Signal Classification in Large-Scale Multi-Sequence Integrative Analysis Under the HMM Dependence

Wendong Li[1], Dongdong Xiang[2], Gongtao Chen[2], and Peihua Qiu[3]

[1]School of Statistics and Management, Shanghai Institute of International Finance and Economics, Shanghai University of Finance and Economics, Shanghai, China

[2] KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China

[3] Department of Biostatistics, University of Florida, Gainesville, USA

Abstract

The integrative analysis of multiple sequences of multiple tests has enjoyed increasing popularity in many applications, especially in large-scale genomics. In the context of large-scale multiple testing, the concept of signal classification has been developed recently for cases when the same features are involved in several independent studies, with the goal of classifying each feature into one of several classes. This paper considers the problem of such signal classification in a generalized compound decision-making framework, where the observed data are assumed to be generated from an underlying four-state Cartesian hidden Markov model. Two oracle procedures are proposed for the total and set-specific control of misclassification rates, respectively, while the number of correct classifications is maximized. Optimal data-driven procedures are also proposed, with their asymptotic properties derived. It is shown that signal-classification could be improved significantly by taking into account the dependence structure among features, and the proposed procedures could have a better performance than their competitors that ignore the dependence structure. The proposed methods are applied to a psychiatric genetics study for detecting genetic variants that affect either or both of bipolar disorder and schizophrenia.

Keywords: Generalized local significance index; Hidden Markov model; Integrative analysis; Signal classification under dependence.

Corresponding author: Dongdong Xiang, email: terryxdd@163.com.

1 Introduction

In recent years, combined datasets from multiple studies become increasingly popular in genomics and genetics research, posing great challenges to the large-scale multiple testing problem. The integrative analysis of such data from multiple studies could provide much useful information by comparing different studies and finding their differences and similarities. Thus, alongside research on a single sequence of multiple tests (Genovese and Wasserman, 2002; Sun and Cai, 2007; Efron, 2007; Basu et al., 2018; Li et al., 2020; Cai, Sun and Xia, 2021), development of effective methods for multiple sequences of multiple tests should be important, which is the focus of this paper.

The large-scale multi-sequence integrative analysis in genomics and genetics research can be formulated generally as a problem of grouping genomic features from multiple studies into different classes based on their test statistics, which is briefly described below. Let $\boldsymbol{X}_i = (X_{i1}, ..., X_{im})$ be the vector of z-scores for the m genomic features in the ith study ($i = 1, ..., J$). Let us focus on the case when $J = 2$ in this paper for simplicity, and the extension to cases when $J > 2$ is quite straightforward. Assume that $\theta_{ij} \in \{0, 1\}$ is the unknown state of X_{ij} ($j = 1, ..., m$), where "$\theta_{ij} = 1$" if X_{ij} is a signal and "$\theta_{ij} = 0$" otherwise. Then, the pair $(\theta_{1j}, \theta_{2j})$ denotes the joint state of (X_{1j}, X_{2j}). Table 1 lists all four possible classes of $(\theta_{1j}, \theta_{2j})$ and their labels, which are also the four possible hidden states of (X_{1j}, X_{2j}). To illustrate, in cases when X_{ij} denotes the jth expression quantitative trait locus in the ith tissue, isolating tissue-specific loci from cross-tissue ones is equivalent to determining whether the related tests belong to the class 1, or 2, or 3 in Table 1.

Table 1: Four possible signal classes and labels for (X_{1j}, X_{2j}).

Class label	$(\theta_{1j}, \theta_{2j})$
0	(0,0)
1	(1,0)
2	(0,1)
3	(1,1)

In the multiple-testing problem described above, one major goal is to assign as many genomic features to their true signal classes as possible. Achieving this goal while controlling the misclassification rate is referred to as signal classification in the multiple-testing literature (Xiang, Zhao and Cai, 2019). Applications of signal classification are popular in modern genomics studies. For instance, to learn the genetic regulation of human gene expressions, data of genotype and gene expressions are often collected for many tissue types from many donors (Lonsdale et al., 2013). One major goal to analyze such data is to identify the specific genes that are regulated by certain genetic variants. Since a genetic variant may be active in a part of tissues only, it is crucial to classify each variant in terms of the related tissues (Torres et al., 2014; GTEx Consortium, 2015), which involves a large number of sequences of multiple tests. Similarly, large-scale genome-wide association studies (GWAS) have enabled researchers to compare the genetics of two clinically indistinguishable diseases that share many symptoms (e.g., bipolar disorder and schizophrenia). Isolating genetic variants that are significantly associated with one disease but not the other is often crucial to the development of effective disease diagnostic methods, which also requires an integrative analysis of two sequences of summary statistics: one from each disease.

As described above, proper analysis of multiple sequences of multiple tests becomes important due mainly to the rapid development of integrative genomics. However, most existing methods for analyzing such data are designed for the simplified binary classification problem in the sense that they typically focus on identifying signals that belong to the class 3 in Table 1 (Benjamini et al., 2009; Chung et al., 2014; Heller and Yekutieli, 2014; Wang et al., 2016; Wang and Zhu, 2019; Zhao and Nguyen, 2020). One main reason for this phenomenon is that the related replicability analysis for detecting replicated signals could obtain replicable scientific findings and provide useful information for genetic association studies. To handle such a binary classification problem, a four-group mixture model is usually used for describing the observed data $\{(X_{1j}, X_{2j}), j = 1, ..., m\}$ with four possible signal classes. It has been studied by many researchers about the optimal decision rule that is based on the local false discovery rate (Lfdr). Urbut et al. (2019) and Li et al. (2018a, b) have extended these results to cases with three or more sequences of multiple tests. In practice, however, identifying signals based on different sequences that belong to the class 3 in Table 1 may not be our primary interest. For instance, in a GWAS study involving bipolar disorder and schizophrenia, signals belonging to the class 1 or 2 in Table 1 should be more useful than those belonging to the class 3, in order to differentiate patients with bipolar disorder from those with schizophrenia. Recently, Xiang, Zhao and Cai (2019) introduced a new framework for the signal classification problem that allows for two or more signal classes of interest. Under the assumption that the test statistics are independent across different dimensions, they proposed total and set-specific indices for measuring misclassification errors and developed asymptotically optimal multiple-testing procedures with their misclassification errors in control, based on a generalized compound decision framework.

While many existing methods rely heavily on the assumption of independence, observed data from large-scale multiple-testing studies often exhibit data dependence (Sun and Cai, 2009). For example, in GWAS, disease-associated SNPs may cluster into groups along biological pathways, indicating a dependence structure. Ignoring this dependence can lead to invalidity or reduced efficiency of methods developed under the assumption of data independence. To the best of our knowledge, it has not been discussed in the literature yet how to accommodate the data dependence structure when handling the signal classification problem with multiple sequences of tests that allow for two or more sets of signal classes of interest. This paper aims to fill this gap by properly modeling the data dependence structure to gain a better understanding of multiple sequences of tests with multiple signal classes of interest.

In this paper, we focus on the four-class signal classification problem with data dependence. The main contributions of the paper are summarized as follows. From the modeling perspective, the signal classification problem under a Cartesian hidden Markov model (HMM) dependence structure is studied using the compound decision framework. The HMM is an effective tool for modeling the data dependence structure and has been widely used in areas such as speech recognition, signal processing and DNA sequence analysis. It assumes that the sequence of unknown joint states forms a four-state Markov chain $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$. When a positive dependence exists in an HMM, the signals belong to the same class would form clusters, which is commonly seen in many real applications.

For example, in GWAS, since the adjacent genomic loci tend to cosegregate in meiosis, the disease-associated SNPs are always clustered and exhibit high correlation. As a result, HMM has been successfully used in GWAS for modelling the clustered and locally dependent data structure (e.g., Colella et al., 2007; Wei et al., 2009; Bercovici et al., 2010; Wang and Zhu, 2019). To proceed, we first develop oracle procedures for signal classification in cases when the HMM parameters are assumed known based on a generalized local index of significance (LIS), which enables us to borrow information from observations in adjacent locations by exploiting the local dependence structure. We then develop data-driven procedures to mimic the oracle ones by plugging in consistent estimates of the HMM parameters. In addition, from the theoretical perspective, we show that the oracle procedures are optimal under some mild conditions in the sense that they can control the corresponding misclassification error while maximizing the number of correct classifications. We also provide the asymptotic optimality of the data-driven procedures.

The remainder of the paper is organized as follows. Section 2 provides a formulation of the signal classification problem under the HMM data dependence. Section 3 introduces the new oracle and data-driven procedures proposed for solving the signal classification problem, along with their related theoretical results in terms of validity and optimality. Simulation studies are presented in Section 4 to compare the proposed methods with some representative existing methods in various settings. In Section 5, the proposed data-driven procedures are applied to a genomic study for understanding the genetic architecture of bipolar disorder and schizophrenia. Finally, a summary of the contributions of this paper and some possible extensions of the proposed methods are discussed in Section 6. Proofs of some theoretical results are given in a supplementary file.

2 Problem Formulation

2.1 Some definitions

Let $\boldsymbol{X}_i = (X_{i1}, ..., X_{im})$ be the vector of m observed values in the ith study ($i = 1, ...J$). We focus on cases with $J = 2$ in this paper for simplicity, and extension of the proposed methods to cases with $J > 2$ should be straightforward. Assume that θ_{ij} is the unknown state of X_{ij}. More specifically, $\theta_{ij} = 1$ if X_{ij} is a signal and $\theta_{ij} = 0$ otherwise. Table 1 lists all four possible configurations of $(\theta_{1j}, \theta_{2j})$ and their labels, denoting four different states of (X_{1j}, X_{2j}). Different from most multiple-testing problems discussed in the literature that are about a single sequence, we are interested in classifying each genomic feature observed in the two sequences into four signal classes. This is a so-called *signal classification problem* (Xiang et al., 2019). To characterize the correlation structure of $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$, we consider using an HMM, in which it is assumed that $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$ is a stationary, irreducible and aperiodic four-state Markov chain with the transition probabilities

$$a_{uv} = P\{(\theta_{1,j+1}, \theta_{2,j+1}) = v|(\theta_{1j}, \theta_{2j}) = u\},$$

where $u, v \in \{(0,0), (1,0), (0,1), (1,1)\}$. It is also assumed that a_{uv}'s do not depend on j, $0 < a_{uv} < 1$ for each u and v, and $\sum_v a_{uv} = 1$ for each u.

Since X_{1j} and X_{2j} are usually observed from two independent studies, it is assumed that they are conditionally independent given $(\theta_{1j}, \theta_{2j})$, and $\{(X_{1j}, X_{2j}), j = 1, ..., m\}$ are independent over different j as well given $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$. Thus, the joint density function of the observed data $\{(X_{1j}, X_{2j}), j = 1, ..., m\}$ is

$$P\left\{\{(X_{1j}, X_{2j})\}_{j=1}^{m}|\{(\theta_{1j}, \theta_{2j})\}_{j=1}^{m}\right\} = \prod_{j=1}^{m} f(X_{1j}|\theta_{1j}) \prod_{j=1}^{m} f(X_{2j}|\theta_{2j}), \tag{1}$$

where $f(X_{ij}|\theta_{ij}) = (1 - \theta_{ij})f_{i0} + \theta_{ij} f_{i1}$, f_{i0} and f_{i1} are respectively the density functions of X_{ij} when $\theta_{ij} = 0$ and $\theta_{ij} = 1$, and $\{f_{i0}, i = 1, 2\}$ are known null distributions. In practice, we usually assume that f_{10} and f_{20} are the densities of $N(0, 1)$. Let $\boldsymbol{\pi} = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$ be the initial distribution of the Markov chain, $\mathcal{A} = \{a_{uv}\}_{4\times4}$ be the transition matrix, and

$\mathcal{F} = \{f_{10}, f_{11}, f_{20}, f_{21}\}$ be the observation distribution. Then, $\vartheta = (\boldsymbol{\pi}, \mathcal{A}, \mathcal{F})$ is the collection of all HMM parameters, where $\pi_{pq} = P\{(\theta_{11}, \theta_{21}) = (p, q)\}$, for $p, q \in \{0, 1\}$.

Remark 1. *In some applications, the four signal classes defined in Table 1 may not be equally important. Generally speaking, the set of four classes can be divided into $K + 1$ disjoint subsets, where K can be 1, 2, or 3. Let G_0 denote the set of unimportant classes (e.g., $G_0 = \{0\}$), and $\{G_k \subset \{0, 1, 2, , 3\} \backslash G_0, k = 1, ..., K\}$ be disjoint subsets of classes of interest with $\bigcup_{k=0}^{K} G_k = \{0, 1, 2, 3\}$. As an example, when $K = 1$, $G_0 = \{0, 1, 2\}$ and $G_1 = \{3\}$, the signal classification problem reduces to the replicability analysis in the literature for discovering significant features in both sequences (Heller and Yekutieli, 2014). In this paper, we focus on cases with $K = 3$ and $G_k = \{k\}$, for $k = 0, 1, 2, 3$. The extension of our proposed methods to cases with $K < 3$ is straightforward.*

2.2 Signal classification in a hidden Markov model

Given $\{G_k, k = 0, 1, 2, 3\}$, the decision rule of a signal classification procedure can be denoted as $\boldsymbol{\delta} = (\delta_1, ..., \delta_m)$, where $\delta_j \in \{0, ..., 3\}$ indicates which class subset the jth genomic feature should be assigned to, for $j = 1, ..., m$. The results of applying this decision rule to the observed data can then be summarized in Table 2. The expected total number of true positives is

$$\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta}) = E\left(\sum_{k=1}^{K} C_{kk}\right),$$

where the subscript "T" of $\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta})$ denotes "total". Obviously, the quantity $\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta})$ denotes the expected total number of tests that correctly classify the m genomic features into the K subsets of interest, which is commonly used for evaluating the power of $\boldsymbol{\delta}$. To measure the misclassification error of $\boldsymbol{\delta}$, there are several choices. One commonly-used metric is the following total marginal false discovery rate (mFDR):

$$\mathrm{mFDR}_{\mathrm{T}}(\boldsymbol{\delta}) = \frac{E(\sum_{k=1}^{K} \sum_{k' \neq k} C_{k'k})}{E(\sum_{k=1}^{K} R_k)}.$$

In the above expression, the denominator is the expected total number of tests that are classified into the K subsets of interest, and the numerator is the expected total number of misclassifications among them. In cases when $K = 2$, $\text{mFDR}_{\text{T}}(\boldsymbol{\delta})$ is just the conventional mFDR used in the binary classification problem. Thus, it can be regarded as a natural generalization to the multi-class signal classification problem, and provides a proper metric of the overall misclassification rate. Set-specific mFDR can be defined similarly for each subset of interest as follows:

$$\text{mFDR}_{\text{S}}^{k}(\boldsymbol{\delta}) = \frac{E(\sum_{k' \neq k} C_{k'k})}{E(R_k)}, \ k = 1, ..., K,$$

where the subscript "S" of $\text{mFDR}_{\text{S}}^{k}(\boldsymbol{\delta})$ denotes "subset", and the superscript "k" denotes the kth subset. This index measures the misclassification error for the subset G_k only, which should be more flexible than the metric $\text{mFDR}_{\text{T}}(\boldsymbol{\delta})$ in the sense that different nominal misclassification error rates can be set for different subsets.

The above two metrics of mFDR lead to the following two different signal classification problems. First, the *total mFDR-control* problem aims to find a decision rule $\boldsymbol{\delta}$ that maximizes $\text{ETP}_{\text{T}}(\boldsymbol{\delta})$ subject to the condition that $\text{mFDR}_{\text{T}}(\boldsymbol{\delta}) \leq \alpha$, for a given $0 < \alpha < 1$. Second, the *set-specific mFDR-control* problem aims to find a decision rule $\boldsymbol{\delta}$ that maximizes $\text{ETP}_{\text{T}}(\boldsymbol{\delta})$ subject to the condition that $\text{mFDR}_{\text{S}}^{k}(\boldsymbol{\delta}) \leq \alpha_k$, for $k = 1, ..., K$ and given $0 < \alpha_1, ..., \alpha_K < 1$. The latter problem enables a proper control of the misclassification errors for individual subsets of interest by setting $\{\alpha_k, k = 1, ..., K\}$. However, in cases when it is unclear how to choose $\{\alpha_k, k = 1, ..., K\}$, as is often the case in practice, the former problem might be more convenient to use by setting the total misclassification error control at a single level. Therefore, the two problems can complement each other well. In the following sections, we focus on both problems and propose the corresponding signal classification procedures.

3 Proposed Statistical Methodology

3.1 Oracle procedures

We first derive the oracle signal classification rules under the HMM model described in Section 2 for the total and set-specific mFDR-control problems, respectively, where the

Table 2: Summary of signal classification results.

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

word “oracle” implies that an ideal situation is considered, in which the HMM parameters $\vartheta = (\boldsymbol{\pi}, \mathcal{A}, \mathcal{F})$ are assumed known. To this end, it is straightforward to check that the total mFDR-control problem is equivalent to the following maximization problem:

$$\max_{\boldsymbol{\delta}} E\left\{\sum_{k=1}^{K}\sum_{j=1}^{m} I(\delta_j = k)[1 - \mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)]\right\}$$

under the condition that

$$E\left\{\sum_{k=1}^{K}\sum_{j=1}^{m} I(\delta_j = k)\left[\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha\right]\right\} \leq 0, \qquad (2)$$

where $\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) = 1 - P\{(\theta_{1j}, \theta_{2j}) = (k_1, k_2)|\boldsymbol{X}_1, \boldsymbol{X}_2\}$ is the conditional probability that the class of the jth genomic feature is not k, and k_1 and k_2 are the two signal indicators of the class k. It is important to notice that the quantity LIS_j^k can be regarded as a generalization of the LIS_j statistic in Sun and Cai (2009) that was designed for binary classification of a single sequence of multiple tests. By pooling information from two sequences, the quantity LIS_j^k should be more effective for signal classification.

The above optimization problem under the inequality constraint (2) can be solved by using the method of Lagrange multipliers by minimizing the Lagrangian objective function

$$L_T(\lambda, \boldsymbol{\delta}) = \sum_{k=1}^{K}\sum_{j=1}^{m} I(\delta_j \neq k)[1 - \mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)] + \sum_{k=1}^{K}\sum_{j=1}^{m} \lambda I(\delta_j = k)[\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha],$$

because any $\boldsymbol{\delta}$ that minimizes $L_T(\lambda, \boldsymbol{\delta})$ also minimizes $E\{L_T(\lambda, \boldsymbol{\delta})\}$. For a given $\lambda > 0$, the decision rule that minimizes $L_T(\lambda, \boldsymbol{\delta})$ is denoted as $\boldsymbol{\delta}_T^\lambda = (\delta_{T1}^\lambda, ..., \delta_{Tm}^\lambda)$, where

$$\delta_{Tj}^\lambda = \arg \min_{k \in \{0,...,K\}} \left\{ \left[\sum_{k' \in \{1,...,K\}, k' \neq k} [1 - \mathrm{LIS}_j^{k'}(\boldsymbol{X}_1, \boldsymbol{X}_2)] \right] + \lambda [\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha] \right\}, \qquad (3)$$

for $j = 1, ..., m$. The following proposition shows two important properties of $\boldsymbol{\delta}_T^\lambda$.

Proposition 1. *Consider an HMM model as defined in model (1) and the related decision rule $\boldsymbol{\delta}_T^\lambda$ defined in (3). Then, we have*

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

(ii) Let $N(\lambda) = E\left\{\sum_{k=1}^K I(\delta_{Tj}^\lambda = k)[\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha]\right\}$, and

$$\lambda^* = \inf\{\lambda : N(\lambda) \leq 0\}.$$

Then, $N(\lambda^*) = 0$ if $N(0) \geq 0$.

The quantity $N(\lambda)$ is obtained by combining the constraint on $\mathrm{mFDR_T}$ in (2) and the decision rule defined in (3), which can be regarded as the difference between the actual and pre-specified misclassification error rates of the decision rule $\boldsymbol{\delta}_T^\lambda$. It is straightforward to check that $N(\lambda)$ is non-increasing in λ (see the proof of Proposition 1 in the Supplementary File). Therefore, the condition $N(0) \geq 0$ is necessary to ensure that α can be achieved when $\lambda = \lambda^*$. Then, the oracle procedure $\boldsymbol{\delta}_T^{\lambda^*}$ for the total error control problem can be defined formally in the following theorem, along with its validity and optimality for the total mFDR-control.

Theorem 1. *Consider an HMM model as defined in model (1) and the related decision rule $\boldsymbol{\delta}_T^\lambda$ defined in (3). Then, oracle procedure is $\boldsymbol{\delta}_T^{\lambda^*}$ where λ^* is defined in Proposition 1. If the pre-specified misclassification error rate α satisfies the condition $N(0) \geq 0$, then we have*

(i) $\mathrm{mFDR_T}(\boldsymbol{\delta}_\mathrm{T}^{\lambda^*}) = \alpha$; and

(ii) For any decision rule $\boldsymbol{\delta}$ satisfying $\mathrm{mFDR_T}(\boldsymbol{\delta}) \leq \alpha$, we have

$$\mathrm{ETP_T}(\boldsymbol{\delta}) \leq \mathrm{ETP_T}(\boldsymbol{\delta}_\mathrm{T}^{\lambda^*}).$$

The set-specific problem can be discussed similarly as follows. First, the constraints on $\{\mathrm{mFDR}_{\mathrm{S}}^{k}, k=1,...,K\}$ can be written as

$$E\left\{\sum_{j=1}^{m} I(\delta_j = k)[\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha_k]\right\} \leq 0, \qquad k=1,...,K,$$

and the optimal decision rule can be obtained by minimizing the Lagrangian

$$L_S(\boldsymbol{\lambda},\boldsymbol{\delta}) = \sum_{k=1}^{K}\sum_{j=1}^{m} I(\delta_j \neq k)[1-\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2)] + \sum_{k=1}^{K}\sum_{j=1}^{m} \lambda_k I(\delta_j = k)[\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha_k].$$

Given $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_K)$ with $\lambda_k > 0$ for all k, define the decision rule $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}} = (\delta_{S1}^{\boldsymbol{\lambda}},...,\delta_{Sm}^{\boldsymbol{\lambda}})$, where

$$\delta_{Sj}^{\boldsymbol{\lambda}} = \arg\min_{k\in\{0,...,K\}}\left\{\left[\sum_{k'\in\{1,...,K\},k'\neq k}[1-\mathrm{LIS}_j^{k'}(\boldsymbol{X}_1,\boldsymbol{X}_2)]\right] + \lambda_k[\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha_k]\right\}. \quad (4)$$

Then, $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ has two important properties given in the following proposition.

Proposition 2. *For given $\boldsymbol{\lambda}$ with each element $\lambda_k > 0$, the decision rule $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ defined in (4) has the following properties:*

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

(ii) Let $N_k(\boldsymbol{\lambda}) = E\left\{I(\delta_{Sj}^{\boldsymbol{\lambda}} = k)[\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha_k]\right\}$, for $k=1,...,K$, and

$$\check{\lambda}_{k,t} = \inf\{\lambda_k \leq \check{\lambda}_{k,t-1} : N_k(\boldsymbol{\check{\lambda}_{k,t-1}}) \leq 0\}, \qquad k=1,...,K,$$

where $t \geq 1$, $\check{\lambda}_{k,0} = \infty$, and $\boldsymbol{\check{\lambda}_{k,t-1}}$ is the $\boldsymbol{\lambda}$ with $\lambda_{k'} = \check{\lambda}_{k',t-1}, k' \neq k$. If $\alpha_k + \alpha_{k'} \leq 1$ holds for any $k \neq k'$ and $\mathbf{0} \in \{(N_1(\boldsymbol{\lambda}),...,N_k(\boldsymbol{\lambda}))\}$, then $\lambda_k^* = \lim_{t\to\infty}\check{\lambda}_{k,t}$ and $N_k(\boldsymbol{\lambda}^*) = 0$, for $k=1,...,K$, where $\boldsymbol{\lambda}^* = (\lambda_1^*,...,\lambda_K^*)$.

Similar to part (ii) of Proposition 1, part (ii) of Proposition 2 provides the necessary conditions that guarantee the existence of $\boldsymbol{\lambda}^*$ such that $N_k(\boldsymbol{\lambda}^*) = 0$, for $k=1,...,K$. Then, the oracle procedure $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}^*}$ for the set-specific error control problem can be defined formally in the following theorem, along with its validity and optimality for the set-specific mFDR-control.

Theorem 2. *Consider an HMM model as defined in model* (1) *and the related decision rule* $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}$ *defined in (4). Then, the oracle procedure is* $\boldsymbol{\delta}_S^{\boldsymbol{\lambda}^*}$ *where* $\boldsymbol{\lambda}^*$ *is defined in Proposition 2. If* $\{\alpha_k, k = 1, ...K\}$ *satisfy the conditions in Proposition 2(ii), then we have*

(i) $\mathrm{mFDR}_{\mathrm{S}}^k(\boldsymbol{\delta}_S^{\boldsymbol{\lambda}}) = \alpha_k,$ for $k = 1, ..., K$*; and*

(ii) For any decision rule $\boldsymbol{\delta}$ *satisfying* $\mathrm{mFDR}_{\mathrm{S}}^k(\boldsymbol{\delta}) \leq \alpha_k$*, for* $k = 1, ..., K$*, we have*

$$\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta}) \leq \mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta}_{\mathrm{S}}^{\boldsymbol{\lambda}}).$$

Remark 2. *For given* ϑ *in the HMM model* (1)*, the oracle statistic* $\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$ *can be expressed in terms of the forward and backward density variables, which are defined respectively as* $\alpha_j(p, q) = P(\theta_{1j} = p, \theta_{2j} = q, \{X_{1i}\}_{i=1}^j, \{X_{2i}\}_{i=1}^j)$ *and* $\beta_j(p, q) = P(\{X_{1i}\}_{i=j+1}^m, \{X_{2i}\}_{i=j+1}^m | \theta_{1j} = p, \theta_{2j} = q)$*. It can be shown that*

$$\mathrm{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{\sum_{(p,q)\neq(k_1,k_2)} \alpha_j(p, q)\beta_j(p, q)}{\sum_{p=0}^1 \sum_{q=0}^1 \alpha_j(p, q)\beta_j(p, q)}.$$

In addition, the quantities $\alpha_j(p, q)$ *and* $\beta_j(p, q)$ *can be calculated recursively by using the forward-backward procedure (Rabiner, 1989; Wang and Zhu, 2019). More specifically, let* $\alpha_1(p, q) = \pi_{pq} f_{1p}(X_{11}) f_{2q}(X_{21})$ *and* $\beta_m(p, q) = 1$*. Then, we have the following recursive formulas:*

$$\begin{aligned}\alpha_{j+1}(p, q) &= \sum_{s=0}^1 \sum_{t=0}^1 \alpha_j(s, t) a_{(s,t)(p,q)} f_{1p}(X_{1,j+1}) f_{2q}(X_{2,j+1}),\\ \beta_j(p, q) &= \sum_{s=0}^1 \sum_{t=0}^1 \beta_{j+1}(s, t) a_{(p,q)(s,t)} f_{1s}(X_{1,j+1}) f_{2t}(X_{2,j+1}).\end{aligned}$$

3.2 Data-driven procedures

In practice, the HMM parameters ϑ are usually unknown, which makes the oracle procedures described in the previous part unusable. To address this issue, our strategy is to first estimate these unknown parameters by $\hat{\vartheta}$, and then plug-in $\hat{\vartheta}$ to the related oracle procedures to obtain the corresponding data-driven signal classification procedures. To this end, the maximum likelihood estimate (MLE) is commonly used in the literature because of their appealing properties (e.g., statistical consistency, and asymptotic normality) under some regularity conditions (Leroux, 1992; Bickel et al., 1998), and the MLE can be computed by using the EM algorithm or other standard numerical optimization schemes, such as the gradient search. Let $\hat{\vartheta}$ be the MLE of ϑ. After it is plugged into $\text{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$, the corresponding plug-in statistic is denoted as $\widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$, which can be computed by using the forward-backward procedure introduced in Remark 2.

Based on $\widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$, we can construct the data-driven procedure for the total error control problem as follows. First, let $\hat{\delta}_{Tj}^{\lambda}$ be the decision rule for the total error minimization problem (3), after $\text{LIS}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is replaced by $\widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2)$. Then, the quantity $N(\lambda)$ in Proposition 1 can be approximated by

$$\hat{N}(\lambda) = \frac{1}{m} \sum_{j=1}^{m} \sum_{k=1}^{K} I(\hat{\delta}_{Tj}^{\lambda} = k)[\widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha].$$

The above expression can be further simplified by using the result that

$$\mathbb{I}(\hat{\delta}_{Tj}^{\lambda} = k) = \mathbb{I}\left\{\widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) \le \frac{\alpha\lambda + 1}{\lambda + 1}, \quad \widehat{\text{LIS}}_j^k(\boldsymbol{X}_1, \boldsymbol{X}_2) \le \min_{k' \neq k} \widehat{\text{LIS}}_j^{k'}(\boldsymbol{X}_1, \boldsymbol{X}_2)\right\}.$$

Then, the data-driven classification rule for the total classification problem is defined to be

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

where

$$\hat{\lambda}^* = \inf\{\lambda : \hat{N}(\lambda) \le 0\}.$$

An adaptive step-up algorithm for calculating $\hat{\boldsymbol{\delta}}_T^{\hat{\lambda}^*}$ is summarized in Table 3. The following theorem establishes the asymptotic validity and optimality of $\hat{\boldsymbol{\delta}}_T^{\hat{\lambda}^*}$ for the total mFDR-control.

Theorem 3. *For the HMM model* (1), *if all assumptions in Theorem 1 and the assumption that $\hat{\vartheta}$ is a consistent estimate of ϑ (called Assumption 1 hereafter) are valid, then we have*

(i) $\text{mFDR}_\text{T}(\hat{\boldsymbol{\delta}}_T^{\hat{\lambda}^*}) = \alpha + o(1)$; *and*

Table 3: Data-driven algorithm for the total classification problem.

Let $\widehat{\mathrm{LIS}}_j^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \min_k \widehat{\mathrm{LIS}}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2)$, and $\widehat{\mathrm{LIS}}_{(j)}^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2)$ be the order statistics of $\widehat{\mathrm{LIS}}_j^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2)$, for $j = 1,...,m$. The quantities $\hat{\delta}_{T(j)}^{\hat{\lambda}^*}$ and $\widehat{\mathrm{LIS}}_{(j)}^k(\boldsymbol{X}_1,\boldsymbol{X}_2)$ are the corresponding decision rules and test statistics. Define $r = \max\{i : (1/i)\sum_{j=1}^i \widehat{\mathrm{LIS}}_{(j)}^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha \leq 0\}$. Then, we have

$$\hat{\delta}_{T(j)}^{\hat{\lambda}^*} = \begin{cases} k & \text{if } j \leq r \text{ and } \widehat{\mathrm{LIS}}_{(j)}^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \widehat{\mathrm{LIS}}_{(j)}^k(\boldsymbol{X}_1,\boldsymbol{X}_2), \\ 0 & \text{if } j \geq r. \end{cases}$$

(ii) $\mathrm{ETP}_{\mathrm{T}}(\hat{\boldsymbol{\delta}}_T^{\hat{\lambda}^*})/\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta}_T^{\lambda^*}) = 1 + o(1)$.

Remark 3. *From the definition of $I(\hat{\delta}_{Tj}^{\boldsymbol{\lambda}} = k)$ and the step-up procedure described in Table 3, we can observe that the step-up procedure is based on ranking the quantities $\widehat{\mathrm{LIS}}_j^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2) = \min_k \widehat{\mathrm{LIS}}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2)$, and considering the discrete cutoffs only in $\widehat{\mathrm{LIS}}_j^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2)$. Such an approach can effectively select the features that are most likely signals. More specifically, even if $\widehat{\mathrm{LIS}}_j^0(\boldsymbol{X}_1,\boldsymbol{X}_2) > \widehat{\mathrm{LIS}}_{j'}^0(\boldsymbol{X}_1,\boldsymbol{X}_2)$ for some $j \neq j'$, which implies that the jth feature is more likely a signal, the j'th feature may still be selected because of $\widehat{\mathrm{LIS}}_{j'}^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2) < \widehat{\mathrm{LIS}}_j^{\min}(\boldsymbol{X}_1,\boldsymbol{X}_2)$. This can happen especially when $\{\widehat{\mathrm{LIS}}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2), k = 1,2,3\}$ are close to each other, which can also be found in the real-data analysis.*

For the set-specific problem, the data-driven procedure can be developed similarly. More specifically, after replacing $\mathrm{LIS}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2)$ with $\widehat{\mathrm{LIS}}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2)$, let $\hat{\delta}_{Sj}^{\boldsymbol{\lambda}}$ be the solution to the set-specific error minimization problem (4). Define

$$\hat{N}_k(\boldsymbol{\lambda}) = \frac{1}{m}\sum_{j=1}^m I(\hat{\delta}_{Sj}^{\boldsymbol{\lambda}} = k)[\widehat{\mathrm{LIS}}_j^k(\boldsymbol{X}_1,\boldsymbol{X}_2) - \alpha_k].$$

Then, for each k, consider a sequence $\{\hat{\lambda}_{k,t}, t \geq 1\}$ satisfying

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

where $\hat{\lambda}_{k,0} = \infty$, and $\hat{\boldsymbol{\lambda}}_{\boldsymbol{k,t-1}}$ is the $\boldsymbol{\lambda}$ with $\lambda_{k'} = \hat{\lambda}_{k',t-1}, k' \neq k$. The convergence of the sequence $\{\hat{\lambda}_{k,t}, t \geq 1\}$ can be proved similarly to the one in Proposition 2. Let $\hat{\lambda}_k^* = \lim_{t\to\infty}\hat{\lambda}_{k,t}$

Table 4: Data-driven algorithm for the set-specific classification problem.

Step 1: Let $\widehat{\mathrm{LIS}}^k_{(j)}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ be the order statistics of $\widehat{\mathrm{LIS}}^k_j(\boldsymbol{X}_1, \boldsymbol{X}_2)$, for each $k = 1, ..., K$. Set $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ to be $\lambda_k = (1 - \alpha_k)/(\widehat{\mathrm{LIS}}^k_{(r_k)}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha_k) - 1$, and $r_k = \max\{i : (1/i)\sum_{j=1}^{i} \widehat{\mathrm{LIS}}^k_{(j)}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha_k \leq 0\}$, for $k = 1, ..., K$.

Step 2: For each k, calculate $\hat{N}_k(\boldsymbol{\lambda})$ and $\hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,r_k+1})$, where $\tilde{\boldsymbol{\lambda}}_{k,j} = (\tilde{\lambda}_{1,j}, ..., \tilde{\lambda}_{K,j})$, $\tilde{\lambda}_{k',j} = \lambda_{k'}$ when $k' \neq k$, and $\tilde{\lambda}_{k,j} = (1 - \alpha_k)/(\widehat{\mathrm{LIS}}^k_{(j)}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \alpha_k) - 1$. If $\hat{N}_k(\boldsymbol{\lambda}) \leq 0$ and $\hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,r_k+1}) > 0$, for all k, then set $\hat{\boldsymbol{\lambda}}^* = \boldsymbol{\lambda}$; otherwise, go to Step 3.

Step 3: Reset $r_k = \tilde{r}_k$, where $\tilde{r}_k = \max\{i \geq r_k : \hat{N}_k(\tilde{\boldsymbol{\lambda}}_{k,i}) \leq 0\}$, for all k. Update $\boldsymbol{\lambda}$ in Step 1 and repeat Steps 2 and 3 until the desired threshold vector $\hat{\boldsymbol{\lambda}}^*$ is obtained.

Step 4: Apply $\hat{\boldsymbol{\lambda}}^*$ to Equation (4) to obtain the classification rule $\hat{\boldsymbol{\delta}}_S^{\hat{\boldsymbol{\lambda}}^*} = (\hat{\delta}_{S1}^{\hat{\boldsymbol{\lambda}}^*}, ..., \hat{\delta}_{Sm}^{\hat{\boldsymbol{\lambda}}^*})$, where $\{\mathrm{LIS}^k_j, k = 1, ..., K\}$ should be replaced by $\{\widehat{\mathrm{LIS}}^k_j, k = 1, ..., K\}$.

and $\hat{\boldsymbol{\lambda}}^* = (\hat{\lambda}_1^*, ..., \hat{\lambda}_K^*)$. Then, the data-driven classification rule is defined to be

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

A simple and fast algorithm for calculating $\hat{\boldsymbol{\delta}}_S^{\hat{\boldsymbol{\lambda}}^*}$ is summarized in Table 4. The following theorem establishes its asymptotic validity and optimality for the set-specific mFDR-control.

Theorem 4. *For the HMM model* (1)*, if all assumptions in Theorems 2 and 3 are valid, then we have the following results: for each* $k = 1, ..., K$*,*

(i) $\mathrm{mFDR}^k_{\mathrm{S}}(\hat{\boldsymbol{\delta}}_S^{\hat{\boldsymbol{\lambda}}^*}) = \alpha_k + o(1)$*; and*

(ii) $\mathrm{ETP}_{\mathrm{T}}(\hat{\boldsymbol{\delta}}_S^{\hat{\boldsymbol{\lambda}}^*})/\mathrm{ETP}_{\mathrm{T}}(\boldsymbol{\delta}_S^{\boldsymbol{\lambda}^*}) = 1 + o(1)$.

4 Simulation Studies

In this section, we investigate the numerical performance of the proposed oracle and data-driven procedures and compare them with some alternative methods. In the simulation study, the number of features to classify is fixed at $m = 10000$. The four-state Markov chain

$\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$ is generated using the initial state distribution $\boldsymbol{\pi} = (1, 0, 0, 0)$ and the transition matrix

$$\mathcal{A} = \begin{pmatrix} 0.94 & 0.02 & 0.02 & 0.02 \\ h/3 & 1-h & h/3 & h/3 \\ h/3 & h/3 & 1-h & h/3 \\ h/3 & h/3 & h/3 & 1-h \end{pmatrix}.$$

Given $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$, the observations $\{(X_{1j}, X_{2j}), j = 1, ..., m\}$ are generated as follows: $X_{ij} \sim N(0, 1)$ when $\theta_{ij} = 0$, and $X_{ij} \sim N(\mu_i, 1)$ when $\theta_{ij} = 1$, for $i = 1, 2, j = 1, ..., m$. As discussed in Remark 1, we focus on classifying the m features into either $S_0 = \{0\}$, $S_1 = \{1\}$, $S_2 = \{2\}$, or $S_3 = \{3\}$. All simulation results presented in this section are based on 200 replicated simulations. The EM algorithm for estimating the HMM parameters in a normal mixture model is provided in the Supplementary file.

The following signal classification procedures are considered:

- the proposed oracle and data-driven procedures for the total and set-specific error control problems, denoted as TO, TD, SO and SD, respectively, where "T" and "S" in the first letter of a notation denote "total" and "set-specific", and "O" and "D" in the second letter of a notation denote "oracle" and "data-driven";

- the oracle and data-driven procedures proposed by Xiang, Zhao and Cai (2019) for the total and set-specific error control problems, in which it is assumed that $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$ are mutually independent over different j values. These procedures are denoted as ITO, ITD, ISO and ISD, where the first letter "I" in the notations denotes "independent", and the second and third letters have the same meanings as those in the notations discussed in the previous item;

- a classification approach to analyze each sequence separately based on the related p-values, denoted as SEP. Here, the procedure discussed in Genovese and Wasserman (2004) is used for each sequence to control the marginal FDR. For the total error control problem, the error levels in both sequences are set to be the same as α. For the set-specific error control problem, the error levels in both sequences are set to be the same as the average of α_k's.

Several simulation settings are considered here. The first setting focuses on total marginal error control. In this setting, the signal strengths μ_i's can vary from 2.5 to 3.4, h can vary from 0.2 to 0.4, and the nominal total mFDR α can vary from 0.05 to 0.2. The results are presented in Figure 1. From the plots of the figure, it can be seen that the proposed oracle and data-driven procedures TO and TD for the total error control problem can both control the total mFDR at the nominal level in all cases considered, and the ETP values of TO is almost the same as those of TD, indicating that the performance of TO could be attained asymptotically by TD. In addition, the ETP values of TO and TD are significantly larger than those of the other three methods, implying their superiority for the total error control problem. As a comparison, although the oracle procedure of Xiang, Zhao and Cai (2019) (i.e., ITO) can control the total mFDR at α, its data-driven version ITD cannot control the mFDR well in the sense that its total mFDR values are a bit larger than the nominal level. The main reason for this phenomenon is that the estimation accuracy involved in this method would not be good when the observed data are correlated. As for the procedure SEP, its total mFDR values are significantly larger than the nominal level, implying its invalidity in this case. From Figure 1, we can have the following additional conclusions. First, we can see from Figures 1(b) and 1(f) that the proposed procedures TO and TD are capable of classifying most true signals correctly even when the signals are weak and the nominal mFDR level is small, which represents a great breakthrough in developing effective signal classification methods. Second, we can see from Figure 1(d) that all ETP curves decrease as h increases, which is reasonable because the number of true signals also decreases in such cases.

The second setting focuses on set-specific error control. Similar to the first simulation setting, we let the signal strengths μ_i's vary from 2.5 to 3.4, h vary from 0.2 to 0.4, and the nominal set-specific mFDR values $\{\alpha_k, k = 1, 2, 3\}$ be the same and their values vary from 0.05 to 0.2. The results are presented in Figure 2. From the plots in the figure, we see that the proposed procedures SO and SD can both control the set-specific mFDR levels well, and have significantly larger ETP values in all cases considered, compared with their competitors.

In many applications, the signal strengths, state transition probabilities, and many other settings may be more complicated than the symmetric settings considered in the above simulation examples. So, we consider the following simulation example to investigate the robustness of the proposed procedures. More specifically, the four-state Markov chain $\{(\theta_{1j}, \theta_{2j}), j = 1, ...m\}$ is first generated using the initial state distribution $\boldsymbol{\pi} = (1, 0, 0, 0)$ and the transition matrix

$$\mathcal{A} = \begin{pmatrix} 0.94 & 0.02 & 0.02 & 0.02 \\ (h+0.1)/3 & 0.9-h & (h+0.1)/3 & (h+0.1)/3 \\ h/3 & h/3 & 1-h & h/3 \\ (h-0.1)/3 & (h-0.1)/3 & (h-0.1)/3 & 1.1-h \end{pmatrix}.$$

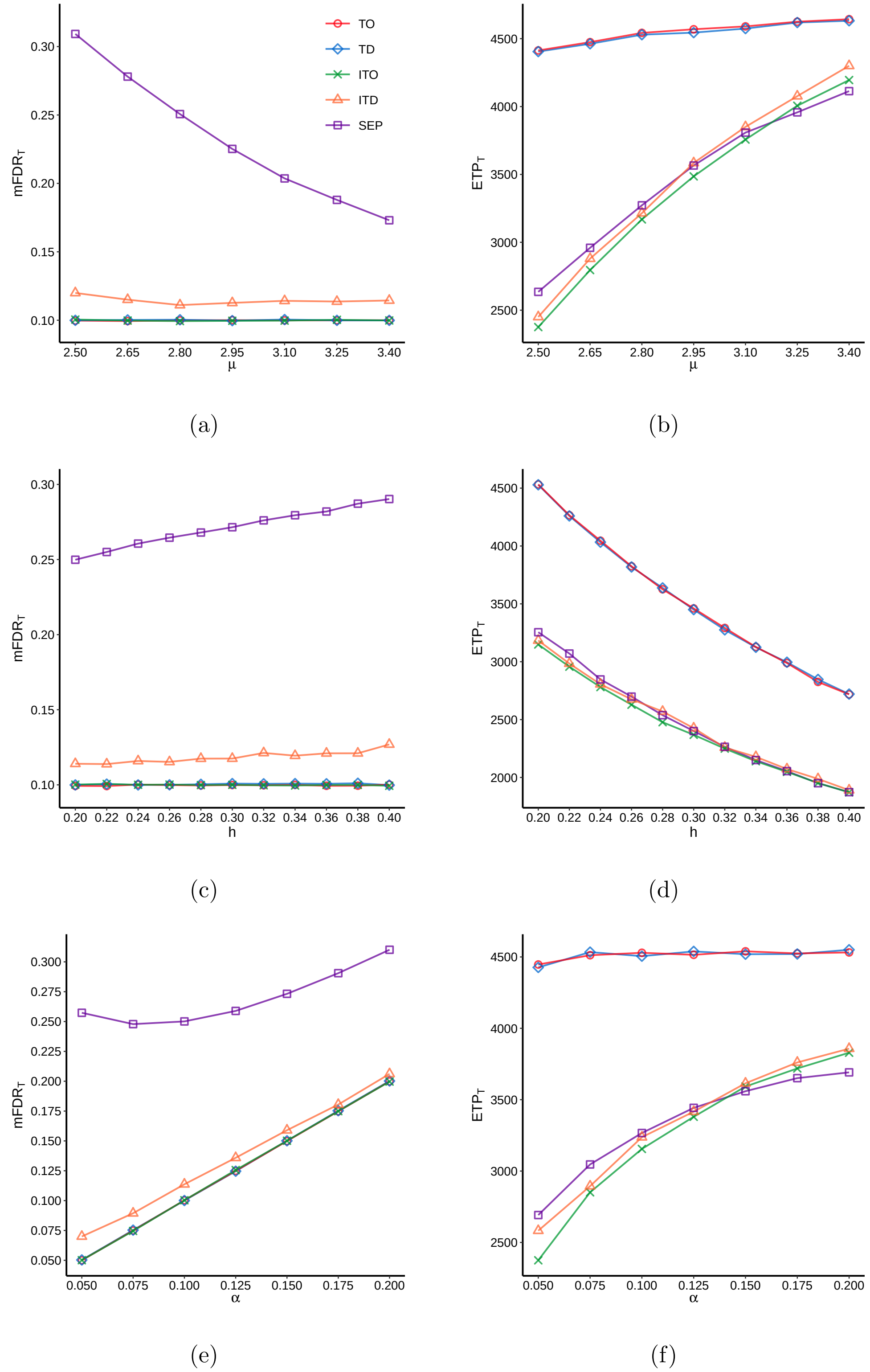
Given $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$, the observations $\{\boldsymbol{X}_{ij}, i = 1, 2, j = 1, ..., m\}$ are generated from the distributions $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im})^T$, $\mu_{1j}+0.5 = \mu$ when $\theta_{1j} = 1$, $\mu_{2j} - 1 = \mu$ when $\theta_{2j} = 1$, $\mu_{1j} = 0$ when $\theta_{1j} = 0$, $\mu_{2j} = 0$ when $\theta_{2j} = 0$, and $\boldsymbol{\Sigma} = (0.5^{|j_1-j_2|})_{j_1,j_2=1}^m$. For the total error control, the nominal total mFDR level is fixed at α. For the set-specific error control, the nominal set-specific mFDR levels are set to be $\alpha_1 = 2\alpha_2 = \alpha_3/2 = \alpha$. The results are presented in Figure 3. It can be seen that the efficacy and robustness of the proposed methods are reasonably good in the sense that both the oracle and data-driven procedures perform satisfactorily.

Finally, we investigate the performance of the proposed procedures in the presence of model misspecification, specifically when the HMM assumption is invalid. To this end, we focus on the total error control problem, and the set-specific error control problem exhibits similar behavior based on our numerical results. More specifically, let us consider a case when $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$ are assumed to be independent of each other, instead of following an HMM. Given $\{(\theta_{1j}, \theta_{2j}), j = 1, ..., m\}$, the observations $\boldsymbol{X}_{ij}, i = 1, 2, j = 1, ..., m$ are generated from the multivariate normal distribution $N_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im})^T$ and $\mu_{ij} = \mu\mathbb{I}(\theta_{ij} = 1)$. The signal proportions are set to be $(\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}) = (1 - h, h/3, h/3, h/3)$, where h varies between 0.15 and 0.36. The signal strength μ varies between 2.8 and 3.4, and the nominal total mFDR α varies between 0.05 and 0.2. Figure 4 presents the results of applying the TD procedure with two different covariance structures: $\boldsymbol{\Sigma} = \boldsymbol{I}_{m\times m}$ and $\boldsymbol{\Sigma} = (0.5^{|j_1-j_2|})_{j_1,j_2=1}^{m}$. As a benchmark, the results of the ITO procedure are also presented, as ITO is theoretically optimal under the independence assumption and asymptotically optimal under short-range dependence (Xiang et al., 2019). From the plots of the figure, we can make the following observations. (i) When there is no correlation among the features, the performance of TD is very close to that of ITO, which demonstrates the strong robustness and effectiveness of TD even in the absence of correlation. (ii) When short-range dependence exists, TD still maintains good performance as long as the signal strength μ is not too small. Based on these observations, we can conclude that the proposed procedures are quite robust to model misspecification.

5 Application

In this section, we apply the proposed methods to study the genetic architectures of schizophrenia (SCZ) and bipolar disorder (BD). It is well known that these two diseases share similar genetic architectures and are affected simultaneously by some common genetic variants. Thus, for more effective disease diagnosis, it is important to study genetic associations between SCZ and BD. For demonstration purposes, we use the single-nucleotide polymorphisms (SNPs) dataset of SCZ and BD that was collected by two GWAS studies performed by Ruderfer et al. (2014). Summary z-scores of the two studies are publicly available at the webpage of the Psychiatric Genomics Consortium (https://www.med.unc.edu/pgc/). After pruning the SNPs at a linkage disequilibrium r^2 threshold of 0.5 obtained from the data of the '1,000 genomes project' (1,000 Genomes Project Consortium, 2015), 439,040 variants remain for further analysis.

It is important to point out that human genome may not be entirely linear, as SNPs are present on physically separate chromosomes. Previous studies in the literature have provided strong evidence that when employing an HMM to capture SNP dependency, it is necessary to treat each chromosome as an independent sequence. Furthermore, combining the testing results from different chromosomes has been shown to yield greater efficiency (Wei et al.,



(a)

(b)

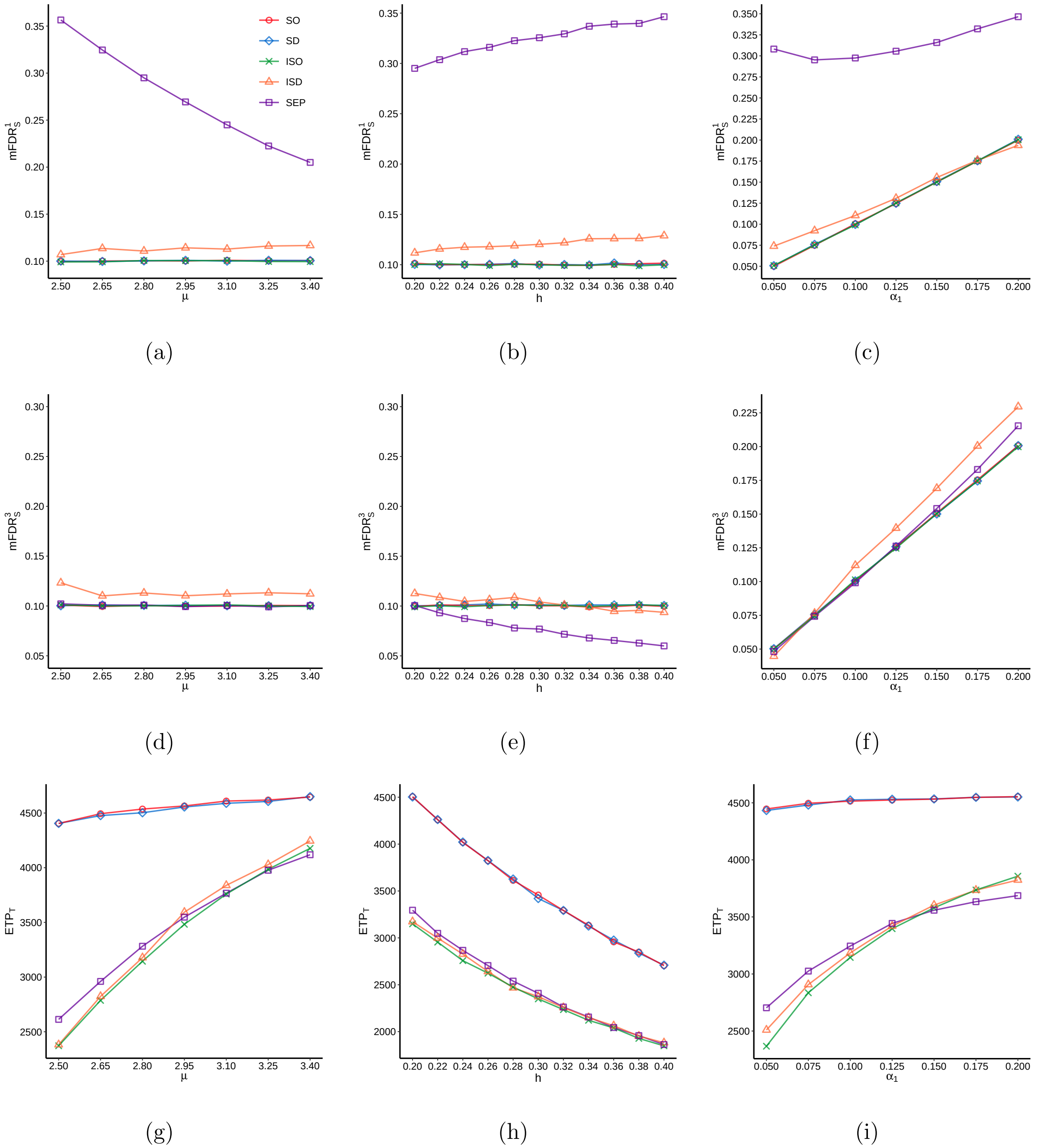
(c)

(d)

(e)

(f)

Figure 1: Simulation results for the total mFDR control. Plots (a) and (b): $h = 0.2$ and $\alpha = 0.1$; plots (c) and (d): $\mu = 2.8$ and $\alpha = 0.1$; plots (e) and (f): $\mu = 2.8$ and $h = 0.2$.

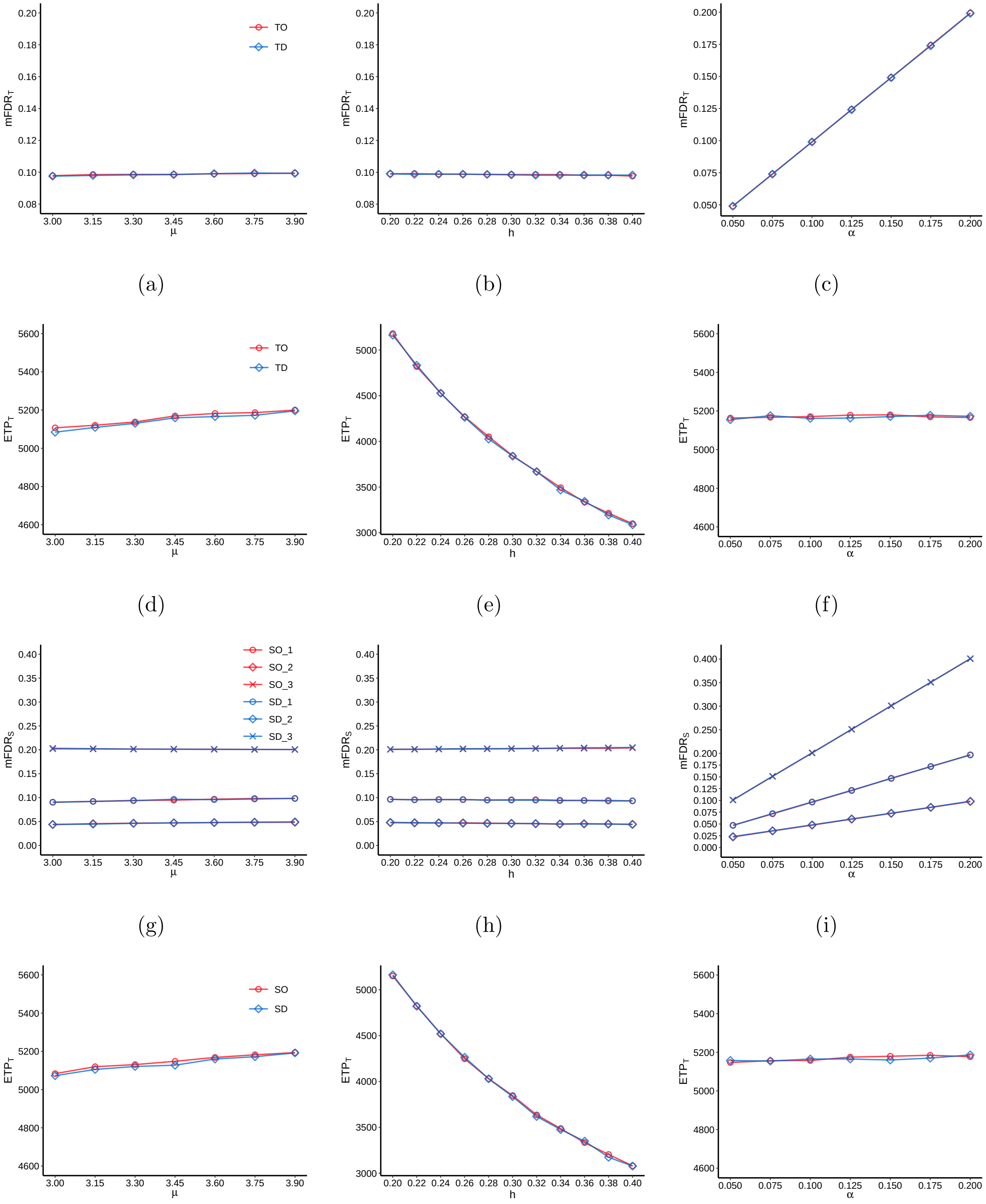


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2: Simulation results for the set-specific mFDR control. Plots (a), (d) and (g): $h = 0.2$ and $\alpha = 0.1$; plots (b), (e) and (h): $\mu = 2.8$ and $\alpha = 0.1$; plots (c), (f) and (i): $\mu = 2.8$ and $h = 0.2$.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

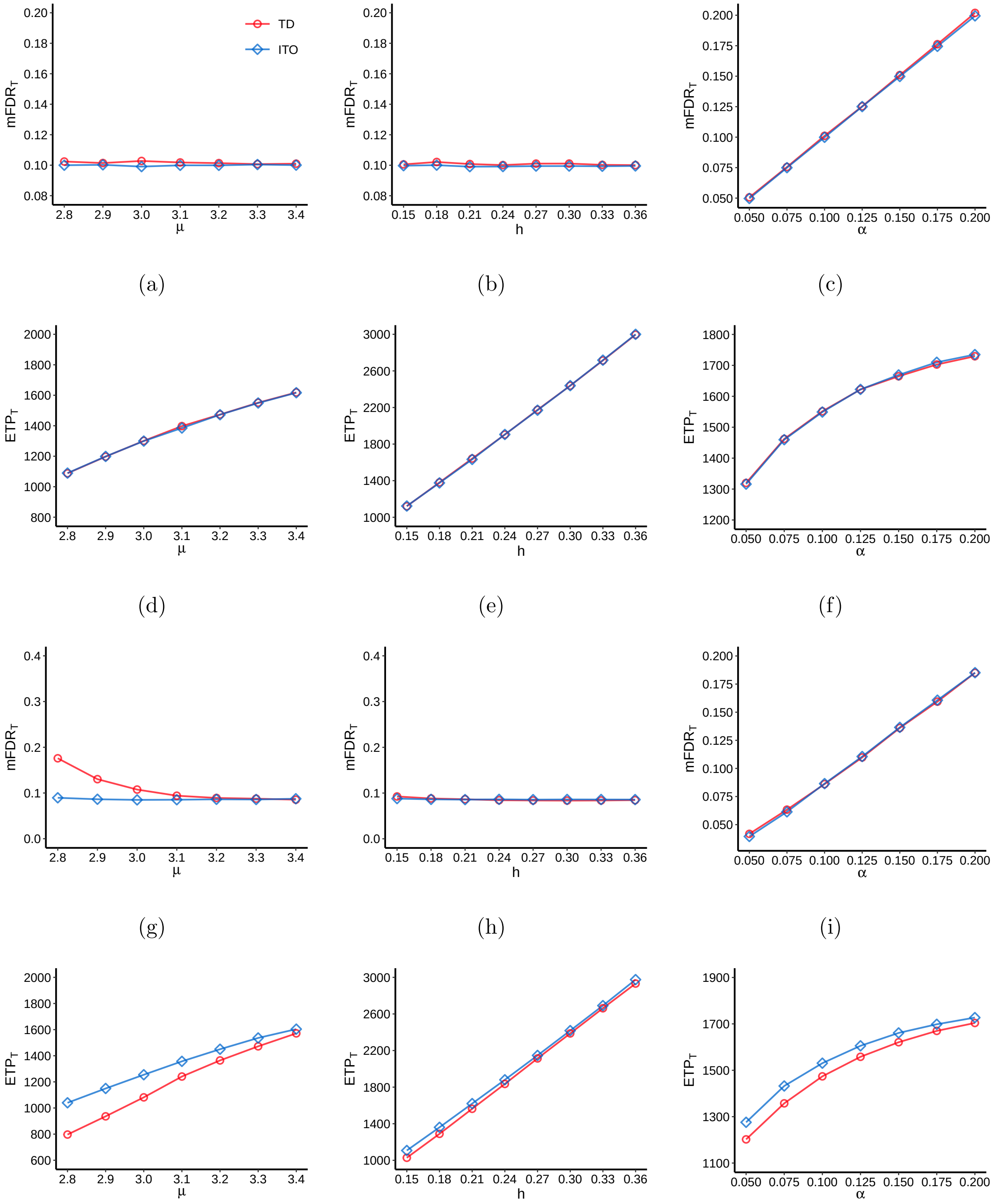
(i)

(j)

(k)

(l)

Figure 3: Robustness of the proposed procedures. Plots (a), (d), (g) and (j): $h = 0.2$ and $\alpha = 0.1$; plots (b), (e), (h) and (k): $\mu = 3.6$ and $\alpha = 0.1$; plots (c), (f), (i) and (l): $\mu = 3.6$ and $h = 0.2$.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 4: Performance of TD in the presence of model misspecification. Results in plots (a)-(f) are under $\boldsymbol{\Sigma} = \boldsymbol{I}_{m\times m}$, and those in plots (g)-(l) are under $\boldsymbol{\Sigma} = (0.5^{|j_1-j_2|})_{j_1,j_2=1}^{m}$. Plots (a), (d), (g) and (j): $h = 0.2$ and $\alpha = 0.1$; plots (b), (e), (h) and (k): $\mu = 3.3$ and $\alpha = 0.1$; plots (c), (f), (i) and (l): $\mu = 3.3$ and $h = 0.2$.

2009; Wang and Zhu, 2019). Therefore, in this study, we use the proposed procedures to calculate the LIS statistics for each chromosome separately, and then the statistics calculated from all the relevant chromosomes are combined to obtain the final results.

We first apply the proposed data-driven procedure TD for signal classification under the total mFDR control problem to classify the SNPs into the sets G_0 (SNPs insignificant in either disease), G_1 (SNPs significant in BD only), G_2 (SNPs significant in SCZ only), and G_3 (SNPs significant in both diseases). The first row of Table 5 shows the numbers of SNPs classified into the sets G_1-G_3 by using TD with $\alpha = 0.05$. It can be seen that majority significant SNPs are categorized into the class G_3, which is consistent with findings in the existing literature indicating a close genetic etiology relationship between SCZ and BD (Huang et al., 2010). Notably, a large portion of the identified SNPs in G_3 are concentrated within certain specific genes, namely, IFI44L (on chromosome 1), ITIH4 (on chromosome 3), ZNF184 (on chromosome 6), CACNB2 and SH3PXD2A (on chromosome 10), CACNA1C (on chromosome 12), and PCNT (on chromosome 21), which are displayed in Figure 5. From Figure 5, it can be seen that some features with high $\widehat{\mathrm{LIS}}^0$ values are considered as non-significant, while others with lower $\widehat{\mathrm{LIS}}^0$ values are identified as signals, which supports Remark 3.

It worths mentioning that the signal classification results of TD are consistent with several previously reported findings. For instance, Ruderfer et al. (2014) identified genome-wide significant SNPs in ITIH4 and CACNA1C that are associated with both BD and SCZ. Ren et al. (2021) demonstrated that the gene ZNF184 is associated with a common factor shared by BD, SCZ, and the major depressive disorder (MDD) at the genomic significance level. Numata et al. (2009) illustrated the interaction between PCNT and disruption-in-schizophrenia 1 (DISC1), a known genetic risk factor for both BD and SCZ. These findings corroborate the signal classification results obtained from the TD procedure.

In some situations, however, SNPs in the sets G_1 and G_2 are more helpful than those in G_3 for differentiating patients with SCZ from those with BD. In such cases, despite all three non-null classes remain of interest, finding SNPs in G_1 and G_2 is more important than finding SNPs in G_3. The proposed data-driven procedure SD for the set-specific mFDR

Table 5: Numbers of SNPs classified by TD/SD and ITD/ISD into different sets of interest.

Method	Nominal mFDR	G_1	G_2	G_3
TD	$\alpha = 0.05$	4	2	155
SD	$\alpha_1 = 0.1, \alpha_2 = 0.1, \alpha_3 = 0.01$	36	18	23
ITD	$\alpha = 0.05$	1	2	54
ISD	$\alpha_1 = 0.1, \alpha_2 = 0.1, \alpha_3 = 0.01$	2	4	8

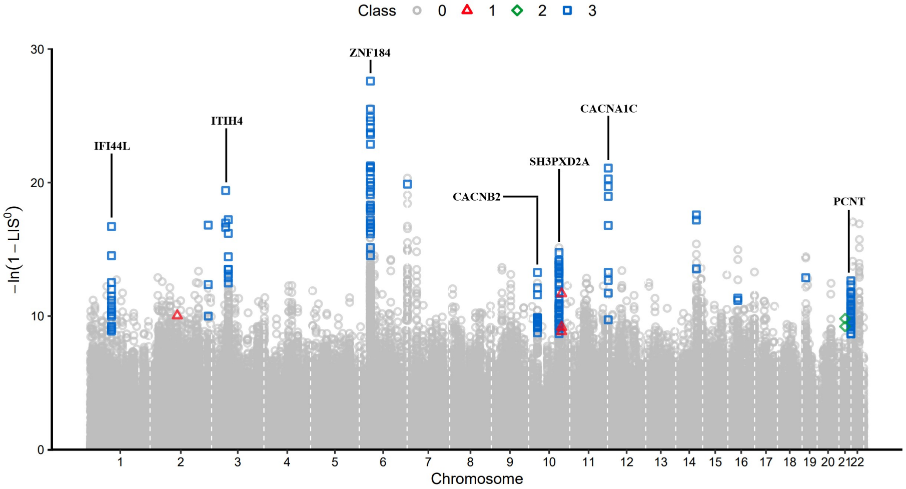


Figure 5: Signal classification results by TD. The y-axis is $-\ln(1-\widehat{\mathrm{LIS}}_j^0)$ with a larger value indicating a larger $\widehat{\mathrm{LIS}}_j^0$.

control problem can be used to solve this problem by assigning different nominal mFDR levels to different classes. To illustrate, the second row of Table 5 presents the numbers of SNPs classified into the sets G_1-G_3 by SD with $\alpha_1 = 0.1, \alpha_2 = 0.1$ and $\alpha_3 = 0.01$. Note that the thresholds for the sets G_1 and G_2 are set to be more liberal than that for the set G_3, to allow more discovered SNPs in the two former sets. It can be seen that compared to TD, more disease-specific SNPs are indeed detected. In addition, most of the 36 SNPs that are associated with BD are clustered within the genes KCNH7 (on chromosome 2), HABP2 (on chromosome 10), and NFIX (on chromosome 19). These findings align with two previous genome-wide association studies. More specifically, Kuo et al. (2014) reported the association between KCNH7 and the risk of developing bipolar I disorder, while Ikeda et al. (2018) identified NFIX as a susceptibility locus for BD. HABP2, also known as FSAP, is involved in blood hemostasis and endothelial function and has been found to be correlated with neurological deficits (Tian et al., 2022). This finding may provide a potential explanation for BD. On the other hand, most of the 18 SNPs that are discovered specific to SCZ are clustered in the genes CSMD1 (on chromosome 8) and NTRK3 (on chromosome 15). The former has association with variation in brain structures or risk of neuropsychiatric disorder and plays a significant role in the etiology of SCZ (Håvik et al., 2011), and the latter was identified to have influence on human hippocampal function, implying a possible role in SCZ pathology (Otnæss et al., 2009).

For comparison purposes, the dataset is also analyzed using ITD and ISD, assuming independence among SNPs. The corresponding results are presented in the last two rows of Table 5. Although it is challenging to validate the true mFDR level in real data analysis, we can draw insights from the simulation results displayed in Figure 4. These simulations imply that the proposed procedures remain valid even in the presence of model misspecification. Here, we define a procedure as valid if it controls the mFDR at the nominal level. From Table 5, it can be observed that TD and SD outperform ITD and ISD in terms of detecting significant SNPs, while maintaining the same nominal mFDR levels. This indicates that the proposed HMM-based TD and SD procedures are more effective in identifying signals than ITD and ISD that assume independence among SNPs.

6 Discussion

This paper focuses on the large-scale signal classification problem under a special form of dependence (i.e., a four-state Cartesian HMM) for paired hypotheses. Under the total and set-specific mFDR control scenarios, we have developed powerful oracle and data-driven procedures under a generalized compound decision theoretic framework based on a generalized local index of significance. It has been shown both theoretically and numerically that the proposed procedures are asymptotically valid and optimal for solving the two-sequence signal classification problem.

Although the HMM model is helpful for describing the data dependence structure in the two-sequence signal classification problem, it should be interesting to extend the proposed procedures to cases with more general data dependence structures. In such cases, a major challenge is that the asymptotic optimality of the data-driven procedures may not be guaranteed since consistent estimates of the unknown model parameters may not be easy to obtain. To the best of our knowledge, consistency of parameter estimates under other general data dependence structures is still unknown in the literature. Thus, the optimality of TO and SO may not be available in such cases, which deserves further research. In addition, the proposed methods can be extended naturally to handle cases with more than two sequences of test statistics. For example, if there are three studies, there would be eight signal classes. This problem can be solved by extending model (1) to have three studies instead of two. Then, the proposed methods can be modified accordingly. However, as the number of studies increases, the implementation of the related signal classification procedures could be difficult. It is therefore important to develop powerful and efficient signal classification methods for cases with a large number of studies, which will be considered in our future research as well.

Supplementary Materials

Supplementary.pdf: The supplementary file contains the proofs of the theoretical results presented in this paper.

CodeAndData.zip: Some computer codes for implementing the proposed methods and the real data used in Section 5.

Acknowledgments

The authors want to thank the Editor, the Associate Editor, and anonymous referees for their constructive comments and suggestions that improved the quality of the paper significantly.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the National Key R&D Program of China [2022YFA1003801; 2021YFA1000101; 2021YFA1000102], National Natural Science Foundation of China [12201382; 12071144; 71931004], Basic Research Project of Shanghai Science and Technology Commission (22JC1400800), and an NSF grant [DMS-1914639].

References

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523), 1172–1183.

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

Bercovici, S., Meek, C., Wexler, Y., and Geiger, D. (2010). Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics*, 26(12), i175–i182.

Bickel, P., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4), 1614–1635.

Cai, T. T., Sun, W., & Xia, Y. (2021). LAWS: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 1–14.

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

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., ... and Ragoussis, J. (2007). QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research*, 35(6), 2013–2025.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477), 93–103.

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

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

Håvik, B., Le Hellard, S., Rietschel, M., Lybæk, H., Djurovic, S., Mattheisen, M., ... and Steen, V. M. (2011). The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. *Biological psychiatry*, 70(1), 35–42.

Heller, R., and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8(1), 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. *American Journal of Psychiatry*, 167(10), 1254–1263.

Ikeda, M., Takahashi, A., Kamatani, Y., Okahisa, Y., Kunugi, H., Mori, N., ... and Iwata, N. (2018). A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Molecular psychiatry*, 23(3), 639–647.

Kuo, P. H., Chuang, L. C., Liu, J. R., Liu, C. M., Huang, M. C., Lin, S. K., ... and Lu, R. B. (2014). Identification of novel loci for bipolar I disorder in a multi-stage genome-wide association study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 51, 58–64.

Leroux, B. (1992) Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1), 127–143.

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 bioinformatics*, 19(1), 1–11.

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(3), 391–406.

Li, W., Xiang, D., Tsung, F., and Pu, X. (2020). A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics*, 62(1), 84–100.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... and Moore, H. F. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6), 580–585.

Numata, S., Iga, J. I., Nakataki, M., Tayoshi, S. Y., Tanahashi, T., Itakura, M., ... and Ohmori, T. (2009). Positive association of the pericentrin (PCNT) gene with major depressive disorder in the Japanese population. *Journal of Psychiatry and Neuroscience*, 34(3), 195–198.

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

Otnæss, M. K., Djurovic, S., Rimol, L. M., Kulle, B., Kähler, A. K., Jönsson, E. G., ... and Andreassen, O. A. (2009). Evidence for a possible association of neurotrophin receptor (NTRK-3) gene polymorphisms with hippocampal function and schizophrenia. *Neurobiology of disease*, 34(3), 518-524.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.

Ren, H., Meng, Y., Zhang, Y., Wang, Q., Deng, W., Ma, X., ... and Li, T. (2021). Spatial expression pattern of ZNF391 gene in the brains of patients with schizophrenia, bipolar disorders or major depressive disorder identifies new cross-disorder biotypes: A trans-diagnostic, top-down approach. *Schizophrenia Bulletin*, 47(5), 1351–1363.

Ruderfer, D. M., Fanous, A. H., Ripke, S., McQuillin, A., Amdur, R. L., Gejman, P. V., ... and Kendler, K. S. (2014). Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular psychiatry*, 19(9), 1017–1024.

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

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

Tian, D. S., Qin, C., Zhou, L. Q., Yang, S., Chen, M., Xiao, J., ... and Wang, W. (2022). FSAP aggravated endothelial dysfunction and neurological deficits in acute ischemic stroke due to large vessel occlusion. *Signal Transduction and Targeted Therapy*, 7(1), 6.

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. *The American Journal of Human Genetics*, 95(5), 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. *Nature genetics*, 51(1), 187–195.

Wang, J., Gui, L., Su, W. J., Sabatti, C., and Owen, A. B. (2016). Detecting multiple replicating signals using adaptive filtering procedures. *arXiv preprint arXiv:1610.03330.*

Wang, P., and Zhu, W. (2019). Replicability analysis in genome-wide association studies via Cartesian hidden Markov models. *BMC bioinformatics*, 20(1), 1–12.

Wei, Z., Sun, W., Wang, K., and Hakonarson, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics*, 25(21), 2802–2808.

Xiang, D., Qiu, P., Wang, D., and Li, W. (2022). Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams. *Technometrics*, 64(3), 323–334.

Xiang, D., Zhao, S. D., and Cai, T. T. (2019). Signal classification for the integrative analysis of multiple sequences of multiple tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81(4), 707–734.

Zhao, S. D., and Nguyen, Y. T. (2020). Nonparametric false discovery rate control for identifying simultaneous signals. *Electronic Journal of Statistics*, 14(1), 110–142.