

Construct a continuous version of $\hat{N}_T(\lambda)$ (denoted as $\hat{N}^C_T(\lambda)$) by linear interpolation, i.e, for $\hat{\lambda}^{(j+1)}_{\min} \leq \lambda < \hat{\lambda}^{(j)}_{\min}$,

$$\hat{N}^C_T(\lambda) = \frac{t - \hat{T}^{(j)}_{\min}}{\hat{T}^{(j+1)}_{\min} - \hat{T}^{(j)}_{\min}} \hat{N}_T(\hat{\lambda}^{(j)}_{\min}) + \frac{\hat{T}^{(j+1)}_{\min} - t}{\hat{T}^{(j+1)}_{\min} - \hat{T}^{(j)}_{\min}} N^{OR}(\hat{\lambda}^{(j+1)}_{\min}),$$

where $\hat{T}^{(m+1)}_{\min} = 1$ and $\hat{\lambda}^{(j)}_{\min} = (1-\alpha)/(\hat{T}^{(j)}_{\min} - \alpha) - 1$. It is easy to check that $\hat{N}^C_T(\lambda)$ is continuous and monotone. Thus, its inverse function, denoted as $\hat{N}^{C,-1}_T(\lambda)$, is well defined, continuous and monotone. Meanwhile, $\mathbb{E}\{|\hat{N}^C_T(\lambda) - \hat{N}_T(\lambda)|\} \leq 1/m \to 0$ as $m \to \infty$. By Markov inequality, we have $\hat{N}^C_T(\lambda) - \hat{N}_T(\lambda) \xrightarrow{p} 0$ holds.

To prove Theorem 3, the following three results shall be discussed in turn.

(1) $\hat{N}_T(\lambda) - \hat{N}^{OR}_T(\lambda) \xrightarrow{p} 0$ holds for any $\lambda > 0$;

(2) $\hat{\lambda}^\star \xrightarrow{p} \lambda^\star$;

(3) $\mathbb{E}\{|I\{\hat{T}_{\min,i} \leq \hat{t}^\star\} - I\{T^{OR}_{\min,i} \leq t^\star\}|\} \to 0$ where $\hat{t}^\star = \alpha + (1-\alpha)/(\hat{\lambda}^\star + 1)$ and $t^\star = \alpha + (1-\alpha)/(\lambda^\star + 1)$.

Proof of result (1):
Since Conditions (C1) and (C2) hold, $\hat{T}_{\min,i} - T^{OR}_{\min,i} = o_P(1)$ holds uniformly for all i. Together with the results $0 \leq \hat{T}_{\min,i} \leq 1$, $0 \leq T^{OR}_{\min,i} \leq 1$ and Lemma 2, it can be shown that

$$\begin{aligned}
& \mathbb{E}\{[\hat{T}_{\min,i} - \alpha]I\{\hat{T}_{\min,i} \leq t\} - [T^{OR}_{\min,i} - \alpha]I\{T^{OR}_{\min,i} \leq t\}\}^2 \\
\leq \; & \mathbb{E}\{[\hat{T}_{\min,i} - T^{OR}_{\min,i}]^2 I\{\hat{T}_{\min,i} \leq t, T^{OR}_{\min,i} \leq t\}\} \\
& + \mathbb{E}\{[\hat{T}_{\min,i} - \alpha]^2 I\{\hat{T}_{\min,i} \leq t, T^{OR}_{\min,i} > t\} + \mathbb{E}\{[T^{OR}_{\min,i} - \alpha]^2 I\{\hat{T}_{\min,i} > t, T^{OR}_{\min,i} \leq t\}\} \\
\leq \; & \mathbb{E}\{[\hat{T}_{\min,i} - T^{OR}_{\min,i}]^2\} + \mathbb{E}\{I\{\hat{T}_{\min,i} \leq t, T^{OR}_{\min,i} > t\} + \mathbb{E}\{I\{\hat{T}_{\min,i} > t, T^{OR}_{\min,i} \leq t\}\} \\
= \; & o(1) + o(1) + o(1) = o(1).
\end{aligned} \tag{16}$$

By the Cauchy-Schwarz inequality, it has

$$\begin{aligned} & \mathbb{E}(|[\hat{T}_{\min,i}-\alpha]I\{\hat{T}_{\min,i}\le t\}-[T^{OR}_{\min,i}-\alpha]I\{T^{OR}_{\min,i}\le t\}|) \\ \le & (\mathbb{E}\{[\hat{T}_{\min,i}-\alpha]I\{\hat{T}_{\min,i}\le t\}-[T^{OR}_{\min,i}-\alpha]I\{T^{OR}_{\min,i}\le t\})^2)^{1/2} \\ = & o(1) \end{aligned} \tag{17}$$

and

$$\begin{aligned} & \mathbb{E}\{[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}] \\ \times & [(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]\} \\ \le & \{E[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]^2\}^{1/2} \\ \times & \{E[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]^2\}^{1/2} \\ = & o(1) \end{aligned} \tag{18}$$

Applying the results (16)-(18), the following results can be obtained.

$$\begin{aligned} & |\mathbb{E}\{1/m\textstyle\sum_{i=1}^m[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]\}| \\ & \le \mathbb{E}\{|(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}|\}=o(1) \end{aligned}$$

and

$$\begin{aligned} & var(1/m\textstyle\sum_{i=1}^m([(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]) \\ & \le 1/mE[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]^2 \\ & +(1-1/m)\mathbb{E}\{[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}] \\ & \times[(\hat{T}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le t\}-(T^{OR}_{\min,i}-\alpha)I\{T^{OR}_{\min,i}\le t\}]\} \\ & =o(1). \end{aligned}$$

Therefore, $\hat{N}_T(\lambda)-\hat{N}^{OR}_T(\lambda)\xrightarrow{p}0$ holds for any $\lambda>0$.
Proof of result (2):
According to the construction of $\hat{N}^C_T(\lambda)$, the result $\hat{\lambda}^\star-\hat{N}^{C,-1}_T(0)\xrightarrow{p}0$ holds immediately. Thus, it suffices to show that $\hat{N}^{C,-1}_T(0)\xrightarrow{p}\lambda^\star$. Since $\hat{N}^{C,-1}_T(\lambda)$ is continuous and $N^{OR}_T(\lambda^\star)=0$, it has that for any $\epsilon>0$, there exist a $\eta>0$, such that

$$\begin{aligned} P\{|\hat{N}^{C,-1}_T(0)-\lambda^\star|>\epsilon\} & =P\{|\hat{N}^{C,-1}_T(N^{OR}_T(\lambda^\star))-\hat{N}^{C,-1}_T(\hat{N}^C_T(\lambda^\star))|>\epsilon\} \\ & \le P\{|N^{OR}_T(\lambda^\star))-\hat{N}^C_T(\lambda^\star)|>\eta\} \end{aligned}$$

Following from the result (1), $\hat{N}^{OR}_T(\lambda)\xrightarrow{p}N^{OR}_T(\lambda)$ and $\hat{N}^C_T(\lambda)-\hat{N}_T(\lambda)\xrightarrow{p}0$, it is easy to see that $P\{|\hat{N}^{C,-1}_T(0)-\lambda^\star|>\epsilon\}$ will tends to 0. Thus, we proved the result (2).
Proof of result (3):
Following from the Lemma 2, the result (3) can be proved immediately.

$$\begin{aligned} & \mathbb{E}\{|I\{\hat{T}_{\min,i}\le\hat{t}^\star\}-I\{T^{OR}_{\min,i}\le t^\star\}|\} \\ \le & \mathbb{E}\{I\{\hat{T}_{\min,i}\le\hat{t}^\star,T^{OR}_{\min,i}>t^\star\}\}+\mathbb{E}\{I\{\hat{T}_{\min,i}>\hat{t}^\star,T^{OR}_{\min,i}\le t^\star\}\} \\ \to & 0,\quad as\ m\to\infty. \end{aligned}$$

The results of Theorem 3 can now be proved. By the result (3), it is easy to show that

$$\begin{aligned} & |\mathbb{E}\{1/m\textstyle\sum_{i=1}^m(T^{OR}_{\min,i}-\alpha)I\{\hat{T}_{\min,i}\le\hat{t}^\star\}\}| \\ = & |\mathbb{E}\{[T^{OR}_{\min,i}-\alpha][I\{\hat{T}_{\min,i}\le\hat{t}^\star\}-I\{T^{OR}_{\min,i}\le t^\star\}]\}| \\ \le & \mathbb{E}\{|I\{\hat{T}_{\min,i}\le\hat{t}^\star\}-I\{T^{OR}_{\min,i}\le t^\star\}|\}=o(1), \end{aligned} \tag{19}$$

$$\begin{array}{rl} & |\mathbb{E}\{1/m \sum_{i=1}^{m} [1 - T_{\min,i}^{OR}][I\{\hat{T}_{\min,i} \le \hat{t}^\star\} - I\{T_{\min,i}^{OR} \le t^\star\}]\}| \\ \le & \mathbb{E}\{|I\{\hat{T}_{\min,i} \le \hat{t}^\star\} - I\{T_{\min,i}^{OR} \le t^\star\}|\} = o(1), \end{array} \tag{20}$$

and

$$E\left\{1/m \sum_{i=1}^{m} I\{\hat{T}_{\min,i} \le \hat{t}^\star\}\right\} = E\left\{1/m \sum_{i=1}^{m} I\{T_{\min,i}^{OR} \le t^\star\}\right\} + o(1) > 0 \tag{21}$$

By (19) and (21), the result that $\text{TMFDR}(\hat{\boldsymbol{\delta}}_T^\star) = \alpha + o(1)$ can be derived. By (20), the result that $\text{TETP}(\hat{\boldsymbol{\delta}}_T^\star)/\text{TETP}(\boldsymbol{\delta}_T^\star) = 1 + o(1)$ can be derived. Then, the proof of Theorem 3 is completed.

B. Two Lemmas and their proofs

B.1. Lemma 1 and Its proof

LEMMA 1. *Suppose that* $\mathbf{0} \in \{(N_1^{OR}(\boldsymbol{\lambda}), \ldots, N_K^{OR}(\boldsymbol{\lambda})) : \boldsymbol{\lambda} \in \{\Re_+ \cup \{0\}\}^K\}$*, then there exists at least one* $\boldsymbol{\lambda}$ *with* $\lambda_k \ge 0$ *such that* $N_k^{OR}(\boldsymbol{\lambda}) = 0$ *for all* $k = 1, \ldots, K$*. Furthermore, denote by* $\Lambda^{**} = \{\boldsymbol{\lambda}^{**} = (\lambda_1^{**}, \ldots, \lambda_K^{**}) : N_k^{OR}(\boldsymbol{\lambda}^{**}) = 0$ *and* $\lambda_k^{**} \ge 0,$ *for all* $k = 1, \ldots, K\}$*, then* $N_k^{OR}(\check{\boldsymbol{\lambda}}'_{k,n}) = 0$ *holds for* $k = 1, \ldots, K$*, where* $\check{\boldsymbol{\lambda}}'_{k,n}$ *is the* $\boldsymbol{\lambda}$ *with* $\lambda_k = \check{\lambda}_{k,n}$ *and* $\lambda_{k'} = \check{\lambda}_{k',n-1}$*,* $k' \ne k$*, and that there exists a* $\boldsymbol{\lambda}^{**} \in \Lambda^{**}$ *such that* $\check{\lambda}_{k,n} \ge \lambda_k^{**}$ *for all* k*.*

Proof: Let $\check{\boldsymbol{\lambda}}_{k0,n-1}$ be the $\boldsymbol{\lambda}$ with $\lambda_k = 0$ and $\lambda_{k'} = \check{\lambda}_{k',n-1}$, $k' \ne k$ and consider the case $n = 1$. First, we claim that $N_k^{OR}(\check{\boldsymbol{\lambda}}_{k0,n-1}) \ge 0$. Because $N_k^{OR}(\check{\boldsymbol{\lambda}}_{k0,n-1}) < 0$ implies a contradiction that

$$N_k^{OR}(\boldsymbol{\lambda}^{**}) \le N_k^{OR}(\check{\boldsymbol{\lambda}}_{k0,n-1}) < 0 = N_k^{OR}(\boldsymbol{\lambda}^{**})$$

for any $\boldsymbol{\lambda}^{**} \in \Lambda^{**}$. The first inequality here follows from the monotonicity of the $N_k^{OR}(\boldsymbol{\lambda})$. The result $N_k^{OR}(\check{\boldsymbol{\lambda}}_{k0,n-1}) \ge 0$ indicates that there exists at least one $\lambda_k \ge 0$ such that $N_k^{OR}(\check{\boldsymbol{\lambda}}_{k,n-1}) = 0$. Since $N_k^{OR}(\boldsymbol{\lambda})$ is a continuous and monotone function, $\check{\lambda}_{k,n}$ exists uniquely and $N_k^{OR}(\check{\boldsymbol{\lambda}}'_{k,n}) = 0$.

Second, let $\boldsymbol{\lambda}_1^{**} = (\lambda_{11}, ..., \lambda_{1K})$ be one of Λ^{**}. Without loss of generality, suppose that for some $k \in \{1, ..., K\}$, $\check{\lambda}_{k,n} < \lambda_{1k}^{**}$ holds. Denote $\boldsymbol{\lambda}^{**}$ to be the $\boldsymbol{\lambda}_1^{**}$ with $\check{\lambda}_{k,n}$ in place of λ_{1k}^{**}. Next, we will show in the following that $\boldsymbol{\lambda}^{**} \in \Lambda^{**}$ and that $\check{\lambda}_{k,n} \ge \lambda_k^{**}$ for all k. If there are multiple $k \in \{1, ..., K\}$ satisfying $\check{\lambda}_{k,n} < \lambda_{1k}^{**}$, we can repeat the following proof many times to find the desired $\boldsymbol{\lambda}^{**}$. The latter result is obvious, so we only need to prove the former result. Repeat the proof of the monotonicity of the $N_k^{OR}(\boldsymbol{\lambda})$, we can further prove the result that $\sum_{k=1}^{K} N_k^{OR}(\boldsymbol{\lambda})$ is also non-increasing in λ_k. With this result, we have $\sum_{k=1}^{K} N_k^{OR}(\boldsymbol{\lambda}^{**}) \ge \sum_{k=1}^{K} N_k^{OR}(\boldsymbol{\lambda}_1^{**}) = 0$. By the monotonicity of the $N_k^{OR}(\boldsymbol{\lambda})$, we have $N_{k'}^{OR}(\boldsymbol{\lambda}^{**}) \le N_{k'}^{OR}(\boldsymbol{\lambda}_1^{**}) = 0$ for $k' \ne k$ and $N_k^{OR}(\boldsymbol{\lambda}^{**}) = 0$ (due to the fact that $0 = N_k^{OR}(\boldsymbol{\lambda}_1^{**}) \le N_k^{OR}(\boldsymbol{\lambda}^{**}) \le N_k^{OR}(\check{\boldsymbol{\lambda}}'_{k,n}) = 0$). Therefore, $N_{k'}^{OR}(\boldsymbol{\lambda}^{**}) = 0$ for $k' \ne k$. Otherwise, $0 \le \sum_{k=1}^{K} N_k^{OR}(\boldsymbol{\lambda}^{**}) < \sum_{k' \ne k}^{K} N_{k'}^{OR}(\boldsymbol{\lambda}^{**}) + N_k^{OR}(\boldsymbol{\lambda}^{**}) < 0$. This is a contradiction and implies that $N_{k'}^{OR}(\boldsymbol{\lambda}^{**}) = 0$ for all $k' \ne k$. Till now, we have found a $\boldsymbol{\lambda}^{**} \in \Lambda^{**}$ such that $\check{\lambda}_{k,n} \ge \lambda_k^{**}$ for all k.

For $n > 1$, if $\check{\lambda}_{k,n-1} \geq \lambda_k^{**}$ for all k, the results that $\check{\lambda}_{k,n}$ exists uniquely and $\check{\lambda}_{k,n-1} \geq \check{\lambda}_{k,n} \geq \lambda_k^{**}$ can be proved similarly. Together with the results for case $n = 1$, we complete the proof of this lemma.

B.2. Lemma 2 and Its proof

LEMMA 2. *Let X_m, Y_m, X and Y be bounded continuous variables. Suppose that $X \neq Y$, $X_m \xrightarrow{p} X$ and $Y_m \xrightarrow{p} Y$ hold, then $P\{X_m \leq Y_m, X > Y\} + P\{X_m > Y_m, X \leq Y\} \to 0$ as $m \to \infty$.*

Proof: For any $\epsilon > 0$, when m is sufficiently large,

$$\begin{aligned}
&P\{X_m \leq Y_m, X > Y\}\} \\
&= P\{X_m \leq Y_m, Y < X \leq Y + \epsilon\} + P\{X_m \leq Y_m, Y + \epsilon < X\} \\
&\leq P\{Y < X \leq Y + \epsilon\} + P\{X_m \leq Y_m, Y_m < Y + \epsilon/2, Y + \epsilon < X\} + P\{Y_m > Y + \epsilon/2\} \\
&\leq P\{0 < X - Y < \epsilon\} + P\{|X_m - X| > \epsilon/2\}\} + P\{|Y_m - Y| > \epsilon/2\}\}
\end{aligned}$$

and similarly

$$\begin{aligned}
& P\{X_m \leq Y_m, X > Y\}\} \\
\leq\ & P\{0 < Y - X < \epsilon\} + P\{|X_m - X| > \epsilon/2\}\} + P\{|Y_m - Y| > \epsilon/2\}\}.
\end{aligned}$$

Let $m \to \infty$ and $\epsilon \to 0$, then $P\{X_m \leq Y_m, X > Y\} + P\{X_m > Y_m, X \leq Y\} \to 0$.

C. Algorithm for adjusted sperate discovery procedure

First, for each class k, fix c_{k1} and find a c_{k2} over interval $(0, 1)$ using bisection searching algorithm such that

$$\sum_{i=1}^{m} I(\delta_i = k)[\hat{T}_k^{OR}(x_{1i}, x_{2i}) - \alpha_k'] \approx 0. \tag{22}$$

where α_k' is set to be the nominal SMFDR level α_k for the set-specific error control problem, or the TMFDR level α for the total error control problem. Second, vary c_{k1} from the smallest test statistic $S_{j(1)}$ to the largest test statistic $S_{j(m)}$, repeat the first step, and select the pair (c_{k1}, c_{k2}) having the largest $\sum_{i=1}^m I(\delta_i = k)[1 - \hat{T}_k^{OR}(x_{1i}, x_{2i})]$. Due to the fact that the signals in the test statistics are sparse, this step can be greatly simplified by varying c_{k1} from the smallest test statistic $S_{1(1)}$ to some test statistic $S_{1(m_1)}$ (e.g., $m_1 = 0.2m$). Third, apply the resulting c_{kj}'s to the decision rules, described in Section 3.3, to obtain the classification outcomes. If the three subsets have no overlaps, the classification outcomes are the desired ones. Otherwise, a remedy to this modification is necessary. This paper suggests keeping the genomic feature classified into several classes in the one having the smallest oracle statistics across classes. For example, if the ith feature is classified into classes 1 and 3, and $\hat{T}_1^{OR}(x_{1i}, x_{2i}) < \hat{T}_3^{OR}(x_{1i}, x_{2i})$, then the ith feature will be kept in class 1. Fourth, adjust the α_k' in (22) properly and repeat the previous three steps such that

$$\sum_{i=1}^{m} I(\delta_i = k)[\hat{T}_k^{OR}(x_{1i}, x_{2i}) - \alpha_k] \approx 0 \tag{23}$$

for the set-specific error control problem, and

$$\sum_{i=1}^{m}\sum_{k=1}^{K} I(\delta_i = k)[\hat{T}_k^{OR}(x_{1i}, x_{2i}) - \alpha] \approx 0 \tag{24}$$

for the total error control problem, are achieved by the resulting classification rules. For example, increase α'_k gradually till equations (23) for all k or equation (24) hold(s).