ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 4, *pp.* 707–734

Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests

Dongdong Xiang,
East China Normal University, Shanghai, People's Republic of China

Sihai Dave Zhao
University of Illinois at Urbana–Champaign, USA

and T. Tony Cai
University of Pennsylvania, Philadelphia, USA

[Received November 2017. Final revision April 2019]

Summary. The integrative analysis of multiple data sets is becoming increasingly important in many fields of research. When the same features are studied in several independent experiments, it can often be useful to analyse jointly 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. The paper studies this signal classification problem, motivated by a range of applications in large-scale genomics. Two new types of misclassification rate are introduced, and two oracle procedures are developed to control each type while also achieving the largest expected number of correct classifications. Corresponding data-driven procedures are also proposed, proved to be asymptotically valid and optimal under certain conditions and shown in numerical experiments to be nearly as powerful as the oracle procedures. In an application to psychiatric genetics, the procedures proposed are used to discover genetic variants that may affect both bipolar disorder and schizophrenia, as well as variants that may help to 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 multiple-testing methods are designed for analysing 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. A large amount of information is contained in the comparison of these studies, and 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, e.g. the effort to understand the genetic regulation of gene expression in humans. The 'Genotype

Address for correspondence: T. Tony Cai, Department of Statistics, Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, USA.
E-mail: tcai@wharton.upenn.edu

1369–7412/19/81707

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 levels of expression of which genes. This is accomplished by significance testing, for each gene in each tissue, of the association between the level of expression and each variant. But, because some regulatory variants may be active in only 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 a large number of sequences of multiple tests.

Similar joint analyses arise in psychiatric genetics. Some disorders, such as schizophrenia and bipolar disorder, share many symptoms and can be difficult to differentiate in clinical diagnoses (Andreassen *et al.*, 2013). Several large genomewide 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, 2013a). Identifying genetic variants that are significantly associated with one disease but not another can pave the way for a molecular diagnostic procedure that can more accurately distinguish the two conditions, whereas 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 analyses abound across genetics and genomics research, and can frequently be formulated in terms of grouping genomic features into different classes on the basis of their corresponding test statistics. 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 level of expression of a given gene. This paper will consider only $J=2$, but extensions to more than two studies are straightforward. Let $\theta_{ji}\in\{0,1\}$ indicate whether X_{ji} represents a signal or not, so $\theta_{ji}=1$ 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.

The present paper studies this signal classification problem, where the goal is to assign correctly as many genomic features to these signal classes as possible while controlling some measure of misclassification error. Signal classification can be viewed as a generalization of the standard multiple-testing problem, which seeks to determine only whether each feature is null or non-null and is therefore equivalent to binary classification. In contrast, signal classification is more similar to multiclass classification, where the results of applying a classification procedure to

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

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

Predicted class	*Matrix for the following true classes:*				*Total*
	0	*1*	*2*	*3*	
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

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, the misclassification error is frequently measured by 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 (Sun and Cai, 2007) 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 procedures proposed, and reveal relationships between the multiclass 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 that are proposed in this paper can be readily extended to other settings where classification into multiple signal classes is necessary, such as in image processing. For example, McHugh *et al.* (2008) proposed controlling the false discovery rate in a motion detection problem where each pixel in an image was to be classified as either background or foreground. In more general image segmentation problems, however, pixels may belong to more than two classes (Forsyth and Ponce, 2003), and controlling the misclassification error would require the methods that are proposed here.

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 formulation 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. Several researchers 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 to obtain a data-driven version of the optimal testing procedure (Chung *et al.*, 2014; Heller and Yekutieli, 2014). Recent work by Urbut *et al.* (2019) and Li *et al.* (2018a,b) has extended 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) defined the set of interest to contain only class 3, to discover features that are significant in both sequences. Alternatively, the set of interest could be defined to contain both classes 1 and 2, 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 which is called 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 1 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 multiclass 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 methods proposed are given in Section 4. In Section 5, the procedures are applied to study the genetic architectures of bipolar disorder and schizophrenia. Section 6 further considers the dependent situation under the multivariate normal distribution. A discussion of more possible extensions is given in Section 7. Proofs and additional results are contained in Appendix A and sections of the on-line supplementary file.

The programs that are were used to analyse the data can be obtained from

```
https://rss.onlinelibraray.wiley.com/hub/journal/14679868/series-
b-datasets
```

2. Problem formulation

2.1. Definitions

As illustrated in Table 1, 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 $\delta_i \in \{0, \ldots, K\}$ 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.*, 2018; Cai and Sun, 2017) and the false discovery rate (Storey, 2002; Benjamini and Hochberg, 1995; Genovese and Wasserman, 2002) need to be generalized to accommodate multiple sets of signal classes of interest. To measure the power of $\boldsymbol{\delta}$, define the total expected true positives number to be

$$_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}) = \mathbb{E}\left(\sum_{k=1}^{K} \sum_{l \in \mathcal{S}_k} C_{ll} \right), \tag{1}$$

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

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

$$_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}) = \frac{\mathbb{E}(\sum_{k=1}^{K} \sum_{l \in \mathcal{S}_k} \sum_{l' \neq l} C_{ll'})}{\mathbb{E}(\sum_{k=1}^{K} \sum_{l \in \mathcal{S}_k} R_l)}, \tag{2}$$

where the $C_{ll'}$ are the off-diagonal entries of Table 2. The numerator of equation (2) is the average number of features that are incorrectly classified into any of the $\mathcal{S}_k$, and the denominator equals the expected value of the total number of features that are 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

$$_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}) = \frac{\mathbb{E}(\sum_{l \in \mathcal{S}_k} \sum_{l' \neq l} C_{ll'})}{\mathbb{E}(\sum_{l \in \mathcal{S}_k} R_l)}, \qquad k = 1, \ldots, K, \tag{3}$$

which measures the proportion of misclassifications only for the kth set of interest.

Remark 1. Instead of the marginal false discovery rate, versions of the false discovery rate of Benjamini and Hochberg (1995) can be employed to measure misclassification error. For example, set-specific misclassification errors could also be measured with

$$_{\mathrm{S}}\mathrm{FDR}_k(\boldsymbol{\delta}) = \mathbb{E}\left\{ \frac{\sum_{l \in \mathcal{S}_k} \sum_{l' \neq l} C_{ll'}}{\max(\sum_{l \in \mathcal{S}_k} R_l, 1)} \right\}$$

for $k = 1, \ldots, K$, which is the Benjamini and Hochberg (1995) false discovery rate for identifying signals to be in $\mathcal{S}_k$ or not. Results from Genovese and Wasserman (2002) for the binary signal classification problem then imply that $_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}) = {}_{\mathrm{S}}\mathrm{FDR}_k(\boldsymbol{\delta}) + O(m^{-1/2})$ for all k, so that these two measures are asymptotically equivalent. Following previous literature (Sun and Cai, 2007; Cai and Sun, 2017), marginal false discovery rates will be used in this paper for technical convenience when obtaining optimality results.

2.2. Signal classification problems

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

Definition 1 (total). Under total error, find the $\boldsymbol{\delta}$ that

$$\text{maximizes }_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}) \qquad \text{subject to }_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}) \leqslant \alpha \tag{4}$$

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

Definition 2 (set specific). Under set-specific error, find the $\boldsymbol{\delta}$ that

$$\text{maximizes }_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}) \qquad \text{subject to }_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}) \leqslant \alpha_k, \qquad \text{for } k = 1, \ldots, K, \tag{5}$$

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 enable fine control over the different types of misclassification errors. For example, if the X_{1i} come from a study with a very large sample size whereas 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 with 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 true positives number in problem (4).

3. Proposed methods

3.1. Oracle procedures

Similarly 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 (IID) across features i. Since in many applications the test statistics X_{1i} and X_{2i} arise from independent data sets, assume that they are independent conditionally on θ_{1i} and θ_{2i}. Let $F_{j0}(x)$ and $F_{j1}(x)$ denote the distribution functions of X_{ji} conditionally on $\theta_{ji} = 0$ and $\theta_{ji} = 1$ respectively, where F_{j0} is known. Throughout, it will be assumed that F_{j0} and F_{ji} admit continuous density functions. Then the test statistics (X_{1i}, X_{2i}) are IID according to the four-group model

$$(X_{1i}, X_{2i}) \overset{\mathrm{IID}}{\sim} \sum_{l=0}^{3} \pi_l F_{1l_1} F_{2l_2}, \tag{6}$$

where, for $l \in \{0, \ldots, 3\}$, 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, $l = 2$ implies that $l_1 = 1$ and $l_2 = 0$. Finally, $\pi_l = \mathbb{P}(\theta_{1i} = l_1, \theta_{2i} = l_2)$.

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

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

subject to

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

where

$$T_k^{\mathrm{OR}}(x_1,x_2)=\frac{\sum_{l\notin\mathcal{S}_k}\pi_l f_{1l_1}(x_1)f_{2l_2}(x_2)}{\sum_{l=0}^{3}\pi_l f_{1l_1}(x_1)f_{2l_2}(x_2)} \tag{7}$$

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

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

since any $\boldsymbol{\delta}$ that minimizes $L_T(\lambda,\boldsymbol{\delta})$ conditionally on the observed test statistics will also minimize $\mathbb{E}\{L_T(\lambda,\boldsymbol{\delta})\}$. This $L_T(\lambda,\boldsymbol{\delta})$ can be regarded as a generalized loss function consisting of three terms:

(a) the cost of misclassifying a signal into the null class $\mathcal{S}_0$,
(b) the cost, weighted by λ, of misclassifying a signal into a class of interest, and
(c) the benefit of identifying features as signals, i.e. $-\lambda\Sigma I(\delta_i\neq 0)$.

In this sense, the solution based on $L_T(\lambda,\boldsymbol{\delta})$ can be regarded as a generalization of the compound decision theoretic treatment of false discovery rate control, proposed by Sun and Cai (2007), to signal classification.

For any $\lambda>0$, define the classification rule $\boldsymbol{\delta}_T^{\lambda}=(\delta_{T1}^{\lambda},\ldots\delta_{Tm}^{\lambda})$ to be the minimizer of $L_T(\lambda,\boldsymbol{\delta})$, where the ith component of $\boldsymbol{\delta}_T^{\lambda}$ is defined as

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

The following result characterizes the behaviour of $\boldsymbol{\delta}_T^{\lambda}$.

Proposition 1. Suppose that the continuous test statistics (X_{1i},X_{2i}) are IID according to the four-group model (6), and $\max_{l\in\bigcup_{k=1}^{K}\mathcal{S}_k}\pi_l>0$. Then, for $\boldsymbol{\delta}_T^{\lambda}$ defined as in equation (8),

(a) $\boldsymbol{\delta}_T^{\lambda}$ minimizes $\mathbb{E}\{L_T(\lambda,\boldsymbol{\delta})\}$;
(b) let $N_T^{\mathrm{OR}}(\lambda)=\mathbb{E}[\Sigma_{k=1}^{K}I(\delta_{Ti}^{\lambda}=k)\{T_k^{\mathrm{OR}}(X_{1i},X_{2i})-\alpha\}]$ and define

$$\lambda^*=\inf\{\lambda:N_T^{\mathrm{OR}}(\lambda)\leqslant 0\}.$$

If $N_T^{\mathrm{OR}}(0)\geqslant 0$ holds, then $N_T^{\mathrm{OR}}(\lambda^*)=0$.

Remark 2. The condition $\max_{l\in\bigcup_{k=1}^{K}\mathcal{S}_k}\pi_l>0$ in proposition 1 ensures that at least one of the signal classes of interest contains a non-zero proportion of features. Intuitively, an optimal decision rule should make the most of the misclassification error that it is allowed, to maximize the number of discoveries that it makes. In other words, $\boldsymbol{\delta}_T^{\lambda}$ should achieve $_{\mathrm{TM}}\mathrm{FDR}=\alpha$. The quantity $N_T^{\mathrm{OR}}(\lambda)$ in proposition 1 derives from the constraint on $_{\mathrm{TM}}\mathrm{FDR}$ and can be interpreted as a measure of how much of the allotted misclassification error has not been used up by $\boldsymbol{\delta}_T^{\lambda}$. It can be shown that $N_T^{\mathrm{OR}}(\lambda)$ is non-increasing in λ, so $N_T^{\mathrm{OR}}(0)<0$ would imply that, for some values of α, there may not be any λ such that $_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}_T^{\lambda})$ exactly attains α. Thus the assumption $N_T^{\mathrm{OR}}(0)\geqslant 0$ is necessary to ensure that the nominal level α can be achieved exactly by some λ.

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

Theorem 1. Suppose that the continuous test statistics (X_{1i}, X_{2i}) are IID according to the four-group model (6) and that $\max_{l \in \bigcup_{k=1}^{K} \mathcal{S}_k} \pi_l > 0$. With δ_{Ti}^{λ} and λ^* defined in proposition 1, define

$$\boldsymbol{\delta}_T^* = (\delta_{T1}^{\lambda^*}, \ldots, \delta_{Tm}^{\lambda^*}).$$

If α satisfies $N_T^{\mathrm{OR}}(0) \geqslant 0$ from proposition 1, part (b), then:

(a) ${}_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}_T^*) = \alpha$;
(b) for any other classification rule $\boldsymbol{\delta}$ that satisfies ${}_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}) \leqslant \alpha$,

$${}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_T^*) \geqslant {}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}).$$

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

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

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

$$L_S(\boldsymbol{\lambda}, \boldsymbol{\delta}) = \sum_{k=1}^{K} \sum_{i=1}^{m} I(\delta_i \neq k) \{1 - T_k^{\mathrm{OR}}(X_{1i}, X_{2i})\} + \sum_{k=1}^{K} \sum_{i=1}^{m} \lambda_k I(\delta_i = k) \{T_k^{\mathrm{OR}}(X_{1i}, X_{2i}) - \alpha_k\}.$$

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

$$\delta_{Si}^{\boldsymbol{\lambda}} = \underset{k \in \{0, \ldots, K\}}{\arg\min} \sum_{k' \in \{1, \ldots, K\}, k' \neq k} [\{1 - T_{k'}^{\mathrm{OR}}(X_{1i}, X_{2i})\} + \lambda_k \{T_k^{\mathrm{OR}}(X_{1i}, X_{2i}) - \alpha_k\}]. \tag{9}$$

An analogue to proposition 1 can be obtained to characterize $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$.

Proposition 2. Suppose that the continuous test statistics (X_{1i}, X_{2i}) are IID according to the four-group model (6), and $\max_{l \in \mathcal{S}_k} \pi_l > 0$ holds for all $k \in \{1, \ldots, K\}$. Then for $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ defined as in equation (9):

(a) $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ minimizes $\mathbb{E}\{L_S(\boldsymbol{\lambda}, \boldsymbol{\delta})\}$;
(b) let $N_k^{\mathrm{OR}}(\boldsymbol{\lambda}) = \mathbb{E}[I(\delta_{Si}^{\boldsymbol{\lambda}} = k)\{T_k^{\mathrm{OR}}(X_{1i}, X_{2i}) - \alpha_k\}]$ and define

$$\check{\lambda}_{k,n} = \inf\{\lambda_k \leqslant \check{\lambda}_{k,n-1} : N_k^{\mathrm{OR}}(\check{\boldsymbol{\lambda}}_{k,n-1}) \leqslant 0\}, \qquad k = 1, \ldots K,$$

where $n \geqslant 1$, $\check{\lambda}_{k,0} = \infty$ and $\check{\boldsymbol{\lambda}}_{k,n-1}$ is the $\boldsymbol{\lambda}$ with $\lambda_{k'} = \check{\lambda}_{k',n-1}$, $k' \neq k$. Suppose that $\alpha_k + \alpha_{k'} \leqslant 1$ holds for any $k \neq k' \in \{1, \ldots, K\}$, and $\mathbf{0} \in \{(N_1^{\mathrm{OR}}(\boldsymbol{\lambda}), \ldots, N_K^{\mathrm{OR}}(\boldsymbol{\lambda})) : \boldsymbol{\lambda} \in \{\Re_+ \cup \{0\}\}^K\}$. Then, the sequence $\{\check{\lambda}_{k,n}, n \geqslant 1\}$ is convergent and $N_k^{\mathrm{OR}}(\boldsymbol{\lambda}^*) = 0$ for all $k = 1, \ldots, K$ where $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_K^*)$ and $\lambda_k^* = \lim_{n \to \infty} \check{\lambda}_{k,n}$.

Remark 3. The condition $\max_{l \in \mathcal{S}_k} \pi_l > 0$ in proposition 2 ensures that each signal class of interest contains a non-zero proportion of features. Proposition 2, part (b), plays the same role as the condition on $N_T^{\mathrm{OR}}(0)$ in proposition 1, part (b). In the set-specific error control problem (5), not all error levels $\alpha_1, \ldots, \alpha_K$ correspond to a $\boldsymbol{\lambda}$ such that ${}_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}_S^{\boldsymbol{\lambda}})$ 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 $0\leqslant\alpha_k\leqslant\frac{1}{2}, k=1,\ldots,K$, which is adequate for many applications.

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

Theorem 2. Suppose that the continuous test statistics (X_{1i},X_{2i}) are IID according to the four-group model (6), and $\max_{l\in\mathcal{S}_k}\pi_l>0$ holds for all $k\in\{1,\ldots,K\}$. With $\delta_{Si}^{\boldsymbol{\lambda}^*}$ and $\boldsymbol{\lambda}^*$ defined in proposition 2, define

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

If $\alpha_1,\ldots,\alpha_K$ satisfy the conditions in proposition 2, part (b), then:

(a) ${}_{\text{SM}}\text{FDR}_k(\boldsymbol{\delta}_S^*)=\alpha_k$ for $k=1,\ldots,K$;
(b) for any other classification rule $\boldsymbol{\delta}$ that satisfies ${}_{\text{SM}}\text{FDR}_k(\boldsymbol{\delta})\leqslant\alpha_k$, $k=1,\ldots,K$,

$${}_{\text{T}}\text{ETP}(\boldsymbol{\delta}_S^*)\geqslant{}_{\text{T}}\text{ETP}(\boldsymbol{\delta}).$$

When class 3 is the only class of interest, the two oracle methods that are described in theorems 1 and 2 are identical to the oracle method that was proposed by Heller and Yekutieli (2014). Otherwise, they are different for more general signal classification problems, which will be further explored in simulations in Section 4.

3.2. Data-driven procedures

The oracle procedures that were described in the previous section cannot be implemented in practice because they are functions of $T_k^{\text{OR}}(X_{1i},X_{2i})$ defined in equation (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 $\pi_{jl_j}=\mathbb{P}(\theta_{ji}=l_j)$ and the marginal densities $f_j(x)=\pi_{j0}f_{j0}(x)+\pi_{j1}f_{j1}(x)$ and rewriting

$$T_k^{\text{OR}}(x_1,x_2)=\frac{\sum_{l\notin\mathcal{S}_k}\{\pi_l/(\pi_{1l_1}\pi_{2l_2})\}\{\pi_{1l_1}f_{1l_1}(x_1)/f_1(x_1)\}\pi_{2l_2}f_{2l_2}(x_2)/f_2(x_2)}{\sum_{l=0}^{3}\{\pi_l/(\pi_{1l_1}\pi_{2l_2})\}\{\pi_{1l_1}f_{1l_1}(x_1)/f_1(x_1)\}\pi_{2l_2}f_{2l_2}(x_2)/f_2(x_2)}.$$

Next, estimates $\hat{\pi}_{j1}$ and $\hat{\pi}_{j0}=1-\hat{\pi}_{j1}$ 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 by using standard kernel-based methods (Silverman, 1986) with the rule-of-thumb bandwidth. The likelihood ratios $\pi_{jl_j}f_{jl_j}(x_j)/f_j(x_j)$, $j=1,2$, in $T_k^{\text{OR}}(x_1,x_2)$ can then be estimated. In practice, each estimated likelihood ratio is set equal to 1 if its calculated value exceeds 1. In some cases, the collected sample could be contaminated and hence the standard kernel density estimator may not work very well. In these situations, robust kernel density estimation (Kim and Scott, 2012) would be a good choice because it enjoys similar theoretical properties to those of the standard kernel density estimator. Finally, an estimate $1-\hat{\pi}_0$ of $1-\mathbb{P}(\theta_{1i}=0,\theta_{2i}=0)$ can be obtained by applying the method of Jin and Cai (2007) to the statistics

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

where $G_{\chi^2,2}$ is the distribution function of a χ^2 random variable with 2 degrees of freedom, and estimates of the other π_l can be calculating by using $\hat{\pi}_0$ and the $\hat{\pi}_{jl_j}$. 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 $\hat{N}_T(\lambda) = m^{-1}\Sigma_{i=1}^m \Sigma_{k=1}^K I(\hat{\delta}_{Ti}^{\lambda} = k)\{\hat{T}_k(X_{1i}, X_{2i}) - \alpha\}$. This expression can be simplified because it can be seen from the definition of the oracle total error control rule (8) that for $k = 1, \ldots, K$

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

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

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

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

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

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

Theorem 3 shows that the data-driven $\hat{\boldsymbol{\delta}}_T^*$ is asymptotically valid and optimal.

Theorem 3. Suppose that all the assumptions in theorem 1 hold. Assume that f_j, $j = 1, 2$, are continuous and positive on the real line, and that the second derivative satisfies $\int (f_j'')^2 \mathrm{d}x_j < \infty$. Let $\hat{\pi}_{jl}$, $\hat{\pi}_k$ and $\hat{f}_j$ be the estimates of π_{jl}, π_k and f_j, for $j = 1, 2$, $l = 0, 1$ and $k = 0, 1, 2, 3$ such that the following conditions hold.

Condition 1. $\hat{\pi}_{jl} \rightarrow^{\mathrm{p}} \pi_{jl}$, $j = 1, 2$, $l = 0, 1$, and $\hat{\pi}_k \rightarrow^{\mathrm{p}} \pi_k$, $k = 0, 1, 2, 3$, as $m \rightarrow \infty$.

Condition 2. $\mathbb{E}\|\hat{f}_j - f_j\| \rightarrow 0$, $j = 1, 2$, as $m \rightarrow \infty$.

Then

(a) $\mathrm{TMFDR}(\hat{\boldsymbol{\delta}}_T^*) = \alpha + o(1)$, and
(b) $\mathrm{TETP}(\hat{\boldsymbol{\delta}}_T^*)/\mathrm{TETP}(\boldsymbol{\delta}_T^*) = 1 + o(1)$.

Remark 4. Conditions 1 and 2 are mild. Jin and Cai (2007) showed that the estimates $\hat{\pi}_{j1}$ and $1 - \hat{\pi}_0$ provided above converge in probability to π_{j1} and $1 - \pi_0$ respectively, and it is straightforward to check that the proposed $\hat{\pi}_k$ satisfy condition 1 as well. Silverman (1986) showed that the standard kernel density estimate $\hat{f}_j(x)$ enjoys the property that $\mathbb{E}\|\hat{f}_j - f_j\|^2 \rightarrow 0$ when the observations are IID, satisfying condition 2.

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 $\hat{\delta}_{T(i)}^*$, $T_k^{(i)}$ be the corresponding decision functions and testing statistics: define $r = \max\{j : (1/j)\Sigma_{i=1}^j \hat{T}_{\min}^{(i)} \leqslant \alpha\}$; then,

$$\hat{\delta}_{T(i)}^* = \begin{cases} k, & i \leqslant r \text{ and } \hat{T}_{\min}^{(i)} = \hat{T}_k^{(i)}, \\ 0, & i > r \end{cases}$$

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} and $\hat{N}_k(\boldsymbol{\lambda}) = (1/m)\Sigma_{i=1}^m I(\hat{\delta}^{\boldsymbol{\lambda}}_{Si} = k)\{\hat{T}_k(x_{1i}, x_{2i}) - \alpha_k\}$ and construct a sequence $\{\hat{\lambda}_{k,n}, n \geqslant 1\}$ that satisfies

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

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 $k' \neq k$. The proof that the sequence $\{\hat{\lambda}_{k,n}, n \geqslant 1\}$ converges is similar to the convergence proof for the sequence $\{\check{\lambda}_{k,n}, n \geqslant 1\}$ from proposition 2. Let $\hat{\lambda}_k^*$ be the value to which $\{\hat{\lambda}_{k,n}, n \geqslant 1\}$ converges. Then the data-driven procedure that solves problem (5) can be defined as

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

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

A fast algorithm for calculating $\hat{\boldsymbol{\delta}}_S^*$ is provided in Table 4, which shows that the algorithm can be regarded as a stagewise multiple-testing procedure for identifying set-specific signals, i.e. 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 $\hat{\boldsymbol{\delta}}_S^*$ is asymptotically valid and optimal.

Theorem 4. Suppose that all the assumptions in theorems 2 and 3 hold. Then, for all $k \in \{1, \ldots, K\}$,

(a) $_{\mathrm{SM}}\mathrm{FDR}_k(\hat{\boldsymbol{\delta}}_S^*) = \alpha_k + o(1)$, and
(b) $_{\mathrm{T}}\mathrm{ETP}(\hat{\boldsymbol{\delta}}_S^*)/_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_S^*) = 1 + o(1)$.

3.3. *Adjusted separate discovery procedure*

As described in Section 1.2, a common existing approach to signal classification is the separate discovery procedure. If P_{ji} is the p-value of the ith feature in sequence j, this procedure would set $\delta_i = 1$ if $P_{1i} > c_1$ and $P_{2i} \leqslant c_2$, $\delta_i = 2$ if $P_{1i} \leqslant c_1$ and $P_{2i} > c_2$, and $\delta_i = 3$ if $P_{1i} \leqslant c_1$ and $P_{2i} \leqslant c_2$, for some cut-offs c_j such that the marginal false discovery rate for sequence j attains α_j. The separate discovery procedure cannot control $_{\mathrm{TM}}\mathrm{FDR}$ and $_{\mathrm{SM}}\mathrm{FDR}$ at desired nominal levels. This cannot be remedied by merely choosing different values for α_j; the key difficulty is that the non-discovery classifications in each sequence are unreliable.

This section proposes an adjusted separate discovery procedure that can provide valid control of the two types of misclassification error that are introduced in this paper. For illustration, the

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 $r_k = \max\{j : (1/j)\Sigma_{i=1}^j \hat{T}_k^{(i)} \leqslant \alpha_k\}$

Step 2: for each k, calculate $\hat{N}_k(\boldsymbol{\lambda})$ and $\hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,r_k+1})$ where $\tilde{\boldsymbol{\lambda}}_{k,j} = (\tilde{\lambda}_{1,j}, \ldots, \tilde{\lambda}_{K,j})$ with $\tilde{\lambda}_{k,j} = (1-\alpha_k)/(\hat{T}_k^{(j)} - \alpha_k) - 1$ and $\tilde{\lambda}_{k',j} = \lambda_{k'}$, $k' \neq k$: if $\hat{N}_k(\boldsymbol{\lambda}) \leqslant 0$ 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 $\tilde{r}_k = \max\{j \geqslant r_k : \hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,j}) \leqslant 0\}$ and reset $r_k = \tilde{r}_k$: then, update the $\boldsymbol{\lambda}$ in step 1 and repeat steps 2 and 3 until this loop is terminated; the $\boldsymbol{\lambda}$ in the last iteration is the desired $\hat{\boldsymbol{\lambda}}^*$

Step 4: apply $\hat{\boldsymbol{\lambda}}^*$ to equation (9) with $\hat{T}_k$ in place of T_k^{OR} to obtain the classification rule $\hat{\boldsymbol{\delta}}_S^* = (\hat{\delta}_{S1}^*, \ldots, \hat{\delta}_{Sm}^*)$

$K=3$ setting is considered below. The main idea is to employ different cut-offs for each set of classes of interest. Specifically, set $\delta_i=1$ if $P_{1i}>c_{11}$ and $P_{2i}\leqslant c_{12}$, $\delta_i=2$ if $P_{1i}\leqslant c_{21}$ and $P_{2i}>c_{22}$, and $\delta_i=3$ if $P_{1i}\leqslant c_{31}$ and $P_{2i}\leqslant c_{32}$, where the c_{kj} can all be unequal. Then the separate discovery procedure can be adjusted by finding the c_{kj} such that $\Sigma_{i=1}^m I(\delta_i=k)\{\hat{T}_k(x_{1i},x_{2i})-\alpha_k\}\approx 0$ for the set-specific error control problem, and $\Sigma_{i=1}^m\Sigma_{k=1}^K I(\delta_i=k)\{\hat{T}_k(x_{1i},x_{2i})-\alpha\}\approx 0$ for the total error control problem. This new procedure may lead to some features being classified into more than one set of interest. Section C in the on-line supplementary file provides an algorithm to find the cut-offs, 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 that are proposed in this paper, it is computationally intensive and its cut-offs are not optimal in the sense of having the largest $_{\text{T}}$ETP-values.

4. Simulations

This section investigates the numerical performances of the proposed oracle and data-driven procedures. Pairs of test statistics (X_{1i},X_{2i}) for $i=1,\ldots,m$ were generated for $m=20000$ features according to the four-group model (6), with class labels defined as in Table 1. Specifically, the null and alternative density functions were

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

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 whereas 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 unadjusted separate discovery approach based on p-values, described in Section 1.2 (for the total error control problem, the error levels in each individual sequence were 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 were all set to equal the average of the desired nominal set-specific marginal false discovery rates; in each sequence, the procedure of Genovese and Wasserman (2004) was used to control the marginal false discovery rate, which is asymptotically equivalent to the false discovery rate of Benjamini and Hochberg (1995));

(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 other 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

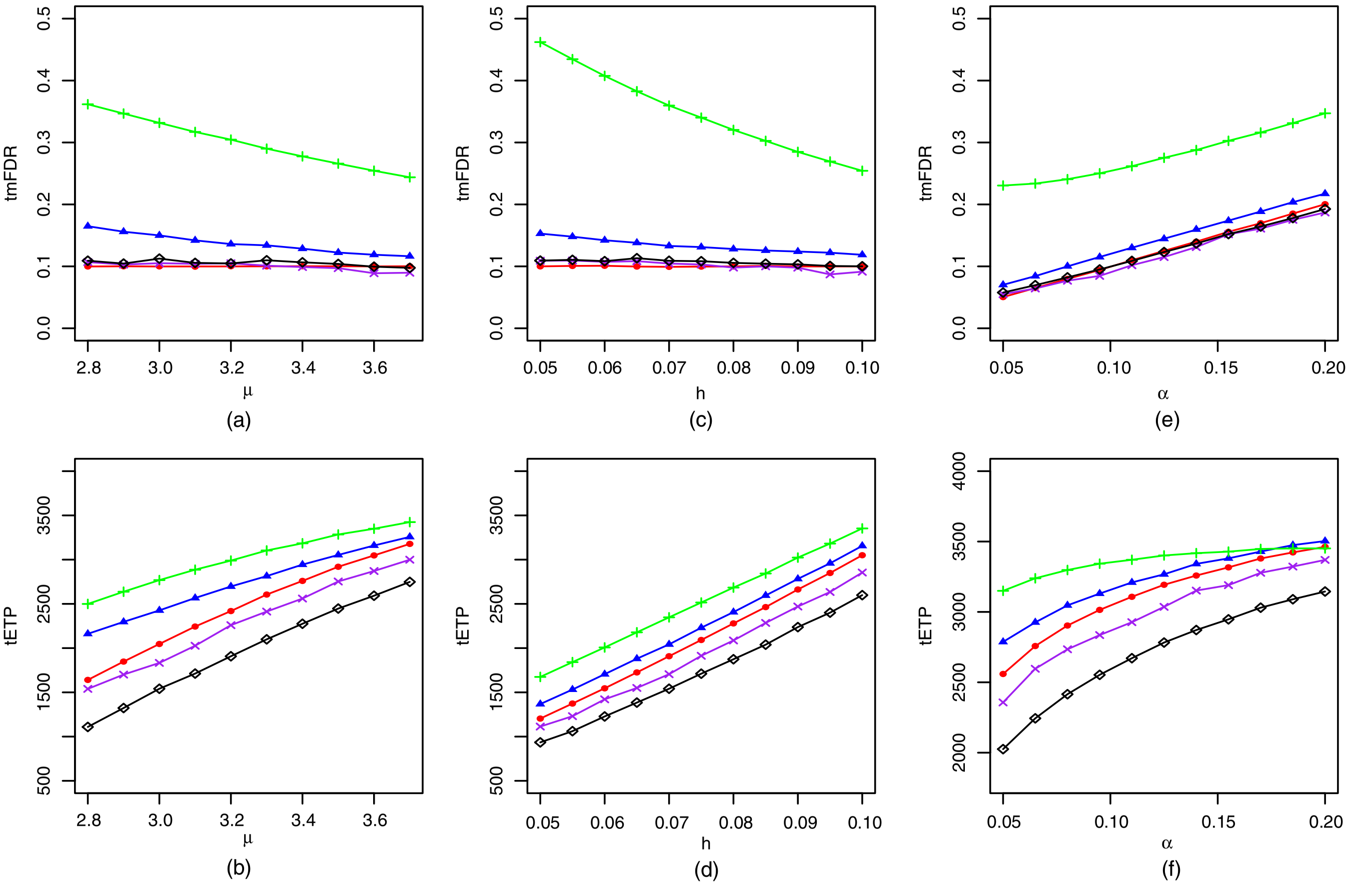


Fig. 1. Total marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0, 3\}$ or $\mathcal{S}_1 = \{1, 2\}$ (tmFDR, empirical total marginal false discovery rate (2); tETP, empirical total expected true positives number (1)) (●, oracle total control error procedure from theorem 1; ×, data-driven total error control procedure from theorem 3; ▲, method of Heller and Yekutieli (2014); +, separate discovery procedure; ◇, adjusted separate discovery procedure from Section 3.3): (a) $h = 0.1, \alpha = 0.1$; (b) $h = 0.1, \alpha = 0.1$; (c) $\mu = 3.6, \alpha = 0.1$; (d) $\mu = 3.6, \alpha = 0.1$; (e) $\mu = 3.6, h = 0.1$; (f) $\mu = 3.6, h = 0.1$

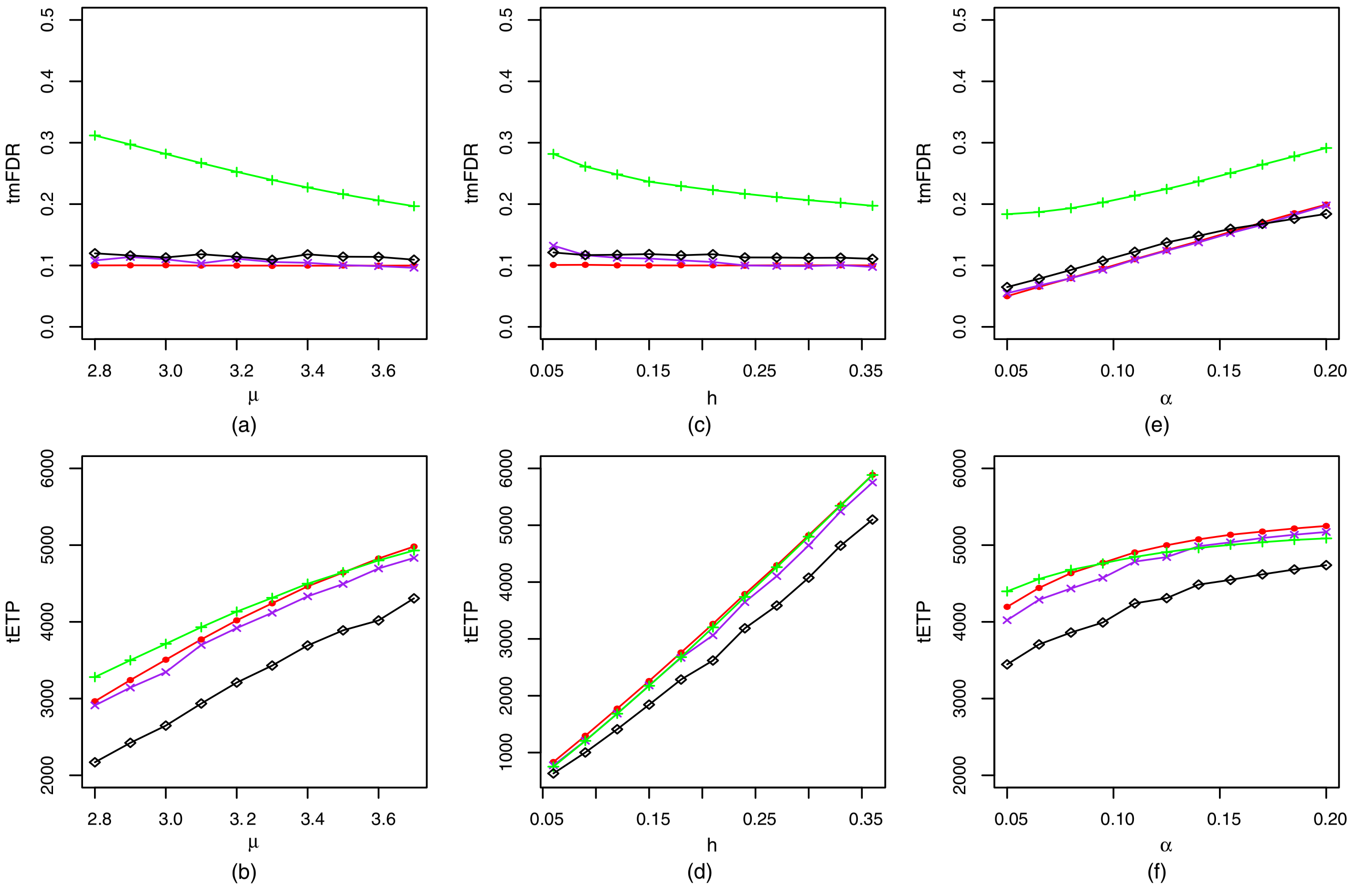


Fig. 2. Total marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$ or $\mathcal{S}_3 = \{3\}$ for independent test statistics (tmFDR, empirical total marginal false discovery rate (2); tETP, empirical total expected true positives number (1)) (●, oracle total error control procedure from theorem 1; ×, data-driven total error control procedure from theorem 3; +, separate discovery procedure; ◇, adjusted separate discovery procedure from Section 3.3): (a) $h = 0.3, \alpha = 0.1$; (b) $h = 0.3, \alpha = 0.1$; (c) $\mu = 3.6, \alpha = 0.1$; (d) $\mu = 3.6, \alpha = 0.1$; (e) $\mu = 3.6, h = 0.3$; (f) $\mu = 3.6, h = 0.3$

$(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) = (0.8 - h, h, h, 0.2 - h)$ 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 Fig. 1 show that the oracle, data-driven and adjusted separate discovery methods could all control the total marginal false discovery rate at the desired nominal level; the unadjusted separate discovery procedure was not. Among the former, 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 it 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 0.06 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 Fig. 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 0.06 and 0.36, and nominal set-specific marginal false discovery rates α_k (3) were varied between 0.05 and 0.2. For simplicity, all α_k were set to be equal for $k = 1, \ldots, 3$. Results are plotted in Fig. 3. The oracle and data-driven procedures again had nearly the same performance and uniformly dominated the adjusted separate discovery procedure. The unadjusted separate discovery procedure could not control the set-specific misclassification errors for all the sets of interest.

The fourth set of simulations explored the relationship between the total (4) and set-specific (5) error control problems for the multiclass problem with sets $\mathcal{S}_k = \{k\}$ for $k = 0, \ldots, 3$. Signal strengths were either set equal to μ in both sequences, or to $\mu - 1$ in sequence 1 and $\mu + 1$ in sequence 2, with μ varying between 4 and 5. This second setting models cases where the test statistics arise from studies with very different sample sizes. Test statistics from the smaller study will tend to have weaker effect sizes, corresponding to sequence 1. 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 $(0.7, 0.05, 0.05, 0.2)$.

Figs 4(a)–4(c) report the empirical set-specific false discovery rates of the oracle total error control procedure when the nominal total error was set to 0.1. In general, signals that were unique to sequence 1, belonging to set $\mathcal{S}_2$, were more difficult to discover compared with signals that were unique to sequence 2, belonging to $\mathcal{S}_1$. The realized set-specific marginal false discovery rate was always higher for $\mathcal{S}_2$ than for $\mathcal{S}_1$, and this difference increased as the signal strengths and the non-uniformity of the signal proportions increased. Furthermore, Figs 4(d)–4(f) show that the average number of features correctly assigned to $\mathcal{S}_2$ was also typically smaller than the number that was correctly assigned to $\mathcal{S}_1$ when the signal strength is weak and finally became larger than the number that was correctly assigned to $\mathcal{S}_2$ when the signal strength increased. These results motivate the proposed set-specific error control procedure.

Figs 4(g)–4(i) report the empirical total false discovery rates of the oracle set-specific error control procedure, where the nominal set-specific errors were set 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 could also control the total error when all α_k equalled the desired total nominal level, or when they equalled the induced

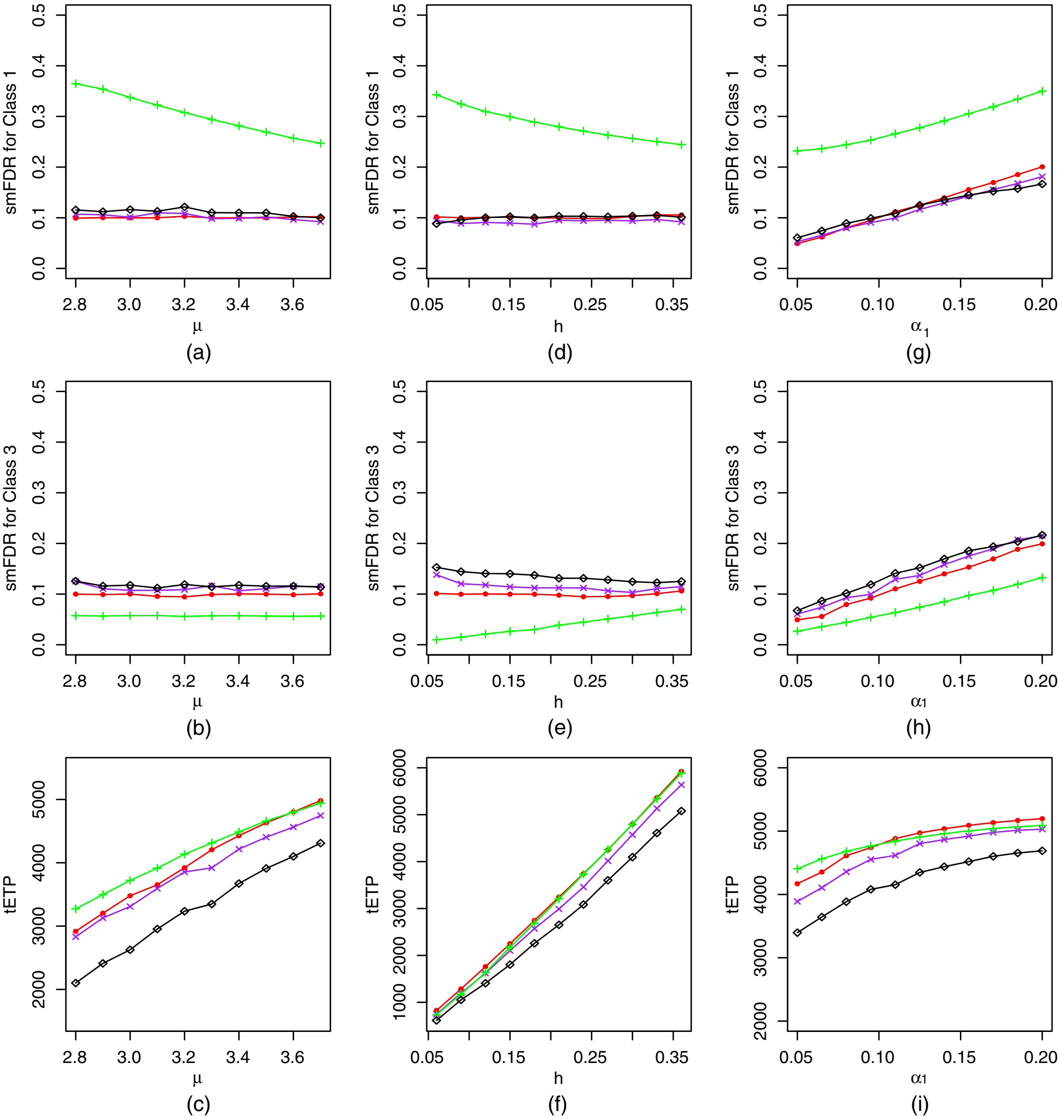
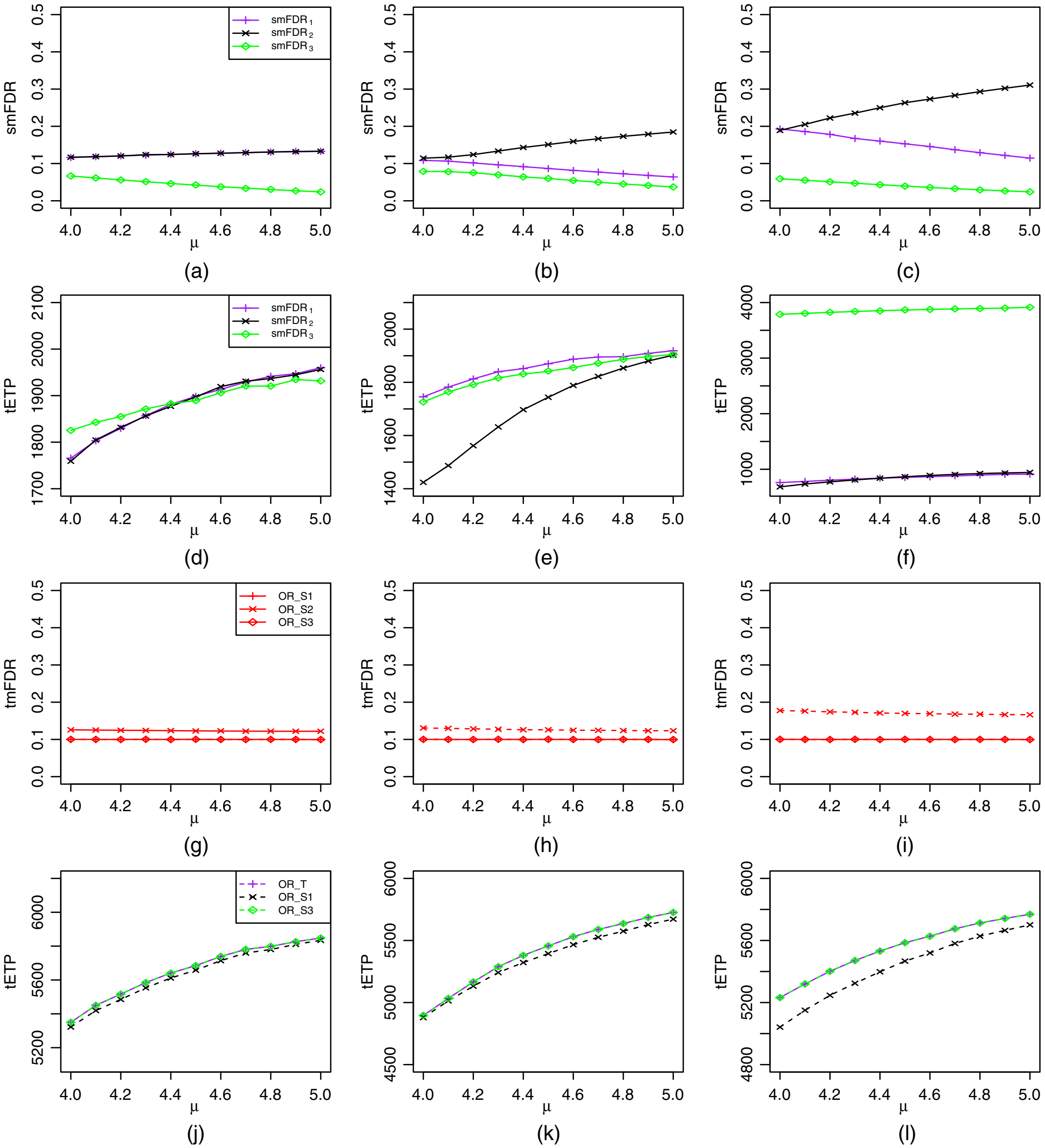


Fig. 3. Set-specific marginal false discovery rate control for classifying signals into $\mathcal{S}_0 = \{0\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$ or $\mathcal{S}_3 = \{3\}$ for independent test statistics (because of symmetry ($\alpha_1 = \alpha_2 = \alpha_3$), plots of the set-specific marginal false discovery rate for classification into $\mathcal{S}_2$ are identical to those for classification into $\mathcal{S}_1$ and therefore have been omitted; smFDR, empirical set-specific marginal false discovery rate (3); tETP, empirical total expected true positives number (1)) (●, oracle set-specific error control procedure from theorem 2; ×, data-driven set-specific error control procedure from theorem 4; +, separate discovery procedure; ◇ adjusted separate discovery procedure from Section 3.3): (a) $h=0.3$, $\alpha_1=0.1$; (b) $h=0.3$, $\alpha_1=0.1$; (c) $h=0.3$, $\alpha_1=0.1$; (d) $\mu=3.6$, $\alpha_1=0.1$; (e) $\mu=3.6$, $\alpha_1=0.1$; (f) $\mu=3.6$, $\alpha_1=0.1$; (g) $\mu=3.6$, $h=0.3$; (h) $\mu=3.6$, $h=0.3$; (i) $\mu=3.6$, $h=0.3$

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.

Figs 4(i)–4(j) report the realized average true positives number for both the total and the set-specific oracle procedures. For the former, the nominal total error was set to 0.1. For the latter, to conduct a fair comparison the nominal set-specific errors were set to be either $\alpha_1=\alpha_2=\alpha_3=0.1$,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

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 $\mathcal{S}_3 = \{3\}$ (smFD, empirical set-specific marginal false discovery rate (3); smFDR$_k$, 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, OR_S procedure with $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$; OR_S2, OR_S procedure with $\alpha_1 = 2\alpha_2 = \alpha_3/2 = 0.1$; OR_S3, OR_S procedure with the nominal set-specific errors induced by the OR_T procedure): (a) OR_T, $\mu_1 = \mu_2 = \mu$, uniform; (b) OR_T, $\mu_1 + 1 = \mu_2 - 2 = \mu$, uniform; (c) OR_T, $\mu_1 + 1 = \mu_2 - 2 = \mu$, non-uniform; (d) OR_T, $\mu_1 = \mu_2 = \mu$, uniform; (e) OR_T, $\mu_1 + 1 = \mu_2 - 2 = \mu$, uniform; (f) OR_T, $\mu_1 + 1 = \mu_2 - 2 = \mu$, non-uniform; (g) OR_S, $\mu_1 = \mu_2 = \mu$, uniform; (h) OR_S, $\mu_1 + 1 = \mu_2 - 2 = \mu$, uniform; (i) OR_S, $\mu_1 + 1 = \mu_2 - 2 = \mu$, non-uniform; (j) $\mu_1 = \mu_2 = \mu$, uniform; (k) $\mu_1 + 1 = \mu_2 - 2 = \mu$, uniform; (l) $\mu_1 + 1 = \mu_2 - 2 = \mu$, non-uniform

or to be 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.

Finally, the number of the features m also affected the performance and variability of the procedures proposed. Larger m gave more accurate data-driven procedures that were closer to their corresponding oracles. Details and additional simulation results are provided in section D of the on-line supplementary material.

5. Application to psychiatric genetics

The methods proposed 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 genomewide association studies: one of bipolar disorder, with 10410 cases and 10700 controls, and the other of schizophrenia, with 9369 cases and 8723 controls. These studies comprised completely independent samples with no shared controls. Summary z-scores are available from the web site of the Psychiatric Genomics Consortium. The data were first preprocessed by pruning the single-nucleotide polymorphisms (SNPs) 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. 439040 variants remained 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 that were associated only with schizophrenia, $\mathcal{S}_2$, containing SNPs that were associated only with bipolar disorder, and $\mathcal{S}_3$, containing SNPs that were 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 that were classified into each of the three classes. The majority of the discovered SNPs were classified into $\mathcal{S}_3$, which is consistent with previous work showing that bipolar disorder and schizophrenia have closely related genetic aetiologies (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

Table 5. Number of SNPs from Ruderfer *et al.* (2014) classified into different sets of interest†

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

†$\mathcal{S}_1$, SNPs associated only with schizophrenia; $\mathcal{S}_2$, SNPs associated only with bipolar disorder.

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 $\alpha_3 = 0.01$†

Results for class $\mathcal{S}_1$			*Result for 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			

†$\mathcal{S}_1$, 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.

(Ruderfer *et al.*, 2014). To address this problem, capturing SNPs that belong to $\mathcal{S}_1$ and $\mathcal{S}_2$ is more important than finding SNPs in $\mathcal{S}_3$, 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 $\mathcal{S}_2$ enable 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 that are specific to schizophrenia are all 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, which have been found to be specific to bipolar disorder, are 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 adenosine monophosphate signalling may differ between the two diseases. The latter SNP appears to regulate levels of a long non-coding ribonucleic acid (Lonsdale *et al.*, 2013), which may point to a new mechanism of action in bipolar disorder.

6. Dependent test statistics

Procedures that were developed in Section 3 rely heavily on the independence assumption, but test statistics in multiple-comparison problems can be dependent. This section further extends the proposed procedures to allow the test statistics X_{ji} within each sequence j to be correlated. Thus, the four-group model that is considered here turns out to be

$$\begin{aligned} &\mathbf{X}_j|\boldsymbol{\theta}_j \sim \mathbf{F}_j(\cdot|\boldsymbol{\theta}_j), && j=1,2, \\ &(\theta_{1i},\theta_{2i}) \overset{\text{IID}}{\sim} \sum_{l=0}^{3} \pi_l \theta_{1i}^{l_1}(1-\theta_{1i})^{1-l_1}\theta_{2i}^{l_2}(1-\theta_{2i})^{1-l_2}, && i=1,\ldots,m, \end{aligned} \tag{12}$$

where $\mathbf{X}_j=(X_{j1},\ldots,X_{jm})$, $\boldsymbol{\theta}_j=(\theta_{j1},\ldots,\theta_{jm})$, $\mathbf{X}_1$ is independent of $\mathbf{X}_2$ and $\boldsymbol{\theta}_2$, $\mathbf{X}_2$ is independent of $\mathbf{X}_1$ and $\boldsymbol{\theta}_1$, and $\mathbf{F}_j(\cdot|\boldsymbol{\theta}_j)$ is the joint distribution function of $\mathbf{X}_j$ conditional on $\boldsymbol{\theta}_j$.

The oracle procedures for the total and set-specific misclassification error control problems are now based on the oracle statistic

$$T_{ki}^{\mathrm{ORC}} = \frac{\sum_{l \notin \mathcal{S}_k} \pi_l \mathbf{f}_1(\mathbf{X}_1|\theta_{1i} = l_1)\mathbf{f}_2(\mathbf{X}_2|\theta_{2i} = l_2)}{\sum_{l=0}^{3} \pi_l \mathbf{f}_1(\mathbf{X}_1|\theta_{1i} = l_1)\mathbf{f}_2(\mathbf{X}_2|\theta_{2i} = l_2)} \tag{13}$$

where $\mathbf{f}_j(\mathbf{X}_j|\theta_{ji} = l_j)$ is the density corresponding to the distribution of $\mathbf{X}_j$ conditionally on $\boldsymbol{\theta}_{ji} = l_j$, $j = 1, 2$. For any $\lambda > 0$, define the total error classification rule $\boldsymbol{\delta}_T^{\lambda} = (\delta_{T1}^{\lambda}, \ldots, \delta_{Tm}^{\lambda})$ where

$$\delta_{Ti}^{\lambda} = \underset{k \in \{0,\ldots,K\}}{\arg\min} \sum_{k' \in \{1,\ldots,K\}, k' \neq k} \{(1 - T_{ki}^{\mathrm{ORC}}) + \lambda(T_{ki}^{\mathrm{ORC}} - \alpha)\}. \tag{14}$$

Similarly, for any $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ with $\lambda_k > 0$, define the set-specific error classification rule $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}} = (\delta_{S1}^{\boldsymbol{\lambda}}, \ldots, \delta_{Tm}^{\boldsymbol{\lambda}})$ where

$$\delta_{Si}^{\boldsymbol{\lambda}} = \underset{k \in \{0,\ldots,K\}}{\arg\min} \sum_{k' \in \{1,\ldots,K\}, k' \neq k} \{(1 - T_{ki}^{\mathrm{ORC}}) + \lambda_k(T_{ki}^{\mathrm{ORC}} - \alpha_k)\}. \tag{15}$$

Properties of these oracle estimators are given by the following theorem, whose proof is similar to those of theorems 1 and 2.

Theorem 5. Suppose that the continuous test statistics $\mathbf{X}_1$ and $\mathbf{X}_2$ are generated from the four-group model (12), and let $\boldsymbol{\delta}_T^{\lambda}$ and $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ be defined as in equations (14) and (15) respectively.

(a) For any $\lambda > 0$, let $N_T^{\mathrm{ORC}}(\lambda) = \mathbb{E}\{\Sigma_{k=1}^K I(\delta_{Ti}^{\lambda} = k)(T_{ki}^{\mathrm{ORC}} - \alpha)\}$ and define

$$\lambda^* = \inf\{\lambda : N_T^{\mathrm{ORC}}(\lambda) \leqslant 0\}.$$

If $\max_{l \in \bigcup_{k=1}^K \mathcal{S}_k} \pi_l > 0$ and $N_T^{\mathrm{ORC}}(0) \geqslant 0$ hold, then:

(i) $_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}_T^*) = \alpha$;

(ii) for any other classification rule $\boldsymbol{\delta}$ that satisfies $_{\mathrm{TM}}\mathrm{FDR}(\boldsymbol{\delta}) \leqslant \alpha$,

$$_{\mathrm{T}}\mathrm{ETP}(\delta_T^*) \geqslant {}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}).$$

(b) For any $\boldsymbol{\lambda}$, let $N_k^{\mathrm{ORC}}(\boldsymbol{\lambda}) = \mathbb{E}\{I(\delta_{Si}^{\boldsymbol{\lambda}} = k)(T_{ki}^{\mathrm{ORC}} - \alpha_k)\}$ and

$$\check{\lambda}_{k,n} = \inf\{\lambda_k \leqslant \check{\lambda}_{k,n-1} : N_k^{\mathrm{ORC}}(\check{\boldsymbol{\lambda}}_{k,n-1}) \leqslant 0\}, \qquad k = 1, \ldots, K,$$

where $n \geqslant 1$, $\check{\lambda}_{k,0} = \infty$, $\check{\boldsymbol{\lambda}}_{k,n-1}$ is the $\boldsymbol{\lambda}$ with $\lambda_{k'} = \check{\lambda}_{k',n-1}$, $k' \neq k$, and $\boldsymbol{\lambda}^* = \lim_{n\to\infty} \boldsymbol{\lambda}_{k,n}$. If $\alpha_k + \alpha_{k'} \leqslant 1$ for any $k \neq k'$, $\mathbf{0} \in \{(N_1^{\mathrm{ORC}}(\boldsymbol{\lambda}), \ldots, N_K^{\mathrm{ORC}}(\boldsymbol{\lambda})) : \boldsymbol{\lambda} \in \{\Re_+ \cup \{0\}\}^K\}$, and $\max_{l \in \mathcal{S}_k} \pi_l > 0$ for all $k \in \{1, \ldots, K\}$, then:

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

(ii) for any other classification rule $\boldsymbol{\delta}$ that satisfies $_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}) \leqslant \alpha_k$, $k = 1, \ldots, K$,

$$_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_S^*) \geqslant {}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}).$$

Because of the computational challenge of calculating $\mathbf{f}_j(\mathbf{X}_j|\theta_{ji} = l_j)$ and $\mathbf{f}_j(\mathbf{X}_j)$, it can be difficult to obtain the joint oracle statistics T_{ki}^{ORC} in equation (13) for $\mathbf{X}_j$ with arbitrary correlation. In these settings, theorem 5 is only of theoretical significance and cannot be used to develop data-driven procedures. However, prior information on correlation between features is sometimes known in practice. In particular, short-range dependence is one such informative structure for characterizing correlation between features within a sequence and has been widely employed for solving various practical problems, e.g. Xie *et al.* (2011).

The following results develop data-driven versions of the oracle procedures when test statistics follow a multivariate normal model with a short-range dependence covariance structure. First, proposition 3 shows that when the $\mathbf{X}_j$ are multivariate normal conditionally on $\boldsymbol{\theta}_j$, $j = 1, 2$, under some mild conditions the computational complexity can be reduced by ignoring the correlation entirely. Specifically, the $\mathbf{f}_j(\mathbf{X}_j|\theta_{ji} = l_j)$s in T_{ki}^{ORC} will be substituted with their corresponding marginal version $f_{jl_j}(X_j)$s, i.e. the marginal oracle statistics $T_k^{\mathrm{OR}}(X_{1i}, X_{2i})$ in equation (7) can be used to approximate the T_{ki}^{ORC}.

Proposition 3. Suppose that the continuous test statistics $\mathbf{X}_1$ and $\mathbf{X}_2$ are generated from the four-group model (12). Define $T_k^{\mathrm{OR}}(X_{1i}, X_{2i})$ as in equation (7) and T_{ki}^{ORC} as in equation (13). Assume that the following conditions hold.

Condition 3. The proportions π_l satisfy $cm^{-\tau_1} \leqslant \pi_l/\pi_{l'} \leqslant cm^{\tau_1}$ for all $l \neq l'$ for some constant $0 < \tau_1 < 1$ and $c > 0$.

Condition 4. The random vector $\mathbf{X}_j|\boldsymbol{\mu}_j, \boldsymbol{\theta}_j \sim N(\boldsymbol{\mu}_j \circ \boldsymbol{\theta}_j, \Sigma_j)$ where $\boldsymbol{\mu}_j \circ \boldsymbol{\theta}_j = (\mu_{j1}\theta_{j1}, \ldots, \mu_{jm}\theta_{jm})$, $j = 1, 2$. Here $\boldsymbol{\mu}_j$ is a random vector, with each μ_{ji} independently following a distribution $G_j(\cdot)$, satisfying, for some constant $\tau_2 > \tau_1$, $G_j[\surd\{2\tau_2 \log(m)\}] - G_j[-\surd\{2\tau_2 \log(m)\}] = 0$.

Condition 5. The covariance matrices Σ_j, $j = 1, 2$, are both positive definite with diagonal elements equal to 1.

Then, for all $\epsilon > 0$ and for all $i = 1, \ldots, m$ and $k = 1, \ldots, K$,

$$\lim_{m\to\infty} P\{|T_{ki}^{\mathrm{ORC}} - T_k^{\mathrm{OR}}(X_{1i}, X_{2i})| > \epsilon\} = 0. \tag{16}$$

Remark 5. Conditions 3–5 are mild and very similar to assumptions (A)–(C) in Xie *et al.* (2015).

With the result of proposition 3, it is reasonable to conjecture that the data-driven procedure that was proposed in Section 3 may be still useful. Theorem 6 in what follows confirms this and shows that the data-driven procedure is asymptotically valid and optimal under the condition that the dependences are short range.

Theorem 6. Suppose that all the assumptions in theorem 5 and proposition 3 hold, and that the marginal density function f_j is continuous and positive on the real line, and that its second derivative satisfies $\int (f_j'')^2 \mathrm{d}x_j < \infty$, $j = 1, 2$. Furthermore, assume that condition 1 holds and that the following condition holds.

Condition 6. The covariance matrices Σ_j obey $\sigma_{j,ik} = 0$ wherever $|i - k| \geqslant m^\tau$ for some constant $0 < \tau < 1$, where $\sigma_{j,ik}$ is the ikth entry of Σ_j.
Then:

(a)
- (i) $_{\mathrm{TM}}\mathrm{FDR}(\hat{\boldsymbol{\delta}}_T^*) = \alpha + o(1)$;
- (ii) $_{\mathrm{T}}\mathrm{ETP}(\hat{\boldsymbol{\delta}}_T^*)/_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_T^*) = 1 + o(1)$;

(b)
- (i) $_{\mathrm{SM}}\mathrm{FDR}_k(\hat{\boldsymbol{\delta}}_S^*) = \alpha_k + o(1)$, for all $k \in \{1, \ldots, K\}$;
- (ii) $_{\mathrm{T}}\mathrm{ETP}(\hat{\boldsymbol{\delta}}_S^*)/_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_S^*) = 1 + o(1)$.

Under short-range dependence, such as under condition 6 of theorem 6, the approach that was proposed by Jin and Cai (2007), with slight modifications, ensures that condition 1 holds. Meanwhile, these estimators attain their optimal rates of convergence. See Cai and Jin (2010) for more detail. In addition, condition 6, combined with the assumption that the second derivatives of the marginal densities are square integrable, implies condition 2 in theorems 3 and 4, leading to $\lim_{m\to\infty} P\{|T_{ki}^{\mathrm{ORC}} - \hat{T}_k(X_{1i}, X_{2i})| > \epsilon\} = 0$ uniformly for all i and k. In fact, $\hat{f}_j$ can be written as $\hat{f}_j = (1/m^\tau)\Sigma_{b=1}^{m^\tau} \hat{f}_{jb}$, where $\hat{f}_{jb}$ is a kernel estimator of f_j based on

$\mathbf{X}_{jb} = (X_{jb,1}, X_{jb,2}, \ldots, X_{jb,m^{1-\tau}})$, with $X_{jb,k} = X_{j(k-1)m^{\tau}+b}$ and $b = 1, \ldots, m^{\tau}$. Because of the independence of $\mathbf{X}_{jb}$ between components, it can be checked that $\mathbb{E}\| \hat{f}_{jb} - f_j \| \rightarrow 0$ as $m \rightarrow \infty$. Thus, $\mathbb{E}\| \hat{f}_j - f_j \| = (1/m^{2\tau}) \Sigma_{b=1}^{m^{\tau}} \mathbb{E}\| \hat{f}_{jb} - f_j \| \rightarrow 0$ as $m \rightarrow \infty$. Simulation results in the on-line supplementary material show that the data-driven procedure approximates the joint oracle procedure very well and performs well with dependent test statistics.

7. Discussion

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

The methods were developed under the assumption that the test statistics are independent across features, and Section 6 shows that they are robust to short-range dependence when the test statistics are normally distributed. Developing a procedure for general distributions and/or dependence structures is still an open problem. Moreover, the procedures proposed are established under the assumption that the $(\theta_{1i}, \theta_{2i})$s are mutually independent across i. However, in some situations, $(\theta_{1i}, \theta_{2i})$s may be correlated. For example, in microarray experiments, genes belonging to the same biological pathway may share similar significance patterns and, in public health surveillance studies, data from different time periods and locations are often serially or spatially correlated. For a single sequence of test statistics, Sun and Cai (2009) developed an optimal testing procedure using a hidden Markov model to model the dependence. This same strategy, extended to a four-state hidden Markov model, could be fruitful for signal classification problems and will be left for future study.

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 accordingly. However, the current implementation of these methods can grow unwieldy as the number of possible signal classes increases. In addition, 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 procedures proposed. 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.*, 2018; Ramdas *et al.*, 2019). However, it is unclear how these methods can be applied to multiple sequences of tests. These issues will be further studied in future work.

Acknowledgements

The authors greatly acknowledge the efforts of the Joint Editor, Associate Editor and two referees that have resulted in great improvements of this paper. Dongdong Xiang was supported in part by National Science Foundation China (grants 11501209, 11571113, 11771145, 71602115 and 71772147), the Postdoctoral Science Foundation of China (grants 20160089 and 2015M570348), KLATASDS-MOE, School of Statistics, East China Normal University and the Fundamental Research Funds for the Central Universities and the 111 Project (B14019). Dave Zhao was supported in part by National Science Foundation grant DMS 1613005. Tony Cai was supported in part by National Science Foundation grant DMS-1712735 and National Institutes of Health grants R01-GM129781 and R01-GM123056.

Appendix A: Proofs of some theoretical results

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

A.1. Proof of proposition 2

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

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

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

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

where $\boldsymbol{\delta}_T^{\boldsymbol{\lambda}}=(\delta_{S1}^{\boldsymbol{\lambda}},\ldots,\delta_{Sm}^{\boldsymbol{\lambda}})$. Take the expectation of both sides; then

$$\mathbb{E}\{L_{\mathrm{S}}(\boldsymbol{\lambda}, \boldsymbol{\delta}_T^{\boldsymbol{\lambda}})\} \leqslant \mathbb{E}\{L_{\mathrm{S}}(\boldsymbol{\lambda}, \boldsymbol{\delta})\}$$

holds for any $\boldsymbol{\delta} \in \{0,1\}^m$.

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

For ease of presentation, in this proof the $T_k^{\mathrm{OR}}(X_{1i}, X_{2i})$ will sometimes be abbreviated as $T_{k,i}^{\mathrm{OR}}$. Let

$$A_{\lambda_k}=\left\{T_{k,i}^{\mathrm{OR}} \leqslant \alpha_k+\frac{1-\alpha_k}{\lambda_k+1}\right\}$$

and

$$B_{\lambda_k}=\{\lambda_k(T_{k,i}^{\mathrm{OR}}-\alpha_k)+T_{k,i}^{\mathrm{OR}}<\min_{k'\neq k}\lambda_{k'}(T_{k',i}^{\mathrm{OR}}-\alpha_k)+T_{k',i}^{\mathrm{OR}}\};$$

then

$$\begin{aligned}N_k^{\mathrm{OR}}(\boldsymbol{\lambda})=E\bigg(I\bigg[&T_{k,i}^{\mathrm{OR}} \leqslant \alpha_k+\frac{1-\alpha_k}{\lambda_k+1}, \lambda_k(T_{k,i}^{\mathrm{OR}}-\alpha_k)+T_{k,i}^{\mathrm{OR}}\\ &<\min_{k'\neq k}\{\lambda_{k'}(T_{k',i}^{\mathrm{OR}}-\alpha_k)+T_{k',i}^{\mathrm{OR}}\}\bigg]\{T_{k,i}^{\mathrm{OR}}(X_{1i}, X_{2i})-\alpha_k\}\bigg)\\ =\mathbb{E}[&I_{A_{\lambda_k}}I_{B_{\lambda_k}}\{T_{k,i}^{\mathrm{OR}}(X_{1i}, X_{2i})-\alpha\}].\end{aligned}$$

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 $T_{k,i}^{\mathrm{OR}}<\alpha_k$, then $1-T_{k',i}^{\mathrm{OR}}+1-T_{k,i}^{\mathrm{OR}} \leqslant 1$ and $T_{k',i}^{\mathrm{OR}}>1-\alpha_k \geqslant \alpha_{k'}$. Thus,

$$\lambda_k(T_{k,i}^{\mathrm{OR}}-\alpha_k)+T_{k,i}^{\mathrm{OR}}<\min_{k'\neq k}\lambda_{k'}(T_{k',i}^{\mathrm{OR}}-\alpha_k)+T_{k',i}^{\mathrm{OR}}$$

will always hold for any λ_k, i.e. $B_{\lambda_k^{(1)}} \cap \{T_{k,i}^{\mathrm{OR}}<\alpha_k\}=B_{\lambda_k^{(2)}} \cap \{T_{k,i}^{\mathrm{OR}}<\alpha_k\}$.

If $T_{k,i}^{\mathrm{OR}} \geqslant \alpha_k$, then $\lambda_k^{(1)}(T_{k,i}^{\mathrm{OR}}-\alpha_k)+T_{k,i}^{\mathrm{OR}} \geqslant \lambda_k^{(2)}(T_{k,i}^{\mathrm{OR}}-\alpha_k)+T_{k,i}^{\mathrm{OR}}$ and thus $B_{\lambda_k^{(1)}} \cap \{T_{k,i}^{\mathrm{OR}} \geqslant \alpha_k\} \subseteq B_{\lambda_k^{(2)}} \cap \{T_{k,i}^{\mathrm{OR}} \geqslant \alpha_k\}$. Till now, $B_{\lambda_k^{(1)}} \subseteq B_{\lambda_k^{(2)}}$ has been 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

$$N_k^{\mathrm{OR}}(\boldsymbol{\lambda}_1)-N_k^{\mathrm{OR}}(\boldsymbol{\lambda}_2)=\mathbb{E}[(I_{A_{\lambda_k^{(1)}}}I_{B_{\lambda_k^{(1)}}}-I_{A_{\lambda_k^{(2)}}}I_{B_{\lambda_k^{(2)}}})I(T_k^{\mathrm{OR}} \geqslant \alpha_k)\{T_{k,i}^{\mathrm{OR}}(X_{1i}, X_{2i})-\alpha_k\}] \leqslant 0,$$

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

The result of part (b) in proposition 2 can now be proved. It follows from lemma 1 in the on-line supplementary file that there is a $\boldsymbol{\lambda}^{**}\in\Lambda^{**}$ such that our constructed K sequences $\{\check{\lambda}_{k,n}, n\geqslant 1\}$ satisfy the relationships $\check{\lambda}_{k,1}\geqslant\ldots\geqslant\check{\lambda}_{k,n}\geqslant\ldots\geqslant\lambda_k^{**}$ and that $N_k^{\mathrm{OR}}(\check{\boldsymbol{\lambda}}'_{k,n})=0$ holds for $k=1,\ldots,K$ and $n\geqslant 1$.

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

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

A.2. Proof of theorem 2

(a) Similarly to the proof of part (a) of theorem 1, the results for ${}_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta}_S^*)=\alpha_k$ for $k=1,\ldots,K$ are very straightforward.

(b) For any $\boldsymbol{\delta}$, if ${}_{\mathrm{SM}}\mathrm{FDR}_k(\boldsymbol{\delta})\leqslant\alpha_k$, for all $k=1,\ldots,K$, then

$$\begin{aligned}
&\mathbb{E}\left(\sum_{i=1}^{m}\sum_{k=1}^{K}[\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}]-I(\delta_{Si}^*=k)\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}]\right)\\
&\qquad=\mathbb{E}\left(\sum_{i=1}^{m}\sum_{k=1}^{K}[I(\delta_{Si}^*\neq k)\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}+\lambda_k^* I(\delta_{Si}^*=k)\{T_k^{\mathrm{OR}}(X_{1i},X_{2i})-\alpha_k\}]\right)\\
&\qquad=\mathbb{E}\{L_S(\boldsymbol{\lambda}^*,\boldsymbol{\delta}_S^*)\}\leqslant\mathbb{E}\{L_S(\boldsymbol{\lambda}^*,\boldsymbol{\delta})\}\\
&\qquad=\mathbb{E}\left(\sum_{i=1}^{m}\sum_{k=1}^{K}[I(\delta_i\neq k)\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}+\lambda_k^* I(\delta_i=k)\{T_k^{\mathrm{OR}}(X_{1i},X_{2i})-\alpha_k\}]\right)\\
&\qquad\leqslant\mathbb{E}\left(\sum_{i=1}^{m}\sum_{k=1}^{K}[\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}-I(\delta_i=k)\{1-T_k^{\mathrm{OR}}(X_{1i},X_{2i})\}]\right).
\end{aligned}$$

Thus, ${}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_S^*)\geqslant{}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta})$ holds.

A.3. Proof of theorem 4

For ease of presentation, in this proof the $T_k^{\mathrm{OR}}(X_{1i},X_{2i})$ and $\hat{T}_k(X_{1i},X_{2i})$ will be denoted as $T_{k,i}^{\mathrm{OR}}$ and $\hat{T}_{k,i}$ respectively. Let $\hat{N}_k^{\mathrm{OR}}(\boldsymbol{\lambda})=(1/m)\Sigma_{i=1}^m I(\delta_{Si}^{\boldsymbol{\lambda}}=k)(T_{k,i}^{\mathrm{OR}}-\alpha_k)$. According to the weak law of large numbers, result (a) $\hat{N}_k^{\mathrm{OR}}(\check{\boldsymbol{\lambda}}_{k,n-1})\to^{\mathrm{p}}N_k^{\mathrm{OR}}(\check{\boldsymbol{\lambda}}_{k,n-1})$ holds where $\check{\boldsymbol{\lambda}}_{k,n-1}$ is defined in the 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 $\hat{N}_k^{\mathrm{C}}(\boldsymbol{\lambda}$, can be defined, which is similar to the definition of $\hat{N}_T^{\mathrm{C}}(\lambda)$ in the proof of theorem 3. It is easy to check that $\hat{N}_k^{\mathrm{C}}(\boldsymbol{\lambda})$ is continuous in λ_k and monotone. Thus, its inverse function, denoted $\hat{N}_k^{\mathrm{C},-1}(\boldsymbol{\lambda})$, is well defined, continuous and monotone. According to the construction of the $\hat{N}_k^{\mathrm{C}}(\boldsymbol{\lambda})$, results (b) $\hat{N}_k(\hat{\boldsymbol{\lambda}}_{k,n-1})-\hat{N}_k^{\mathrm{C}}(\hat{\boldsymbol{\lambda}}_{k,n-1})\to^{\mathrm{p}}0$ and (c) $\hat{\lambda}_{k,n}-\hat{N}_k^{\mathrm{C},-1}(\hat{\boldsymbol{\lambda}}_{k0,n-1})\to^{\mathrm{p}}0$ hold for all k.

Suppose that $\hat{\lambda}_{k',n-1}\to^{\mathrm{p}}\check{\lambda}_{k',n-1}$ for all $k'\neq k$; results (d) and (e) can then be derived immediately, respectively $\hat{N}_k^{\mathrm{C}}(\check{\boldsymbol{\lambda}}_{k,n-1})-\hat{N}_k^{\mathrm{C}}(\hat{\boldsymbol{\lambda}}_{k,n-1})\to^{\mathrm{p}}0$ and $\hat{N}_k^{\mathrm{C},-1}(\hat{\boldsymbol{\lambda}}_{k0,n-1})-\hat{N}_k^{\mathrm{C},-1}(\check{\boldsymbol{\lambda}}_{k0,n-1})\to^{\mathrm{p}}0$ 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 the 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 will be discussed in turn. Suppose that $\hat{\lambda}_{k',n-1}\to^{\mathrm{p}}\check{\lambda}_{k',n-1}$ for all $k'\neq k$, then:

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

(b) $\hat{N}_k^{\mathrm{C},-1}(\check{\boldsymbol{\lambda}}_{k0,n-1})\to^{\mathrm{p}}\check{\lambda}_{k,n}$ and $\hat{\lambda}_{k,n}\to^{\mathrm{p}}\check{\lambda}_{k,n}, n\geqslant 1$.

A.3.1. *Proof of result (a)*

From the proof of result (1) in the proof of theorem 3 in the on-line supplementary material, it suffices to show that

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

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

Following from lemma 2 in the on-line supplementary file,

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

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,

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

where $\mathbb{E}\{(\hat{T}_{k,i}-T_{k,i}^{\mathrm{OR}})^2\}=o(1)$ follows from the results that $\hat{T}_{k,i}-T_{k,i}^{\mathrm{OR}}\rightarrow^{\mathrm{p}}0$ and $|\hat{T}_{k,i}-T_{k,i}^{\mathrm{OR}}|\leqslant 1$ uniformly for i.

A.3.2. Proof of result (b)

Similarly to the proof of result (2) in the proof of theorem 3 in the on-line supplementary material it suffices to prove that

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

which follows from the above results (1), (a), (b) and (d). Therefore,

$$\hat{N}_k^{\mathrm{C},-1}(\check{\boldsymbol{\lambda}}_{k0,n-1})\xrightarrow{\mathrm{p}}\check{\lambda}_{k,n}.$$

By this result, together with results (c) and (e), $\hat{\lambda}_{k,n}\rightarrow^{\mathrm{p}}\check{\lambda}_{k,n}$ can be obtained.

The result of theorem 4 can now be proved. When $n=1$, $\hat{\lambda}_{k'}=\check{\lambda}_{k'}=\infty$ holds for all $k'\neq k$; thus $\hat{\lambda}_{k,n}\rightarrow^{\mathrm{p}}\check{\lambda}_{k,n}$. Repeating to apply result (b), we have

$$\hat{\lambda}_{k,n}\xrightarrow{\mathrm{p}}\check{\lambda}_{k,n},\qquad n\geqslant 1.$$

Taking the limitations on both sides leads to $\hat{\lambda}_k^*\rightarrow^{\mathrm{p}}\lambda_k^*$.

Following from lemma 2 in the on-line supplementary file, we have

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

and similarly

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

Then,

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

By the above result, it is easy to show that

$$\begin{aligned}|\mathbb{E}\{(1/m)\sum_{i=1}^{m}(T_{k,i}^{\mathrm{OR}} - \alpha_k) I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k)\}| &= |\mathbb{E}[(T_{k,i}^{\mathrm{OR}} - \alpha)\{I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k) - I(\delta_{Si}^{\boldsymbol{\lambda}^*} = k)\}]| \\ &\leqslant \mathbb{E}\{|I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k) - I(\delta_{Si}^{\boldsymbol{\lambda}^*} = k)|\} = o(1), \qquad (17)\end{aligned}$$

$$|\mathbb{E}[(1/m)\sum_{i=1}^{m}(1 - T_{k,i}^{\mathrm{OR}})\{I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k) - I(\delta_{Si}^{\boldsymbol{\lambda}^*} = k)\}]| \leqslant \mathbb{E}\{|I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k) - I(\delta_{Si}^{\boldsymbol{\lambda}^*} = k)|\} = o(1), \qquad (18)$$

and

$$E\left\{(1/m)\sum_{i=1}^{m} I(\delta_{Si}^{\hat{\boldsymbol{\lambda}}^*} = k)\right\} = E\left\{(1/m)\sum_{i=1}^{m} I(\delta_{Si}^{\boldsymbol{\lambda}^*} = k)\right\} + o(1) > 0. \qquad (19)$$

By expressions (17) and (19), the result that ${}_{\mathrm{SM}}\mathrm{FDR}_k(\hat{\boldsymbol{\delta}}_S^*) = \alpha_k + o(1)$ can be derived. By expression (18), the result that ${}_{\mathrm{T}}\mathrm{ETP}(\hat{\boldsymbol{\delta}}_S^*)/{}_{\mathrm{T}}\mathrm{ETP}(\boldsymbol{\delta}_S^*) = 1 + o(1)$ can be derived. Then, the proof of theorem 4 is completed.

References

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

Basu, P., Cai, T. T., Das, K. and Sun, W. (2018) Weighted false discovery rate control in large-scale multiple testing. *J. Am. Statist. Ass.*, **113**, 1172–1183.

Benjamini, Y., Heller, R. and Yekutieli, D. (2009) Selective inference in complex research. *Phil. Trans. R. Soc. Lond.* A, **367**, 4255–4271.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Bogomolov, M. and Heller, R. (2013) Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Am. Statist. Ass.*, **108**, 1480–1492.

Cai, T. T. and Jin, J. (2010) Optimal rates of convergence for estimating the null density and proportion of non-null effects in large-scale multiple testing. *Ann. Statist.*, **38**, 100–145.

Cai, T. T. and Sun, W. (2017) Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *J. R. Statist. Soc.* B, **79**, 197–223.

Chi, Z. (2008) False discovery rate control with multivariate *p*-values. *Electron. J. Statist.*, **2**, 368–411.

Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2009) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLOS Genet.*, **10**, article e1004787.

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

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

Du, L. and Zhang, C. (2014) Single-index modulated multiple testing. *Ann. Statist.*, **42**, 30–79.

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

Forsyth, D. A. and Ponce, J. (2003) *Computer Vision: a Modern Approach*. Englewood Cliffs: Prentice Hall.

Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B, **64**, 499–517.

Genovese, C. and Wasserman, L. (2004) A stochastic process approach to false discovery control. *Ann. Statist.*, **32**, 1035–1061.

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

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

Hawkins, R. D., Hon, G. C. and Ren, B. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, **11**, 476–486.

Heller, R., Bogomolov, M. and Benjamini, Y. (2014) Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc. Natn. Acad. Sci. USA*, **111**, 16262–16267.

Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. *Ann. Appl. Statist.*, **8**, 481–498.

Huang, J., Perlis, R. H., Lee, P. H., Rush, A. J., Fava, M., Sachs, G. S., Lieberman, J., Hamilton, S. P., Sullivan, P., Sklar, P., Purcell, S. and Smoller, J. W. (2010) Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. *Am. J. Psychiatr.*, **167**, 1254–1263.

Jin, J. and Cai, T. T. (2007) Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Am. Statist. Ass.*, **102**, 495–506.

Kim, J. and Scott, C. (2012) Robust kernel density estimation. *J. Mach. Learn. Res.*, **13**, 2529–2565.

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

Li, H. (2013) Systems biology approaches to epidemiological studies of complex diseases. *Syst. Biol. Med.*, **5**, 677–686.

Li, G., Jima, D., Wright, F. A. and Nobel, A. B. (2018a) Ht-eqtl: integrative expression quantitative trait loci analysis in a large number of human tissues. *BMC Bioinform.*, **19**, article 95.

Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A. and Nobel, A. B. (2018b) An empirical bayes approach for multiple tissue eqtl analysis. *Biostatistics*, **19**, 391–406.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J. M., Wilder, E. L., Derr, L. K., Green, E. D., Struewing, J. P., Temple, G., Volpi, S., Boyer, J. T., Thomson, E. J., Guyer, M. S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T. R., Koester, S. E., Little, A. R., Bender, P. K., Lehner, T., Yao, Y., Compton, C. C., Vaught, J. B., Sawyer, S., Lockhart, N. C., Demchok, J. and Moore, H. F. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

McHugh, J. M., Konrad, J., Saligrama, V., Jodoin, P.-M. and Castanón, D. (2008) Motion detection with false discovery rate control. In *Image Processing, 2008*, pp. 873–876. New York: Institute of Electrical and Electronics Engineers.

Mühleisen, T. W., Leber, M., Schulze, T. G., Strohmaier, J., Degenhardt, F., Treutlein, J., Mattheisen, M., Forstner, A. J., Schumacher, J., Breuer, R., Meier, S., Herms, S., Hoffmann, P., Lacour, A., Witt, S. H., Reif, A., Müller-Myhsok, B., Lucae, S., Maier, W., Schwarz, M., Vedder, H., Kammerer-Ciernioch, J., Pfennig, A., Bauer, M., Hautzinger, M., Moebus, S., Priebe, L., Czerski, P. M., Hauser, J., Lissowska, J., Szeszenia-Dabrowska, N., Brennan, P., McKay, J. D., Wright, A., Mitchell, P. B., Fullerton, J. M., Schofield, P. R., Montgomery, G. W., Medland, S. E., Gordon, S. D., Martin, N. G., Krasnow, V., Chuchalin, A., Babadjanova, G., Pantelejeva, G., Abramova, L. I., Tiganov, A. S., Polonikov, A., Khusnutdinova, E., Alda, M., Grof, P., Rouleau, G. A., Turecki, G., Laprise, C., Rivas, F., Mayoral, F., Kogevinas, M., Grigoroiu-Serbanescu, M., Propping, P., Becker, T., Rietschel, M., Nöthen, M. M. and Cichon, S. (2014) Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Communs*, **5**, article 3339.

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

Ramdas, A., Barber, R. F., Wainwright, M. J. and Jordan, M. I. (2019) A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.*, to be published.

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

Roeder, K. and Wasserman, L. (2009) Genome-wide significance levels and weighted hypothesis testing. *J. Am. Statist. Ass.*, **24**, 398–413.

Roquain, E. and de Wiel, M. A. (2009) Optimal weighting for false discovery rate control. *Electron. J. Statist.*, **3**, 678–711.

Ruderfer, D. M., Fanous, A. H., Ripke, S., McQuillin, A., Amdur, R. L., 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, Gejman, P. V., O'Donovan, M. C., Andreassen, O. A., Djurovic, S., Hultman, C. M., Kelsoe, J. R., Jamain, S., Landén, M., Leboyer, M., Nimgaonkar, V., Numberger, J., Smoller, J. W., Craddock, N., Corvin, A., Sullivan, P. F., Holmans, P., Sklar, P. and Kendler, K. S. (2014) Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molec. Psychiatr.*, **19**, 1017–1024.

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

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

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Sun, W. and Cai, T. T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Ass.*, **102**, 901–912.

Sun, W. and Cai, T. T. (2009) Large-scale multiple testing under dependence. *J. R. Statist. Soc.* B, **71**, 393–424.

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

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

Urbut, S. M., Wang, G., Carbonetto, P. and Stephens, M. (2019) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions.*Nat. Genet.*, **51**, 187–195.

Xie, J., Cai, T. and Li, H. (2015) Correction to the paper "Optimal false discovery rate control for dependent data". *Statist. Interfc.*, **9**, 33–35.

Xie, J., Cai, T., Maris, J. and Li, H. (2011) Optimal false discovery rate control for dependent data. *Statist. Interfc.*, **4**, 417–430.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary file of "Signal classification for the integrative analysis of multiple sequences of large-scale multiple tests"'.