Submitted to the Annals of Statistics

OPTIMAL FALSE DISCOVERY RATE CONTROL FOR LARGE SCALE MULTIPLE TESTING WITH AUXILIARY INFORMATION

By Hongyuan Cao[†] Jun Chen[‡] and Xianyang Zhang[§]

Florida State University[†], Mayo Clinic[‡] and Texas A&M University[§]

Large-scale multiple testing is a fundamental problem in high dimensional statistical inference. It is increasingly common that various types of auxiliary information, reflecting the structural relationship among the hypotheses, are available. Exploiting such auxiliary information can boost statistical power. To this end, we propose a framework based on a two-group mixture model with varying probabilities of being null for different hypotheses *a priori*, where a shape-constrained relationship is imposed between the auxiliary information and the prior probabilities of being null. An optimal rejection rule is designed to maximize the expected number of true positives when average false discovery rate is controlled. Focusing on the ordered structure, we develop a robust EM algorithm to estimate the prior probabilities of being null and the distribution of p-values under the alternative hypothesis simultaneously. We show that the proposed method has better power than state-of-the-art competitors while controlling the false discovery rate, both empirically and theoretically. Extensive simulations demonstrate the advantage of the proposed method. Datasets from genome-wide association studies are used to illustrate the new methodology.

1. Introduction. Large scale multiple testing refers to simultaneously testing of many hypotheses. Given a pre-specified significance level, family-wise error rate (FWER) controls the probability of making one or more false rejections, which can be unduly conservative in many applications. The false discovery rate (FDR) controls the expected value of the false discovery proportion, which is defined as the ratio of the number of false rejections divided by the number of total rejections. Benjamini and Hochberg (BH) [5] proposed a FDR control procedure that sets adaptive thresholds for the p-values. It turns out that the actual FDR level of the BH procedure is the multiplication of the proportion of null hypotheses and the pre-specified

*The authors are alphabetically ordered. Address correspondence to Xianyang Zhang (zhangxiany@stat.tamu.edu).

MSC 2010 subject classifications: Primary 62G07, 62G10; secondary 62C12

Keywords and phrases: EM algorithm, False discovery rate, Isotonic regression, Local false discovery rate, Multiple testing, Pool-Adjacent-Violators algorithm.

significance level. Therefore, the BH procedure can be overly conservative when the proportion of null hypotheses is far from one. To address this issue, [43] proposed a two-stage procedure (ST), which first estimates the proportion of null hypotheses and uses the estimated proportion to adjust the threshold in the BH procedure at the second stage. From an empirical Bayes perspective, [17] proposed the notion of local FDR (Lfdr) based on the two-group mixture model. [45] developed a step-up procedure based on Lfdr and demonstrated its optimality from the compound decision viewpoint.

The aforementioned methods are based on the premise that the hypotheses are exchangeable. However, in many scientific applications, particularly in genomics, auxiliary information regarding the pattern of signals is available. For instance, in differential expression analysis of RNA-seq data, which tests for difference in the mean expression of the genes between conditions, the sum of read counts per gene across all samples could be the auxiliary data since it is informative of the statistical power [35]. In differential abundance analysis of microbiome sequencing data, which tests for difference in the mean abundance of the detected bacterial species between conditions, the genetic divergence among species is important auxiliary information, since closely-related species usually have similar physical characteristics and tend to covary with the condition of interest [50]. In genome-wide association studies, the major objective is to test for association between the genetic variants and a phenotype of interest. The minor allele frequency and the pathogenicity score of the genetic variants, which are informative of the statistical power and the prior null probability, respectively, are potential auxiliary data, which could be leveraged to improve the statistical power as well as enhance interpretability of the results.

Accommodating auxiliary information in multiple testing has recently been a very active research area. Many methods have been developed adapting to different types of structure among the hypotheses. The basic idea is to relax the p-value thresholds for hypotheses that are more likely to be alternative and tighten the thresholds for the other hypotheses so that the overall FDR level can be controlled. For example, [19] proposed to weight the p-values with different weights, and then apply the BH procedure to the weighted p-values. [23] developed a group BH procedure by estimating the proportion of null hypotheses for each group separately. [34] generalized this idea by using the censored p-values (i.e., the p-values that are greater than a pre-specified threshold) to adaptively estimate the weights that can be designed to reflect any structure believed to be present. [25; 26] proposed the independent hypothesis weighting (IHW) for multiple testing with covariate information. The idea is to use cross-weighting to achieve finite-sample

FDR control. Note that the binning in IHW is only to operationalize the procedure and it can be replaced by the proposed EM algorithm below.

The above procedures can be viewed to some extent as different variants of the weighted-BH procedure. Another closely related method was proposed in [30], which iteratively estimates the p-value threshold using partially masked p-values. It can be viewed as a type of Knockoff procedure [2] that uses the symmetry of the null distribution to estimate the false discovery proportion. A similar idea was explored in [51] which proposed a covariate adaptive multiple testing procedure.

Along a separate line, Lfdr-based approaches have been developed to accommodate various forms of auxiliary information. For example, [9] considered multiple testing of grouped hypotheses. The authors proposed an optimal data-driven procedure that uniformly improves the pooled and separate analyses. [44] developed an Lfdr-based method to incorporate spatial information. [40; 47] proposed EM-type algorithms to estimate the Lfdr by taking into account covariate and spatial information, respectively.

Other related works include [18], which considers the two-group mixture models with side-information. [13] develops a method for estimating the constrained optimal weights for Bonferroni multiple testing. [7] proposes an FDR-controlling procedure based on the covariate-dependent null probabilities.

In this paper, we develop a new method along the line of research on Lfdr-based approaches by adaptively estimating the prior probabilities of being null in Lfdr that reflect auxiliary information in multiple testing. The proposed Lfdr-based procedure is built on the optimal rejection rule as shown in Section 2.1 and thus is expected to be more powerful than the weighted-BH procedure when the underlying two-group mixture model is correctly specified. Compared to existing work on Lfdr-based methods, our contributions are three-fold. (i) We outline a general framework for incorporating various forms of auxiliary information. This is achieved by allowing the prior probabilities of being null to vary across different hypotheses. We propose a data-adaptive step-up procedure and show that it provides asymptotic FDR control when relevant consistent estimates are available. (ii) Focusing on the ordered structure, where auxiliary information generates a ranked list of hypotheses, we develop a new EM-type algorithm [12] to estimate the prior probabilities of being null and the distribution of p-values under the alternative hypothesis simultaneously. Under monotone constraint on the density function of p-values under the alternative hypothesis, we utilize the Pool-Adjacent-Violators Algorithm (PAVA) to estimate both the prior probabilities of being null and the density function of p-values under the al-

ternative hypothesis (see [20] for early work on this kind of problems). Due to the efficiency of PAVA, our method is scalable to large datasets arising in genomic studies. (iii) We prove asymptotic FDR control for our procedure and obtain some consistency results for the estimates of the prior probabilities of being null and the alternative density, which is of independent theoretical interest. Finally, to allow users to conveniently implement our method and reproduce the numerical results reported in Sections 5-6, we make our code publicly available at `https://github.com/jchen1981/OrderShapeEM`.

The problem we considered is related but different from the one in [21; 33], where the authors seek the largest cutoff k so that one rejects the first k hypotheses while accepts the remaining ones. So their method always rejects an initial block of hypotheses. In contrast, our procedure allows researchers to reject the kth hypothesis but accept the $k-1$th hypothesis in the ranked list. In other words, we do not follow the order restriction strictly. Such flexibility could result in a substantial power increase when the order information is not very strong or even weak, as observed in our numerical studies. Also see the discussions on monotonicity in Section 1.1 of [40].

To account for the potential mistakes in the ranked list or to improve power by incorporating external covariates, alternative methods have been proposed in the literature. For example, [36] extends the fixed sequence method to allow more than one acceptance before stopping. [32] modifies AdaPT in [30] by giving analysts the power to enforce the ordered constraint on the final rejection set. Though aiming for addressing a similar issue, our method is motivated from the empirical Bayes perspective, and it is built on the two-group mixture model that allows the prior probabilities of being null to vary across different hypotheses. The implementation and theoretical analysis of our method are also quite different from those in [32; 36].

Finally, it is also worth highlighting the difference with respect to the recent work [11] which is indeed closely related to ours. First of all, our Theorem 3.3 concerns about the two-group mixture models with decreasing alternative density, while Theorem 3.1 in [11] focuses on a mixture of Gaussians. We generalize the arguments in [48] by considering a transformed class of functions to relax the boundedness assumption on the class of decreasing densities. A careful inspection of the proof of Theorem 3.3 reveals that the techniques we develop are quite different from those in [11]. Second, we provide a more detailed empirical and theoretical analysis of the FDR-controlling procedure. In particular, we prove that the step-up procedure based on our Lfdr estimates asymptotically controls the FDR and provide the corresponding power analysis. We also conduct extensive simulation studies to evaluate the finite sample performance of the proposed

Lfdr-based procedure.

The rest of the paper proceeds as follows. Section 2 proposes a general multiple testing procedure that incorporates auxiliary information to improve statistical power, and establishes its asymptotic FDR control property. In Section 3, we introduce a new EM-type algorithm to estimate the unknowns and study the theoretical properties of the estimators. We discuss two extensions in Section 4. Section 5 and Section 6 are devoted respectively to simulation studies and data analysis. We conclude the paper in Section 7. All the proofs of the main theorems and technical lemmas are collected in the Appendix.

2. Covariate-adjusted multiple testing. In this section, we describe a covariate-adjusted multiple testing procedure based on Lfdr.

2.1. *Optimal rejection rule.* Consider simultaneous testing of m hypotheses H_i for $i = 1, \ldots, m$ based on m p-values $x_1, \ldots, x_m$, where x_i is the p-value corresponding to the ith hypothesis H_i. Let $\theta_i, i = 1, \ldots, m$ indicate the underlying truth of the ith hypothesis. In other words, $\theta_i = 1$ if H_i is non-null/alternative and $\theta_i = 0$ if H_i is null. We allow the probability that $\theta_i = 0$ to vary across i. In this way, auxiliary information can be incorporated through

$$P(\theta_i = 0) = \pi_{0i}, \quad i = 1, \ldots, m. \tag{1}$$

Consider the two-group model for the p-values (see e.g., [15] and Chapter 2 of [16]):

$$x_i \mid \theta_i \sim (1 - \theta_i) f_0 + \theta_i f_1, \quad i = 1, \ldots, m, \tag{2}$$

where f_0 is the density function of the p-values under the null hypothesis and f_1 is the density function of the p-values under the alternative hypothesis. The marginal probability density function of x_i is equal to

$$f^i(x) = \pi_{0i} f_0(x) + (1 - \pi_{0i}) f_1(x). \tag{3}$$

We briefly discuss the identifiability of the above model. Suppose f_0 is known and bounded away from zero and infinity. Consider the following class of functions:

$$\mathbf{F}_m = \big\{ \tilde{\mathbf{f}} = (\tilde{f}^1, \ldots, \tilde{f}^m) \text{ with } \tilde{f}^i = \tilde{\pi}_i f_0 + (1 - \tilde{\pi}_i) \tilde{f}_1 : \min_{x \in [0,1]} \tilde{f}_1(x) = 0, \\ 0 \le \tilde{\pi}_i \le 1, \min_i \tilde{\pi}_i < 1 \big\}.$$

Suppose $\tilde{\mathbf{f}}, \breve{\mathbf{f}} \in \mathbf{F}_m$, where the ith components of $\tilde{\mathbf{f}}$ and $\breve{\mathbf{f}}$ are given by $\tilde{f}^i = \tilde{\pi}_i f_0 + (1-\tilde{\pi}_i)\tilde{f}_1$ and $\breve{f}^i = \breve{\pi}_i f_0 + (1-\breve{\pi}_i)\breve{f}_1$ respectively. We show that if $\tilde{f}^i(x) = \breve{f}^i(x)$ for all x and i, then $\tilde{f}_1(x) = \breve{f}_1(x)$ and $\tilde{\pi}_i = \breve{\pi}_i$ for all x and i. Suppose $\tilde{f}_1(x') = 0$ for some $x' \in [0,1]$. If $\tilde{\pi}_i < \breve{\pi}_i$ for some i, then we have

$$0 = \frac{\tilde{f}_1(x')}{f_0(x')} = \frac{\breve{\pi}_i - \tilde{\pi}_i}{1-\tilde{\pi}_i} + \frac{(1-\breve{\pi}_i)\breve{f}_1(x')}{(1-\tilde{\pi}_i)f_0(x')} > 0, \tag{4}$$

which is a contradiction. Similarly, we get a contradiction when $\tilde{\pi}_i > \breve{\pi}_i$ for some i. Thus we have $\tilde{\pi}_i = \breve{\pi}_i$ for all i. As there exists a i such that $1-\tilde{\pi}_i = 1-\breve{\pi}_i > 0$, it is clear that $\tilde{f}^i(x) = \breve{f}^i(x)$ implies that $\tilde{f}_1(x) = \breve{f}_1(x)$.

In statistical and scientific applications, the goal is to separate the alternative cases ($\theta_i = 1$) from the null cases ($\theta_i = 0$). This can be formulated as a multiple testing problem, with solutions represented by a decision rule $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \{0,1\}^m$. It turns out that the optimal decision rule is closely related to the Lfdr defined as

$$\mathrm{Lfdr}_i(x) := P(\theta_i = 0 \mid x_i = x) = \frac{\pi_{0i} f_0(x)}{\pi_{0i} f_0(x) + (1-\pi_{0i}) f_1(x)} = \frac{\pi_{0i} f_0(x)}{f^i(x)}.$$

In other words, $\mathrm{Lfdr}_i(x)$ is the posterior probability that a case is null given the corresponding p-value is equal to x. It combines the auxiliary information (π_{0i}) and data from the current experiment. Information across tests is used in forming $f_0(\cdot)$ and $f_1(\cdot)$.

Optimal decision rule under mixture model has been extensively studied in the literature, see e.g., [46; 30; 3]. For completeness, we present the derivations below and remark that they follow somewhat directly from existing results. Consider the expected number of false positives (EFP) and true positives (ETP) of a decision rule. Suppose that x_i follows the mixture model (2) and we intend to reject the ith null hypothesis if $x_i \le c_i$. The size and power of the ith test are given respectively by

$$\alpha_i(c_i) = \int_0^{c_i} f_0(t)dt \quad \text{and} \quad \beta_i(c_i) = \int_0^{c_i} f_1(t)dt.$$

It thus implies that

$$\mathrm{EFP}(\mathbf{c}) = \sum_{i=1}^m \pi_{0i}\alpha_i(c_i) \quad \text{and} \quad \mathrm{ETP}(\mathbf{c}) = \sum_{i=1}^m (1-\pi_{0i})\beta_i(c_i),$$

where $\mathbf{c} = (c_1, \ldots, c_m)$. We wish to maximize ETP for a given value of the marginal FDR (mFDR) defined as

$$\mathrm{mFDR}(\mathbf{c}) = \frac{\mathrm{EFP}(\mathbf{c})}{\mathrm{ETP}(\mathbf{c}) + \mathrm{EFP}(\mathbf{c})}, \tag{5}$$

by an optimum choice of the cutoff value $\mathbf{c}$. Formally, consider the problem

$$\max_{\mathbf{c}} \mathrm{ETP}(\mathbf{c}) \quad \text{subject to} \quad \mathrm{mFDR}(\mathbf{c}) \leq \alpha. \tag{6}$$

A standard Lagrange multiplier argument gives the following result which motivates our choice of thresholds.

PROPOSITION 2.1. *Assume that f_1 is continuously non-increasing, and f_0 is continuously non-decreasing and uniformly bounded from above. Further assume that for a pre-specified $\alpha > 0$,*

$$\min_i \frac{(1-\pi_{0i})f_1(0)}{\pi_{0i}f_0(0)} > \frac{1-\alpha}{\alpha}. \tag{7}$$

Then (6) has at least one solution and every solution $(\tilde{c}_1, \ldots, \tilde{c}_m)$ satisfies

$$Lfdr_i(\tilde{c}_i) = \tilde{\lambda}$$

for some $\tilde{\lambda}$ that is independent of i.

The proof of Proposition 2.1 is similar to that of Theorem 2 in [30] and we omit the details. Under the monotone likelihood ratio assumption [45; 10]:

$$f_1(x)/f_0(x) \text{ is decreasing in } x, \tag{8}$$

we obtain that $\mathrm{Lfdr}_i(x)$ is monotonically increasing in x. Therefore, we may reduce our attention to the rejection rule $\mathbf{I}\{x_i \leq c_i\}$ as

$$\delta_i = \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \leq \lambda\} \tag{9}$$

for a constant λ to be determined later.

2.2. *Asymptotic FDR control.* To fully understand the proposed method, we gradually investigate its theoretical properties through several steps, starting with an oracle procedure which provides key insights into the problem. Assume that $\{\pi_{0i}\}_{i=1}^m$, $f_0(\cdot)$ and $f_1(\cdot)$ are known. The proposed method utilizes auxiliary information through $\{\pi_{0i}\}_{i=1}^m$ and information from the alternative through $f_1(\cdot)$ in addition to information from the null, upon which conventional approaches are based. In view of (9), the number of false rejections equals to

$$V_m(\lambda) = \sum_{i=1}^m \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \leq \lambda\}(1-\theta_i)$$

and the total number of rejections is given by

$$D_{m,0}(\lambda) = \sum_{i=1}^{m} \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\}.$$

Write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We aim to find the critical value λ in (9) that controls the FDR, which is defined as $\mathrm{FDR}_m(\lambda) = E\{V_m(\lambda)/(D_{m,0}(\lambda) \vee 1)\}$ at a pre-specified significance level α. Note that

(10)
$$E[V_m(\lambda)] = \sum_{i=1}^{m} \pi_{0i} P(\mathrm{Lfdr}_i(x_i) \le \lambda | \theta_i = 0) = \sum_{i=1}^{m} E[\mathrm{Lfdr}_i(x_i) \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\}].$$

An estimate of the $\mathrm{FDR}_m(\lambda)$ is given by

$$\mathrm{FDR}_m(\lambda) = \frac{\sum_{i=1}^{m} \mathrm{Lfdr}_i(x_i) \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\}}{\sum_{i=1}^{m} \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\}}.$$

Let $\lambda_m = \sup\{\lambda \in [0, 1] : \mathrm{FDR}_m(\lambda) \le \alpha\}$. Then reject H_i if $\mathrm{Lfdr}_i(x_i) \le \lambda_m$. Below we show that the above (oracle) step-up procedure provides asymptotic control on the FDR under the following assumptions.

(C1) Assume that for any $\lambda \in [0, 1]$,

$$\frac{1}{m} \sum_{i=1}^{m} \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\} \to^p D_0(\lambda),$$
$$\frac{1}{m} \sum_{i=1}^{m} \mathrm{Lfdr}_i(x_i) \mathbf{I}\{\mathrm{Lfdr}_i(x_i) \le \lambda\} \to^p D_1(\lambda),$$

and

(11)
$$\frac{1}{m} V_m(\lambda) \to^p D_1(\lambda),$$

where D_0 and D_1 are both continuous functions over $[0, 1]$.

(C2) Write $R(\lambda) = D_1(\lambda)/D_0(\lambda)$, where D_0 and D_1 are defined in (C1). There exists a $\lambda_\infty \in (0, 1]$ such that $R(\lambda_\infty) < \alpha$.

We remark that (C1) is similar to those for Theorem 4 in [42]. In view of (10), (11) follows from the weak law of large numbers. Note that (C1) allows certain forms of dependence, such as m-dependence, ergodic dependence and certain mixing type dependence. (C2) ensures the existence of the critical value λ_m to asymptotically control the FDR at level α. The following proposition shows that the oracle step-up procedure provides asymptotic FDR control.

PROPOSITION 2.2. *Under conditions (C1)-(C2),*

$$\limsup_{m\to\infty} FDR_m(\lambda_m) \le \alpha.$$

The proof of Proposition 2.2 is relegated in the Appendix. In the following, we mimic the operation of the oracle procedure and provide an adaptive procedure. In the inference problems that we are interested in, the p-value distribution under the null hypothesis is assumed to be known (e.g., the uniform distribution on $[0,1]$, or can be obtained from the distributional theory of the test statistic in question). Below we assume f_0 is known and remark that our result still holds provided that f_0 can be consistently estimated. In practice, f_1 and $\{\pi_{0i}\}_{i=1}^m$ are often unknown and replaced by their sample counterparts. Let $\hat{f}_1(\cdot)$ and $\{\hat{\pi}_{0i}\}_{i=1}^m$ be the estimators of $f_1(\cdot)$ and $\{\pi_{0i}\}_{i=1}^m$ respectively. Define

$$\widehat{\mathrm{Lfdr}}_i(x) = \frac{\hat{\pi}_{0i} f_0(x)}{\hat{\pi}_{0i} f_0(x) + (1-\hat{\pi}_{0i})\hat{f}_1(x)} = \frac{\hat{\pi}_{0i} f_0(x)}{\hat{f}^i(x)},$$

where $\hat{f}^i(x) = \hat{\pi}_{0i} f_0(x) + (1-\hat{\pi}_{0i})\hat{f}_1(x)$. A natural estimate of λ_m can be obtained through

$$\hat{\lambda}_m = \sup\left\{\lambda \in [0,1] : \frac{\sum_{i=1}^m \widehat{\mathrm{Lfdr}}_i(x_i)\mathbf{I}\{\widehat{\mathrm{Lfdr}}_i(x_i) \le \lambda\}}{\sum_{i=1}^m \mathbf{I}\{\widehat{\mathrm{Lfdr}}_i(x_i) \le \lambda\}} \le \alpha\right\}.$$

Reject the ith hypothesis if $\widehat{\mathrm{Lfdr}}_i(x_i) \le \hat{\lambda}_m$. This is equivalent to the following step-up procedure that was originally proposed in [45]. Let $\widehat{\mathrm{Lfdr}}_{(1)} \le \cdots \le \widehat{\mathrm{Lfdr}}_{(m)}$ be the order statistics of $\{\widehat{\mathrm{Lfdr}}_1(x_1), \ldots, \widehat{\mathrm{Lfdr}}_m(x_m)\}$ and denote by $H^{(1)}, \ldots, H^{(m)}$ the corresponding ordered hypotheses. Define

$$\hat{k} := \max\left\{1 \le i \le m : \frac{1}{i}\sum_{j=1}^{i} \widehat{\mathrm{Lfdr}}_{(j)} \le \alpha\right\};$$

then reject all $H^{(i)}$ for $i = 1, \ldots, \hat{k}$.

We show that this step-up procedure provides asymptotic control on the FDR. To facilitate the derivation, we make the following additional assumption.

(C3) Assume that

$$\frac{1}{m}\sum_{i=1}^m |\widehat{\mathrm{Lfdr}}_i(x_i) - \mathrm{Lfdr}_i(x_i)| \to^p 0.$$

(C3) requires the Lfdr estimators to be consistent in terms of the empirical L_1 norm. We shall justify Condition (C3) in Section 3.3.

THEOREM 2.3. *Under Conditions (C1)-(C3),*

$$\limsup_{m\to\infty} FDR_m(\hat{\lambda}_m) \le \alpha.$$

Theorem 2.3 indicates that we can obtain asymptotic control on the FDR using the data-adaptive procedure when relevant consistent estimates are available. Similar algorithm has been obtained in [45], where it is assumed that the hypotheses are exchangeable in the sense that $\pi_{01} = \cdots = \pi_{0m}$.

3. Estimating the unknowns.

3.1. *The density function $f_1(\cdot)$ is known.* We first consider the case that $f_0(\cdot)$ and $f_1(\cdot)$ are both known. Under such setup, we need to estimate m unknown parameters $\pi_{0i}, i = 1, \ldots, m$, which is prohibitive without additional constraints. One constraint that makes the problem solvable is the monotone constraint. In statistical genetics and genomics, investigators can use auxiliary information (e.g., p-values from previous or related studies) to generate a ranked list of hypotheses $H_1, \ldots, H_m$ even before performing the experiment, where H_1 is the hypothesis that the investigator believes to most likely correspond to a true signal, while H_m is the one believed to be least likely. Specifically, let $\Pi_0 = (\pi_{01}, \ldots, \pi_{0m}) \in (0,1)^m$. Define the convex set

$$\mathcal{M} = \{\Pi = (\pi_1, \ldots, \pi_m) \in (0,1)^m : 0 \le \pi_1 \le \ldots \le \pi_m \le 1\}.$$

We illustrate the motivation for the monotone constraint with an example.

EXAMPLE 3.1. Suppose that we are given data consisting of a pair of values (x_{i1}, x_{i2}), where x_{i1} represents the p-value, x_{i2} represents auxiliary information and they are independent conditional on the hidden true state θ_i for $i = 1, \ldots, m$. Suppose

$$x_{ij} \mid \theta_i \overset{\text{ind}}{\sim} (1-\theta_i) f_{0,j}(x_{ij}) + \theta_i f_{1,j}(x_{ij}), \qquad i = 1, \ldots, m, \quad j = 1, 2, \tag{12}$$

where $\theta_i = 1$ if H_i is alternative and $\theta_i = 0$ if H_i is null, $f_{0,j}(\cdot)$ is the density function of p-values or auxiliary variables under the null hypothesis and $f_{1,j}(\cdot)$ is the density function of p-values or auxiliary variables under the alternative hypothesis. Suppose $P(\theta_i = 0) = \tau_0$ for all $i = 1, \ldots, m$.

Using the Bayes rule and the independence between x_{i1} and x_{i2} given $\theta_i, i = 1, \ldots, m$, we have the conditional distribution of $x_{i1} \mid x_{i2}$ as follows:

$$\begin{aligned}
& f(x_{i1} \mid x_{i2}) \\
=& \frac{f(x_{i1}, x_{i2} \mid \theta_i = 0)\tau_0 + f(x_{i1}, x_{i2} \mid \theta_i = 1)(1-\tau_0)}{f(x_{i2} \mid \theta_i = 0)\tau_0 + f(x_{i2} \mid \theta_i = 1)(1-\tau_0)} \\
=& \frac{f(x_{i1} \mid \theta_i = 0)f(x_{i2} \mid \theta_i = 0)\tau_0 + f(x_{i1} \mid \theta_i = 1)f(x_{i2} \mid \theta_i = 1)(1-\tau_0)}{f(x_{i2} \mid \theta_i = 0)\tau_0 + f(x_{i2} \mid \theta_i = 1)(1-\tau_0)} \\
=& \frac{f_{0,1}(x_{i1})f_{0,2}(x_{i2})\tau_0 + f_{1,1}(x_{i1})f_{1,2}(x_{i2})(1-\tau_0)}{f_{0,2}(x_{i2})\tau_0 + f_{1,2}(x_{i2})(1-\tau_0)} \\
=& f_{0,1}(x_{i1})\gamma_0(x_{i2}) + f_{1,1}(x_{i1})(1-\gamma_0(x_{i2})),
\end{aligned}$$

where

$$\gamma_0(x) = \frac{f_{0,2}(x)\tau_0}{f_{0,2}(x)\tau_0 + f_{1,2}(x)(1-\tau_0)} = \frac{\tau_0}{\tau_0 + \frac{f_{1,2}(x)}{f_{0,2}(x)}(1-\tau_0)}.$$

If $f_{1,2}(x)/f_{0,2}(x)$ is a monotonic function, so is $\gamma_0(x)$. Therefore, the order of x_{i2} generates a ranked list of the hypotheses $H_1, \ldots, H_m$ through the conditional prior probability $\gamma_0(x)$.

We estimate Π_0 by solving the following maximum likelihood problem:

$$\begin{aligned}
\hat{\Pi}_0 &= (\hat{\pi}_{01}, \ldots, \hat{\pi}_{0m}) = \operatorname*{argmax}_{\Pi=(\pi_1,\ldots,\pi_m)\in\mathcal{M}} l_m(\Pi), \\
l_m(\Pi) &:= \sum_{i=1}^{m} \log\left\{\pi_i f_0(x_i) + (1-\pi_i) f_1(x_i)\right\}.
\end{aligned} \tag{13}$$

It is easy to see that (13) is a convex optimization problem. Let $\phi(x, a) = af_0(x) + (1-a)f_1(x)$. To facilitate the derivations, we shall assume that $f_0(x_i) \neq f_1(x_i)$ for all i, which is a relatively mild requirement. Under this assumption, it is straightforward to see that for any $1 \leq k \leq l \leq m$, $\sum_{i=k}^{l} \log \phi(x_i, a)$ is a strictly concave function for $0 < a < 1$. Let $\hat{a}_{kl} = \operatorname{argmax}_{a\in[0,1]} \sum_{i=k}^{l} \log \phi(x_i, a)$ be the unique maximizer. According to Theorem 3.1 of [37], we have

$$\hat{\pi}_{0i} = \max_{1\leq k\leq i} \min_{i\leq l\leq m} \hat{a}_{kl}. \tag{14}$$

However, this formula is not practically useful due to the computational burden when m is very large. Below we suggest a more efficient way to solve problem (13). A general algorithm when f_1 is unknown is provided in the

next subsection. The main computational tools are the EM algorithm for two-group mixture model and the Pool-Adjacent-Violator-Algorithm from isotonic regression for the monotone constraint on the prior probability of null hypothesis [12; 38]. We provide the derivation of the EM algorithm from the full data likelihood in the appendix. In particular, let $\Pi^{(t)} = (\hat{\pi}_{01}^{(t)}, \ldots, \hat{\pi}_{0m}^{(t)})$ be the solution at the tth iteration. Define

$$Q_j^{(t)} := Q_j^{(t)}(\hat{\pi}_{0j}^{(t)}) = \frac{\hat{\pi}_{0j}^{(t)} f_0(x_j)}{\hat{\pi}_{0j}^{(t)} f_0(x_j) + (1 - \hat{\pi}_{0j}^{(t)}) f_1(x_j)},$$

$$Q(\Pi|\Pi^{(t)}) = \sum_{j=1}^{m} \{Q_j^{(t)} \log(\pi_j) + (1 - Q_j^{(t)}) \log(1 - \pi_j)\}.$$

At the $(t+1)$th iteration of the EM algorithm, we solve the following problem,

$$\Pi^{(t+1)} = \underset{\Pi=(\pi_1,\ldots,\pi_m)\in\mathcal{M}}{\operatorname{argmax}} Q(\Pi|\Pi^{(t)}). \tag{15}$$

By Theorem 1.5.1 of [38] or Theorem 3.1 of [37], we only need to solve the isotonic regression problem,

$$\Pi^{(t+1)} = \underset{\Pi=(\pi_1,\ldots,\pi_m)\in\mathcal{M}}{\operatorname{argmin}} \sum_{j=1}^{m} \left\{Q_j^{(t)} - \pi_j\right\}^2. \tag{16}$$

The solution to (16) has an explicit form given by the max-min formula,

$$\hat{\pi}_{0i}^{(t+1)} = \max_{a\le i} \min_{b\ge i} \frac{\sum_{j=a}^{b} Q_j^{(t)}}{b-a+1},$$

which can be obtained conveniently using the Pool-Adjacent-Violators Algorithm (PAVA) [38]. Note that if $Q_1^{(t)} \ge Q_2^{(t)} \ge \cdots Q_m^{(t)}$, then the solution to (16) is simply given by $\hat{\pi}_{0i}^{(t+1)} = \sum_{j=1}^{m} Q_j^{(t)}/m$ for all $1 \le i \le m$. As the EM algorithm is a hill-climbing algorithm, it is not hard to show that $l_m(\Pi^{(t)})$ is a non-decreasing function of t.

We study the asymptotic consistency of the true maximum likelihood estimator $\hat{\Pi}_0$ which can be represented as (14). To this end, consider the model

$$x_i \overset{\text{ind}}{\sim} \pi_{0i} f_0 + (1 - \pi_{0i}) f_1, \quad \pi_{0i} = \pi_0(i/m)^{[1]},$$

for some non-decreasing function $\pi_0 : [0,1] \to [0,1]$. Our first result concerns the point-wise consistency for each $\hat{\pi}_{0i}$. For a set A, denote by $\text{card}(A)$ its cardinality.

[1]For the ease of presentation, we suppress the dependence on m in π_{0i}.

THEOREM 3.1. *Assume that* $\int (\log f_i(x))^2 f_j(x)dx < \infty$ *for* $i, j = 0, 1$, *and* $P(f_0(x_i) = f_1(x_i)) = 0$. *Suppose* $0 < \pi_0(0) \le \pi_0(1) < 1$. *For any* $\epsilon > 0$, *let* $0 \le t' < i_0/m < t'' \le 1$ *such that* $|\pi_0(t') - \pi_0(i_0/m)| \vee |\pi_0(t'') - \pi_0(i_0/m)| < \epsilon/2$. *Denote* $A_1 = \{i : t' \le i/m \le i_0/m\}$ *and* $A_2 = \{i : i_0/m \le i/m \le t''\}$. *For* $card(A_1) \wedge card(A_2) \ge N$, *we have*

$$P\left(|\hat{\pi}_{0,i_0} - \pi_{0,i_0}| < \epsilon\right) \ge 1 - O\left(\frac{1}{\epsilon^2 N}\right).$$

The condition on the cardinalities of A_1 and A_2 guarantees that there are sufficient observations around i_0/m, which allows us to borrow information to estimate π_{0,i_0} consistently. The assumption $P(f_0(x_i) = f_1(x_i)) = 0$ ensures that the maximizer $\hat{a}_{kl}$ is unique for $1 \le k \le l \le m$. It is fulfilled if the set $\{x \in [0,1] : f_0(x) = f_1(x)\}$ has zero Lebesgue measure. As a direct consequence of Theorem 3.1, we have the following uniform consistency result of $\hat{\Pi}_0$. Due to the monotonicity, the uniform convergence follows from the pointwise convergence.

COROLLARY 3.2. *For* $\epsilon > 0$, *suppose there exists a set* $i_1 < i_2 < \cdots < i_l$, *where each* i_k *satisfies the assumption for* i_0 *in Theorem 3.1 and that* $\max_{2 \le k \le l}(\pi_{0,i_k} - \pi_{0,i_{k-1}}) < \epsilon$. *Then we have*

$$P\left(\max_{i_1 \le i \le i_l} |\hat{\pi}_{0,i} - \pi_{0,i}| < \epsilon\right) \ge 1 - O\left(\frac{l}{\epsilon^2 N}\right).$$

REMARK 3.1. Suppose π_0 is Lipschitz continuous with the Lipschitz constant K. Then we can set $t'' = (i_0 - 1)/m + \epsilon/(2K)$, $t' = (i_0 + 1)/m - \epsilon/(2K)$ and thus $N = \lfloor m\epsilon/(2K) \rfloor$. Our result suggests that

$$P\left(|\hat{\pi}_{0,i_0} - \pi_{0,i_0}| < \epsilon\right) \ge 1 - O\left(\frac{K}{\epsilon^3 m}\right),$$

which implies that $|\hat{\pi}_{0,i_0} - \pi_{0,i_0}| = O_p(m^{-1/3})$.

3.2. *The density function* $f_1(\cdot)$ *is unknown.* In practice, f_1 and Π_0 are both unknown. We propose to estimate f_1 and Π_0 by maximizing the likelihood, i.e.,

$$(\hat{\Pi}_0, \hat{f}_1) = \underset{\Pi \in \mathcal{M}, \tilde{f}_1 \in \mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^{m} \log\left\{\pi_i f_0(x_i) + (1 - \pi_i)\tilde{f}_1(x_i)\right\}, \tag{17}$$

where $\mathcal{H}$ is a pre-specified class of density functions. In (17), $\mathcal{H}$ might be the class of beta mixtures or the class of decreasing density functions. Problem

(17) can be solved by Algorithm 1. A derivation of Algorithm 1 from the full data likelihood that has access to latent variables is provided in the Appendix. Our algorithm is quite general in the sense that it allows users to specify their own updating scheme for the density components in (19). Both parametric and non-parametric methods can be used to estimate f_1.

Algorithm 1

0. Input the initial values $(\Pi^{(0)}, f_1^{(0)})$.
1. **E-step:** Given $(\hat{\Pi}^{(t)}, \hat{f}_1^{(t)})$, let

$$Q_i^{(t)} = \frac{\hat{\pi}_{0i}^{(t)} f_0(x_i)}{\hat{\pi}_{0i}^{(t)} f_0(x_i) + (1 - \hat{\pi}_{0i}^{(t)}) \hat{f}_1^{(t)}(x_i)}.$$

2. **M-step:** Given $Q_i^{(t)}$, update (Π, f_1) through

$$(\hat{\pi}_{01}^{(t+1)}, \ldots, \hat{\pi}_{0m}^{(t+1)}) = \underset{\Pi=(\pi_1,\ldots,\pi_m)\in\mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{m} \left(Q_i^{(t)} - \pi_i\right)^2, \tag{18}$$

and

$$\hat{f}_1^{(t+1)} = \underset{\tilde{f}_1\in\mathcal{H}}{\operatorname{argmax}} \sum_{i=1}^{m} (1 - Q_i^{(t)}) \log \tilde{f}_1(x_i). \tag{19}$$

3. Repeat the above E-step and M-step until the algorithm converges.

In the multiple testing literature, it is common to assume that f_1 is a decreasing density function (e.g., smaller p-values imply stronger evidence against the null), see e.g. [29]. As an example of the general algorithm, let $\mathcal{H}$ denote the class of decreasing density functions. We shall discuss how (19) can be solved using the PAVA. The key recipe is to use Theorem 3.1 of [4] in obtaining f_1 evaluated at the observed p-values. Specifically, it can be accomplished by a series of steps outlined below. Define the order statistics of $\{x_i\}$ as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(m)}$. Let $Q_{(i)}^{(t)}$ be the corresponding $Q_i^{(t)}$ that is associated with $x_{(i)}$.

Step 1: The objective function in (19) only looks at the value of f_1 at $x_{(i)}$. The objective function increases if $f_1(x_{(i)})$ increases, and the value of f_1 at $(x_{(i-1)}, x_{(i)})$ has no impact on the objective function (where $x_{(0)} = 0$). Therefore, if f maximizes the objective function, there is a solution that is constant on $(x_{(i-1)}, x_{(i)}]$.

Step 2: Let $y_i = f_1(x_{(i)})$. We only need to find y_i which maximizes

$$\sum_{i=1}^{m} (1 - Q_{(i)}^{(t)}) \log(y_i),$$

subject to $y_1 \geq y_2 \geq \cdots \geq y_m \geq 0$ and $\sum_{i=1}^m y_i(x_{(i)} - x_{(i-1)}) = 1$. It can be formulated as a convex programming problem which is tractable. In Steps 3 and 4 below, we further translate it into an isotonic regression problem.
Step 3: Write $Q^{(t)} = \sum_{i=1}^m (1 - Q_{(i)}^{(t)})$. Consider the problem:

$$\min \sum_{i=1}^m \left\{ -(1 - Q_{(i)}^{(t)}) \log(y_i) + Q^{(t)} y_i (x_{(i)} - x_{(i-1)}) \right\}.$$

The solution is given by $\hat{y}_i = \frac{1 - Q_{(i)}^{(t)}}{Q^{(t)}(x_{(i)} - x_{(i-1)})}$, which satisfies the constraint $\sum_{i=1}^m y_i(x_{(i)} - x_{(i-1)}) = 1$ in Step 2.
Step 4: Rewrite the problem in Step 3 as

$$\min \sum_{i=1}^m (1 - Q_{(i)}^{(t)}) \left\{ -\log(y_i) - \frac{-Q^{(t)}(x_{(i)} - x_{(i-1)})}{(1 - Q_{(i)}^{(t)})} y_i \right\}.$$

This is the generalized isotonic regression problem considered in Theorem 3.1 of [4]. Let

$$(\hat{u}_1, \ldots, \hat{u}_m) = \operatorname{argmin} \sum_{i=1}^m (1 - Q_{(i)}^{(t)}) \left(-\frac{Q^{(t)}(x_{(i)} - x_{(i-1)})}{(1 - Q_{(i)}^{(t)})} - u_i \right)^2$$

subject to $u_1 \geq u_2 \geq \cdots \geq u_m$. The solution is given by the max-min formula

$$\hat{u}_i = \max_{b \geq i} \min_{a \leq i} \frac{-Q^{(t)} \sum_{j=a}^b (x_{(j)} - x_{(j-1)})}{\sum_{j=a}^b (1 - Q_{(j)}^{(t)})},$$

which can be obtained using the PAVA. By Theorem 3.1 of [4], we arrive at the solution to the original problem (19) by letting $\tilde{y}_i = -\frac{1}{\hat{u}_i}$. Therefore, in the EM-algorithm, one can employ the PAVA to estimate both the prior probabilities of being null and the p-value density function under the alternative hypothesis. Because of this, our algorithm is fast and tuning parameter free, and is very easy to implement in practice.

3.3. *Asymptotic convergence and verification of Condition (C3).* In this subsection, we present some convergence results regarding the proposed estimators in Section 3.2. Furthermore, we propose a refined estimator for π_0, and justify Condition (C3) for the corresponding Lfdr estimator. Throughout the following discussions, we assume that

$$x_i \sim f^i = \pi_0(i/m) f_0 + (1 - \pi_0(i/m)) f_1$$

independently for $1 \leq i \leq m$ and $\pi_0 : [0,1] \to [0,1]$ with $\pi_0(i/m) = \pi_{0i}$. Let $\mathcal{F}$ be the class of densities defined on $[0,1]$. For $f, g \in \mathcal{F}$, we define the squared Hellinger-distance as

$$H^2(f,g) = \frac{1}{2}\int_0^1 (\sqrt{f(x)} - \sqrt{g(x)})^2 dx = 1 - \int_0^1 \sqrt{f(x)g(x)}dx.$$

Suppose the true alternative density f_1 belongs to a class of decreasing density functions $\mathcal{H} \subset \mathcal{F}$. Let $\Xi = \{\pi : [0,1] \to [0,1], 0 < \varepsilon < \pi(0) \leq \pi(1) < 1 - \varepsilon < 1, \text{ and } \pi(\cdot) \text{ is nondecreasing}\}$ and assume that $\pi_0 \in \Xi$. Consider $\tilde{f}^i = \tilde{\pi}(i/m)f_0 + (1 - \tilde{\pi}(i/m))\tilde{f}_1$ and $\breve{f}^i = \breve{\pi}(i/m)f_0 + (1 - \breve{\pi}(i/m))\breve{f}_1$ for $1 \leq i \leq m$, $\tilde{f}_1, \breve{f}_1 \in \mathcal{H}$ and $\tilde{\pi}, \breve{\pi} \in \Xi$. Define the average squared Hellinger-distance between $(\tilde{\pi}, \tilde{f}_1)$ and $(\breve{\pi}, \breve{f}_1)$ as

$$H_m^2((\tilde{\pi}, \tilde{f}_1), (\breve{\pi}, \breve{f}_1)) = \frac{1}{m}\sum_{i=1}^m H^2(\tilde{f}^i, \breve{f}^i).$$

Suppose $(\hat{\pi}_0, \hat{f}_1)$ is an estimator of (π_0, f_1) such that

$$\sum_{i=1}^m \log\left(\frac{2\hat{f}^i(x_i)}{\hat{f}^i(x_i) + f^i(x_i)}\right) \geq 0,$$

where $\hat{f}^i(x) = \hat{\pi}_0(i/m)f_0(x) + (1 - \hat{\pi}_0(i/m))\hat{f}_1(x)$. Note that we do not require $(\hat{\pi}_0, \hat{f}_1)$ to be the global maximizer of the likelihood. We have the following result concerning the convergence of $(\hat{\pi}_0, \hat{f}_1)$ to (π_0, f_1) in terms of the average squared Hellinger-distance.

THEOREM 3.3. *Suppose $\pi_0 \in \Xi$, $f_0 \equiv 1$, and $f_1 \in \mathcal{H}$. Under the assumption that $\int_0^1 f_1^{1+a}(x)dx < \infty$ for some $0 < a \leq 1$, we have*

$$P\left(H_m((\pi_0, f_1), (\hat{\pi}_0, \hat{f}_1)) > Mm^{-1/3}\right) \leq M_1 \exp(-M_2 m^{1/3}),$$

for some M, M_1 and $M_2 > 0$. We remark that $f_1(x) = (1-\gamma)x^{-\gamma}$ with $0 < \gamma < 1$ satisfies $\int_0^1 f_1^{1+a}(x)dx < \infty$ for $0 < a < (1/\gamma - 1) \wedge 1$.

Theorem 3.3 follows from an application of Theorem 8.14 in [48]. By the Cauchy-Schwarz inequality, it is known that

$$\int_0^1 |f(x) - g(x)|dx \leq 2H(f,g)\sqrt{2 - H^2(f,g)}.$$

Under the conditions in Theorem 3.3, we have

$$\frac{1}{m}\sum_{i=1}^{m}\int_0^1 |\hat{f}^i(x) - f^i(x)|dx = O_p(m^{-1/3}). \tag{20}$$

However, π_0 and f_1 are generally unidentifiable without extra conditions. Below we focus on the case $f_0 \equiv 1$. The model is identifiable in this case if there exists an $a_0 \leq 1$ such that $f_1(a_0) = 0$. If f_1 is decreasing, then $f_1(x) = 0$ for $x \in [a_0, 1]$. Suppose $a_0 < 1$. For a sequence $b_m \in (0, 1)$ such that

$$\frac{\int_{b_m}^1 f_1(x)dx}{1 - b_m} = o(1), \quad \frac{m^{-1/3}}{1 - b_m} = o(1), \tag{21}$$

as $m \to +\infty$, we define the refined estimator for $\pi_0(i/m)$ as

$$\breve{\pi}_0(i/m) = \frac{1}{1 - b_m}\int_{b_m}^1 \hat{f}^i(x)dx = \hat{\pi}_0(i/m) + (1 - \hat{\pi}_0(i/m))\frac{\int_{b_m}^1 \hat{f}_1(x)dx}{1 - b_m}.$$

Under (21), we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}|\breve{\pi}_0(i/m) - \pi_0(i/m)| \\
=&\frac{1}{m(1 - b_m)}\sum_{i=1}^{m}\left|\int_{b_m}^1 \hat{f}^i(x)dx - \int_{b_m}^1 f^i(x)dx\right| + o_p(1) \\
\leq&\frac{1}{m(1 - b_m)}\sum_{i=1}^{m}\int_0^1 |\hat{f}^i(x) - f^i(x)|dx + o_p(1) = o_p(1).
\end{aligned} \tag{22}$$

Given the refined estimator $\breve{\pi}_0$, the Lfdr can be estimated by

$$\widehat{\text{Lfdr}}_i(x_i) = \frac{\breve{\pi}_0(i/m)}{\hat{f}^i(x_i)}.$$

As $\hat{\pi}_0, \pi_0 \in \Xi$ and thus are bounded from below, by (20) and (22), it is not hard to show that

$$\frac{1}{m}\sum_{i=1}^{m}\int_0^1 |\widehat{\text{Lfdr}}_i(x) - \text{Lfdr}_i(x)|dx = o_p(1). \tag{23}$$

Moreover, we have the following result which justifies Condition (C3).

COROLLARY 3.4. Suppose $\pi_0 \in \Xi$, $f_0 \equiv 1$, and $f_1 \in \mathcal{H}$. Further assume D_0 in Condition (C1) is continuous at zero and (21) holds. Then Condition (C3) is fulfilled.

REMARK 3.2. Although b_m needs to satisfy (21) theoretically, the rate condition is of little use in selecting b_m in practice. We use a simple heuristic procedure that performs reasonably well in our simulations. To motivate our procedure, we let θ indicate the underlying truth of a randomly selected hypothesis from $\{H_i\}_{i=1}^m$. Then we have

$$P(\theta = 0) = \frac{1}{m}\sum_{i=1}^{m} P(\theta_i = 0) = \frac{1}{m}\sum_{i=1}^{m} \pi_0(i/m) := \bar{\pi}_m.$$

Without knowing the order information, the p-values follow the mixture model $\bar{\pi}_m f_0(x) + (1-\bar{\pi}_m) f_1(x)$. The overall null proportion $\bar{\pi}_m$ can be estimated by classical methods such as those proposed by [41] (in practice, we use the maximum of the two Storey's global null proportion estimates in the `qvalue` package for more conservativeness). Denote the corresponding estimator by $\hat{\pi}$. Also denote $\breve{\pi} = m^{-1}\sum_{i=1}^m \breve{\pi}_0(i/m)$, where $\breve{\pi}_0(i/m) = \hat{\pi}_0(i/m) + \delta(1-\hat{\pi}_0(i/m))$ is the calibrated null probability and δ is the amount of calibration, which is a function of b_m. Then it makes sense to choose $b_m \in [0,1]$ such that the difference $|\breve{\pi} - \hat{\pi}|$ is minimized. This results in the procedure that if the mean of $\hat{\pi}_0(i/m)$'s from the EM algorithm (denote as $\tilde{\pi}$) is greater than the global estimate $\hat{\pi}$, $\breve{\pi}_0(i/m) = \hat{\pi}_0(i/m)$, and if the mean is less than $\hat{\pi}$, then $\breve{\pi}_0(i/m) = \hat{\pi}_0(i/m) + \delta(1-\hat{\pi}_0(i/m))$, where $\delta = (\hat{\pi} - \tilde{\pi})/(1-\tilde{\pi})$.

3.4. *Asymptotic power analysis.* We provide asymptotic power analysis for the proposed method. In particular, we have the following result concerning the asymptotic power of the Lfdr procedure in Section 2.2.

THEOREM 3.5. *Suppose Conditions (C1)-(C3) hold and additionally assume that*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\theta_i = 0\} \to \kappa_0,$$
$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\theta_i = 1, Lfdr_i(x_i) \le \lambda\} \to^p D_2(\lambda),$$

for a continuous function D_2 of λ on [0,1]. Let λ_0 be the largest $\lambda \in [0,1]$ such that $R(\lambda) \le \alpha$ and for any small enough ϵ, $R(\lambda_0 - \epsilon) < \alpha$. Then we

have

$$Power_{Lfdr} := \frac{\sum_{i=1}^m \mathbf{1}\{\theta_i = 1, \widehat{Lfdr}_i(x_i) \le \hat{\lambda}_m\}}{\sum_{i=1}^m \mathbf{1}\{\theta_i = 1\} \vee 1} \to^p \frac{D_2(\lambda_0)}{1-\kappa_0}.$$

Recall that in Section 2.1, we have shown that the step-up procedure has the highest expected number of true positives amongst all α-level FDR rules. This result thus sheds some light on the asymptotic optimal power amongst all α-level FDR rules when the number of hypothesis tests goes to infinity.

REMARK 3.3. Under the two-group mixtue model (1)-(2) with $\pi_{0i} = \pi_0(i/m)$ for some non-decreasing function π_0, we have $m^{-1}\sum_{i=1}^m P(\theta_i = 0) = m^{-1}\sum_{i=1}^m \pi_0(i/m) \to \int_0^1 \pi_0(x)dx$ as monotonic functions are Riemann integrable. Thus $\kappa_0 = \int_0^1 \pi_0(x)dx$. Define $g(x) = \sup\{t \in [0,1] : f_1(t)/f_0(t) \ge x\}$ and $w(\lambda, x) = \frac{\pi_0(x)(1-\lambda)}{(1-\pi_0(x))\lambda}$. Denote by F_1 the distribution function of f_1. Then we have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m P(\theta_i = 1, \mathrm{Lfdr}_i(x_i) \le \lambda) =& \frac{1}{m}\sum_{i=1}^m P(\theta_i = 1)P(\mathrm{Lfdr}_i(x_i) \le \lambda | \theta_i = 1) \\ =& \frac{1}{m}\sum_{i=1}^m (1-\pi_0(i/m))F_1 \circ g \circ w(\lambda, i/m) \\ \to& \int_0^1 (1-\pi_0(x))F_1 \circ g \circ w(\lambda, x)dx,\end{aligned}$$

where "$\circ$" denotes the composition of two functions, and we have used the fact that $F_1 \circ g \circ w$ is monotonic and thus Riemann integrable. So $D_2(\lambda) = \int_0^1 (1-\pi_0(x))F_1 \circ g \circ w(\lambda, x)dx$.

4. Two extensions.

4.1. *Grouped hypotheses with ordering.* Our idea can be extended to the case where the hypotheses can be divided into $d \ge 2$ groups within which there is no explicit ordering but between which there is an ordering. One can simply modify (18) by considering the problem,

$$(\hat{\pi}_{01}^{(t+1)}, \ldots, \hat{\pi}_{0d}^{(t+1)}) = \mathrm{argmin} \sum_{j=1}^m \left\{\tilde{Q}_j^{(t)} - \pi_{s(j)}\right\}^2, \quad (24)$$

[2]This optimization problem can be solved by slightly modifying the PAVA by averaging the estimators within each group.

subject to $0 \leq \pi_1 \leq \cdots \leq \pi_d \leq 1$, where $s(j) \in \{1, 2, \ldots, d\}$ is the group index for the jth hypothesis. A particular example is about using the sign to improve power while controlling the FDR. Consider a two-sided test where the null distribution is symmetric and the test statistic is the absolute value of the symmetric statistic. The sign of the statistic is independent of the p-value under the null. If we have *a priori* belief that among the alternatives, more hypotheses have true positive effect sizes than negative ones or vice versa, then sign could be used to divide the hypotheses into two groups such that $\pi_1 \leq \pi_2$ (or $\pi_1 \geq \pi_2$).

4.2. *Varying alternative distributions.* In model (1), we assume that the success probabilities $\pi_{0i}, i = 1, \ldots, m$ vary with i while F_1 is independent of i. This assumption is reasonable in some applications but it can be restrictive in other cases. We illustrate this point via a simple example described below.

EXAMPLE 4.1. For $1 \leq i \leq m$, let $\{x_{ik}\}_{k=1}^{n_i}$ be n_i observations generated independently from $N(\mu_i, 1)$. Consider the one sided z-test $Z_i = \sqrt{n_i}\bar{x}_i$ with $\bar{x}_i = n_i^{-1} \sum_{k=1}^{n_i} x_{ik}$ for testing

$$H_{i0} : \mu_i = 0 \quad \text{vs} \quad H_{ia} : \mu_i < 0.$$

The p-value is equal to $p_i = \Phi(\sqrt{n_i}\bar{x}_i)$ and the p-value distribution under the alternative hypothesis is given by

$$F_{1i}(x) = \Phi\left(\Phi^{-1}(x) - \sqrt{n_i}\mu_i\right),$$

with the density

$$f_{1i}(x) = \frac{\phi(\Phi^{-1}(x) - \sqrt{n_i}\mu_i)}{\phi(\Phi^{-1}(x))} = \exp\left(\frac{2\sqrt{n_i}\mu_i\Phi^{-1}(x) - n_i\mu_i^2}{2}\right).$$

By prioritizing the hypotheses based on the values of $\sqrt{n_i}\mu_i$, one can expect more discoveries. Suppose

$$n_1\mu_1^2 \leq n_2\mu_2^2 \leq \cdots \leq n_m\mu_m^2.$$ [3]

One can consider the following problem to estimate π and μ_i simultaneously,

$$\underset{\pi \in [0,1], r_m \leq r_{m-1} \leq \cdots \leq r_1 < 0}{\operatorname{argmax}} \sum_{i=1}^{m} \log\left\{\pi + (1-\pi)\exp\left(\frac{2r_i\Phi^{-1}(p_i) - r_i^2}{2}\right)\right\}.$$

This problem can again be solved using the EM algorithm together with the PAVA.

[3] This is the case if $\mu_i = \mu$ and $n_1 \leq n_2 \leq \cdots \leq n_m$.

Generally, if the p-value distribution under the alternative hypothesis, denoted by F_{1i}, is allowed to vary with i, model (1)-(2) is not estimable without extra structural assumptions as we only have one observation that is informative about F_{1i}. On the other hand, if we assume that $F_{1i} := F_{1,i/m}$ which varies smoothly over i, then one can use non-parametric approach to estimate each $F_{1,i/m}$ based on the observations in a neighborhood of i/m. However, this method requires the estimation of m density functions at each iteration, which is computationally expensive for large m. To reduce the computational cost, one can divide the indices into K consecutive bins, say $S_1, S_2, \ldots, S_K$, and assume that the density remains unchanged within each bin. In the M-step, we update f_{1i} via

$$f_{1i}^{(t+1)} = \underset{\tilde{f}_1 \in \mathcal{H}}{\operatorname{argmax}} \sum_{j \in S_i} (1 - Q_j^{(t)}) \log \tilde{f}_1(x_j), \tag{25}$$

for $i = 1, 2, \ldots, K$. For small K, the computation is relatively efficient. We note that this strategy is related to the independent hypothesis weighting proposed in [25; 26], which divides the p-values into several bins and estimate the cumulative distribution function (CDF) of the p-values in each stratum. Our method is different from theirs in the following aspect: the estimated densities will be used in constructing the optimal rejection rule, while in their procedure, the varying CDF is used as an intermediate quantity to determine the thresholds for p-values in each stratum. In other words, the estimated CDFs are not utilized optimally in constructing the rejection rule.

5. Simulation studies.

5.1. *Simulation setup.* We conduct comprehensive simulations to evaluate the finite-sample performance of the proposed method and compare it to competing methods. For simplicity, we directly simulate z-values for m=10,000 hypotheses. All simulations are replicated 100 times except for the global null, where the results are based on 2,000 Monte Carlo replicates. We simulate different combinations of signal density (the percentage of alternative) and signal strength (the effect size of alternative) since these are two main factors affecting the power of multiple testing procedures. We first generate the hypothesis-specific null probability (π_{0i}), upon which the truth, i.e., null or alternative, is simulated. Afterwards, we generate z-values based on the truth of the hypothesis. We first use π_{0i} as the auxiliary covariate. Later, we will study the effect of using noisy π_{0i} as auxiliary covariate. Three scenarios, representing weakly, moderately and highly informative auxiliary information, are simulated based on the distribution of π_{0i} (Figure 1(a)),

where the informativeness of the auxiliary covariate is determined based on its ability to separate alternatives from nulls (Figure 1(b)). In the weakly informative scenario, we make π_{0i}'s similar for all hypotheses by simulating π_{0i}'s from a highly concentrated normal distribution (truncated on the unit interval $[0, 1]$)

$$\pi_{0i} \sim N_C(\mu_w, 0.005^2).$$

In the moderately informative scenario, we allow π_{0i} to vary across hypotheses with moderate variability. This is achieved by simulating π_{0i}'s from a beta distribution

$$\pi_{0i} \sim \text{Beta}(a, b).$$

In the highly informative scenario, π_{0i}'s are simulated from a mixture of a truncated normal and a highly concentrated truncated normal distribution

$$\pi_{0i} \sim \pi_h N_C(\mu_{h1}, \sigma_{h1}^2) + (1 - \pi_h) N_C(\mu_{h2}, 0.005^2),$$

which represents two groups of hypotheses with strikingly different probabilities of being null. Since the expected alternative proportion is $\sum_{i=1}^{m}(1 - \pi_{0i})/m$, we adjust the parameters $\mu_w, a, b, \pi_h, \mu_{h1}, \sigma_{h1}^2$ and μ_{h2} to achieve approximately 5%, 10% and 20% signal density level. Figure 1(a) shows the distribution of π_{0i} for the three scenarios. Based on π_{0i}, the underlying truth θ_i is simulated from

$$\theta_i \sim \text{Bernoulli}(1 - \pi_{0i}).$$

Figure 1(b) displays the distribution of π_{0i} for $\theta_i = 1$ and $\theta_i = 0$ from one simulated dataset. As the difference in π_{0i} between H_1 and H_0 gets larger, the auxiliary covariate becomes more informative. Finally, we simulate independent z-values using

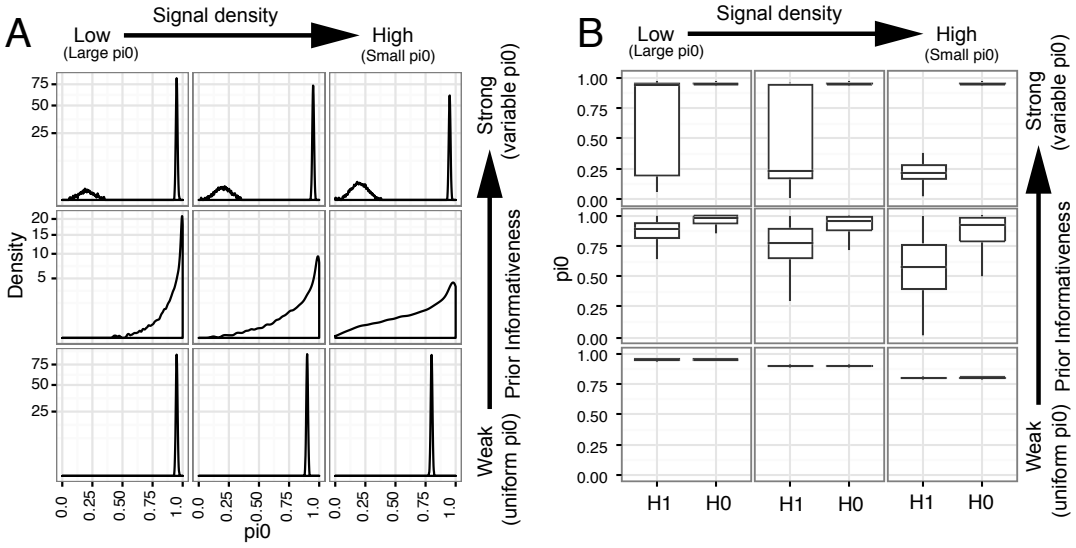
$$z_i \sim N(k_s \theta_i, 1),$$

where k_s controls the signal strength and k_s=2, 2.5 and 3 are chosen to represent weak, moderate and strong signal, respectively. We convert z-values to p-values using the formula $p_i = 1 - \Phi(z_i)$. The proposed method accepts p-values and π_{0i}s as input. The specific parameter values mentioned above could be found in https://github.com/jchen1981/OrderShapeEM.

To examine the robustness of the proposed method, we vary the simulation setting in different ways. Specifically, we investigate:

1. *Skewed alternative distribution.* Instead of simulating normal z-values for the alternative group, we simulate z-values from a non-central gamma distribution with the shape parameter $k = 2$. The scale and non-centrality parameters of the non-central gamma distribution are

Fig 1: Simulation Strategy. (a) The distribution of probabilities of being null ($\pi_{0i}, i = 1, \ldots, m$) for three scenarios representing weakly, moderately and highly informative auxiliary information (from bottom to top). Different levels of signal density are simulated. (b) Distribution of the realized π_{0i} for alternatives and nulls from one simulated dataset.



chosen to match the mean and variance of the normal distribution for the alternative group under the basic setting.

2. *Correlated hypotheses.* Our theory allows certain forms of dependence. We then simulate correlated z-values, which are drawn from a multivariate normal distribution with a block correlation structure. The order of π_{0i} is random with respect to the block structure. Specifically, we divide the $10,000$ hypotheses into 100 blocks and each block is further divided into two sub-blocks of equal size. Within each sub-block, there is a constant positive correlation (ρ=0.5). Between the sub-blocks in the same block, there is a constant negative correlation (ρ=-0.5). Hypotheses in different blocks are independent. We use $p = 8$ to illustrate. The correlation matrix is

$$\begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ 0.5 & 1 & 0.5 & 0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & 1 & 0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & 0.5 & 1 & -0.5 & -0.5 & -0.5 & -0.5 \\ -0.5 & -0.5 & -0.5 & -0.5 & 1 & 0.5 & 0.5 & 0.5 \\ -0.5 & -0.5 & -0.5 & -0.5 & 0.5 & 1 & 0.5 & 0.5 \\ -0.5 & -0.5 & -0.5 & -0.5 & 0.5 & 0.5 & 1 & 0.5 \\ -0.5 & -0.5 & -0.5 & -0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}.$$

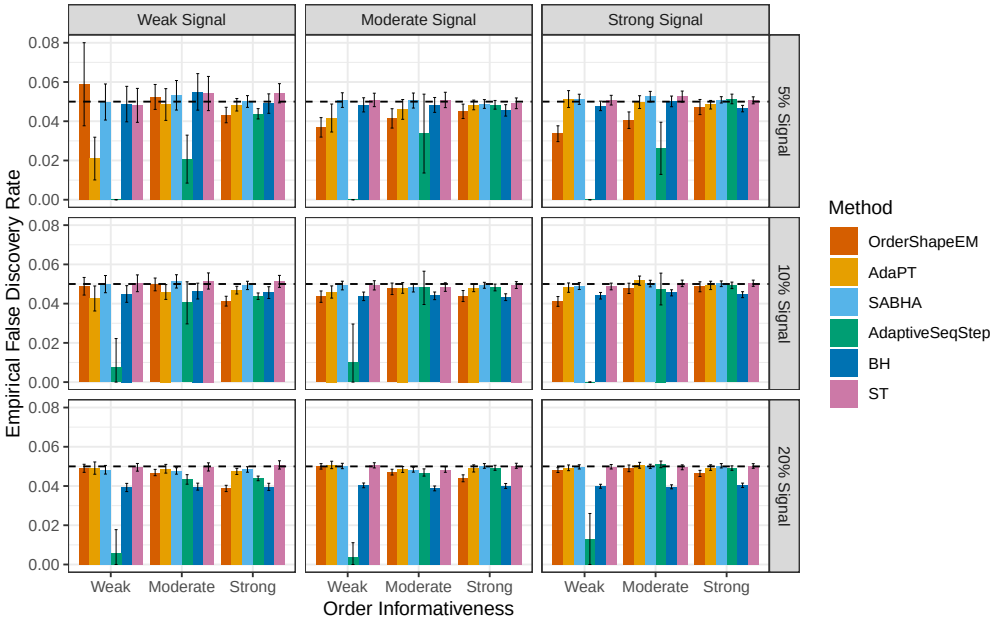
3. *Noisy auxiliary information.* In practice, the auxiliary data can be very noisy. To examine the effect of noisy auxiliary information, we shuffle half or all the π_{0i}, representing moderately and completely noisy order.
4. *A smaller number of alternative hypotheses and a global null.* It is interesting to study the robustness of the proposed method under an even more sparse signal. We thus simulate 1% alternatives out of 10,000 features. We also study the error control under a global null, where all the hypotheses are nulls. Under the global null, We increased the number of Monte Carlo simulations to 2,000 times to have a more accurate estimate of the FDR.
5. *Varying f_1 across alternative hypotheses.* We consider the case where among the alternative hypotheses, the most promising 20% hypotheses (i.e., those with the lowest prior order) follow Unif$(0, 0.02)$ and the remaining p-values are derived from the z-values (see the setting of Figure 2).
6. *Varying f_0 across null hypotheses.* Similar to the case of varying f_1, we sample the p-values of 20% of the null hypotheses with the highest prior order from Unif$(0.5, 1)$, which mimics the composite null situations. The remaining p-values are derived from the z-values as above.

We compare the proposed method (OrderShapeEM) with classical multiple testing methods that do not utilize external covariates (BH and ST) and recent multiple testing procedures that exploit auxiliary information (AdaPT, SABHA, AdaptiveSeqStep). Detailed descriptions of these methods are provided in the appendix. The FDP estimate of AdaPT involves a finite-sample correction term +1 in the numerator. The +1 term yields a conservative procedure and could lose power when the signal density is low. To study the effect of the correction term, we also compared to AdaPT+, where we removed the correction term +1 in the numerator. However, we observed a significant FDR inflation when the signal density is low, see Figure 14 in the Appendix. We thus compared to AdaPT procedure with correction term throughout the simulations.

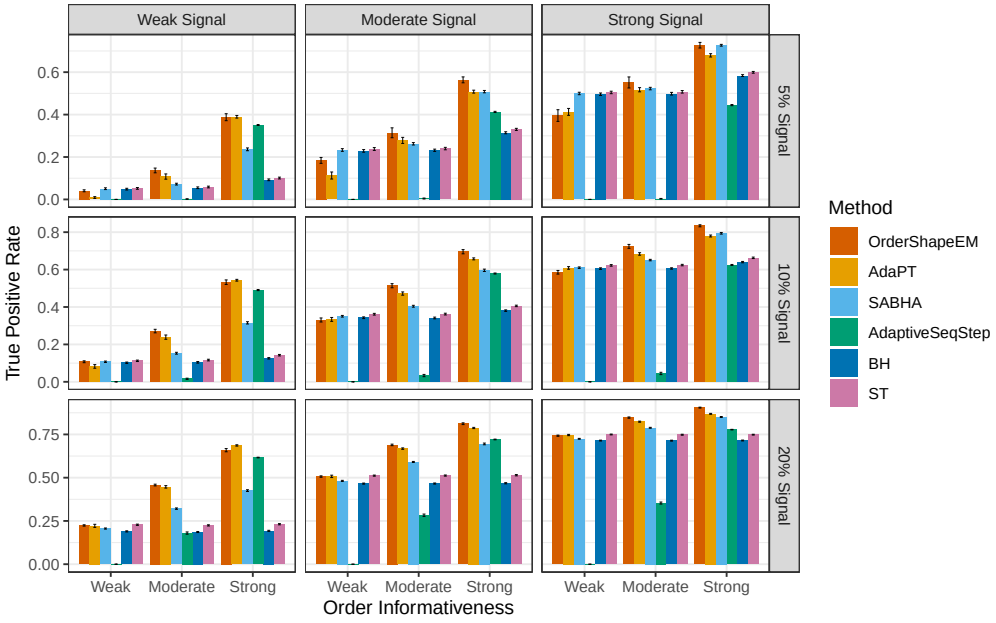
5.2. *Simulation results.* We first discuss the simulation results of *Normal alternative distribution*. In Figure 2, we present FDR control and power comparison with different methods when z-values under the null hypothesis follow $N(0,1)$ and z-values under the alternative hypothesis follow a normal distribution. In Figure 2(a), the dashed line indicates the pre-specified FDR control level 0.05 and the error bars represent empirical 95% confidence intervals. We observe that all procedures control the FDR sufficiently well across settings and no FDR inflation has been observed. Adaptive SeqStep is conservative most of the time especially when the signal is sparse and the auxiliary information is weak or moderate. AdaPT is conservative under sparse signal and weak auxiliary information. The proposed procedure OrderShapeEM generally controls the FDR at the target level with some conservativeness under some settings. As expected, ST procedure controls the FDR at the target level while BH procedure is more conservative under dense signal. In Figure 2(b), we observe that OrderShapeEM is overall the most powerful when the auxiliary information is not weak. When the auxiliary information is weak and the signal is sparse, OrderShapeEM could be less powerful than BH/ST. Close competitors are AdaPT and SABHA. However, AdaPT is significantly less powerful when the signal is sparse and the auxiliary information is weak. AdaPT is also computationally more intensive than the other methods. SABHA performs well when the signal is strong but becomes much less powerful than OrderShapeEM and AdaPT as the signal weakens. Adaptive SeqStep has good power for dense signal and moderate to strong auxiliary information. However, it is powerless when auxiliary information is weak. If auxiliary information is weak, SABHA, ST and BH have similar power, while Adaptive SeqStep has little power. Under this scenario, incorporating auxiliary information does not help much.

Fig 2: Performance under normal alternative distribution.

(a) FDR control



(b) power comparison



All methods become more powerful with the increase of signal density and signal strength.

Results for the other settings are included in the Appendix. Briefly, results based on *Skewed alternative distribution* and *Noisy auxiliary information* (Figures 5-7) have similar patterns. OrderShapeEM has adequate FDR control under *a smaller number of alternative hypotheses* and *a global null* (Figures 9-10). Under *varying* f_1, we observe slight inflation for the proposed method under some scenarios especially when the signal density is low (Figure 11). On the other hand, under *varying* f_0, the proposed method suffers from severe power deterioration (Figure 12). Although our method offers asymptotic FDR control and we have observed adequate FDR control at $m = 10,000$, it is interesting to study the performance at a smaller m. We thus tried $m = 500, 1000$ and 2000 under the same setup as in Figure 2 with a medium signal density. The results are summarized in Figure 13. We observe small FDR inflation for these sample sizes and the inflation increased with smaller sample sizes, particularly for a weaker signal and less informative prior. We thus recommend using our method when m is not small (e.g. $m > 1000$). Since our theory depends on the independence between hypotheses, we also study the robustness of OrderShapeEM to correlated hypotheses. The simulation setup is described in Section 5.1. From Figure 6(a), we observe that there is more variability across the replications indicated by a wider confidence interval of the empirical FDR and power. OrderShapeEM is more conservative under the correlated hypotheses. With respect to power (Figure 6(b)), when the signal is strong, it could be less powerful than BH/ST. However, when the signal becomes weaker and the auxiliary data is informative, OrderShapeEM is more powerful than BH/ST but is less powerful than AdaPT and SABHA.

Since OrderShapeEM consists of two components: (1) the estimation of the mixing probabilities and the alternative distribution using PAVA, and (2) the optimal rejection rule, it is interesting to study the contribution of each component. We thus apply the SABHA rejection rule using the mixing probabilities from OrderShapeEM (denoted as "SABHA+"), and compare to SABHA and OrderShapeEM. In Figure 8 (see "Additional simulation results"), we observe that SABHA and SABHA+ have a similar performance across settings, while the OrderShapeEM, which uses the optimal rejection rule, is much more powerful than SABHA+ and SABHA under weak signal. The results suggest that the performance improvement of OrderShapeEM is largely contributed by the proposed optimal rejection rule. Therefore, the power loss of SABHA under weak signal is likely due to the inefficiency of its rejection rule.

6. Data Analysis. We illustrate the application of our method by analyzing data from publicly available genome-wide association studies (GWAS). We use datasets from two large-scale GWAS of coronary artery disease (CAD) in different populations (CARDIoGRAM and C4D). CARDIoGRAM is a meta-analysis of 14 CAD genome-wide association studies, comprising $22,233$ cases and $64,762$ controls of European descent [39]. The study includes 2.3 million single nucleotide polymorphisms (SNP). In each of the 14 studies and for each SNP, a logistic regression of CAD status was performed on the number of copies of one allele, along with suitable controlling covariates. C4D is a meta-analysis of 5 heart disease genome-wide association studies, totaling $15,420$ CAD cases and $15,062$ controls [8]. The samples did not overlap those from CARDIoGRAM. The analysis steps were similar to CARDIoGRAM. A total of $514,178$ common SNPs were tested in both the CARDIoGRAM and C4D association analyses. Dataset can be downloaded from `http://www.cardiogramplusc4d.org`. Available data comprise of a bivariate p-value sequence (x_{1i}, x_{2i}), where x_{1i} represents p-values from the CARDIoGRAM dataset and x_{2i} represents p-values from the C4D dataset, $i = 1, \ldots, 514,178$.

We are interested in identifying SNPs that are associated with CAD. Due to the shared genetic polymorphisms between populations, information contained in x_{i1} can be helpful in the association analysis of x_{2i} and vice versa. We thus performed two separate analyses, where we conducted FDR control on x_{1i} and x_{2i} respectively, using x_{2i} and x_{i1} as the auxiliary covariate.

In the analysis, we compare the proposed OrderShapeEM, robust method that incorporates auxiliary information (SABHA) and method that does not incorporate auxiliary information (ST). As BH was outperformed by ST and Adaptive SeqStep by SABHA, we only included ST and SABHA in the comparison. AdaPT was not able to complete the analysis within 24 hours and was not included either. The results are summarized in Figure 3. From Figure 3(a), we observe that at the same FDR level, the proposed OrderShapeEM made significantly more discoveries than SABHA and ST. SABHA procedure, which incorporates the auxiliary information, picked up more SNPs than the ST procedure. The performance of OrderShapeEM is consistent with the weak signal scenario, where a significant increase in power has been observed (Figure 2(b)). Due to disease heterogeneity, signals in the genetic association studies are usually very weak. Thus, it can be extremely helpful to incorporate auxiliary information to improve power. The power difference becomes even larger at higher target FDR level. Figure 3(b) shows similar patterns.

To further examine the identified SNPs based on different methods, Figure

Fig 3: Comparison of the number of discoveries at different pre-specified FDR level (left panels) as well as the estimates of π_0 (middle panels) and f_1 (right panels).

(a) Analysis of C4D data with CARDIoGRAM data as auxiliary information; (b) Analysis of CARDIoGRAM data with C4D data as auxiliary information.

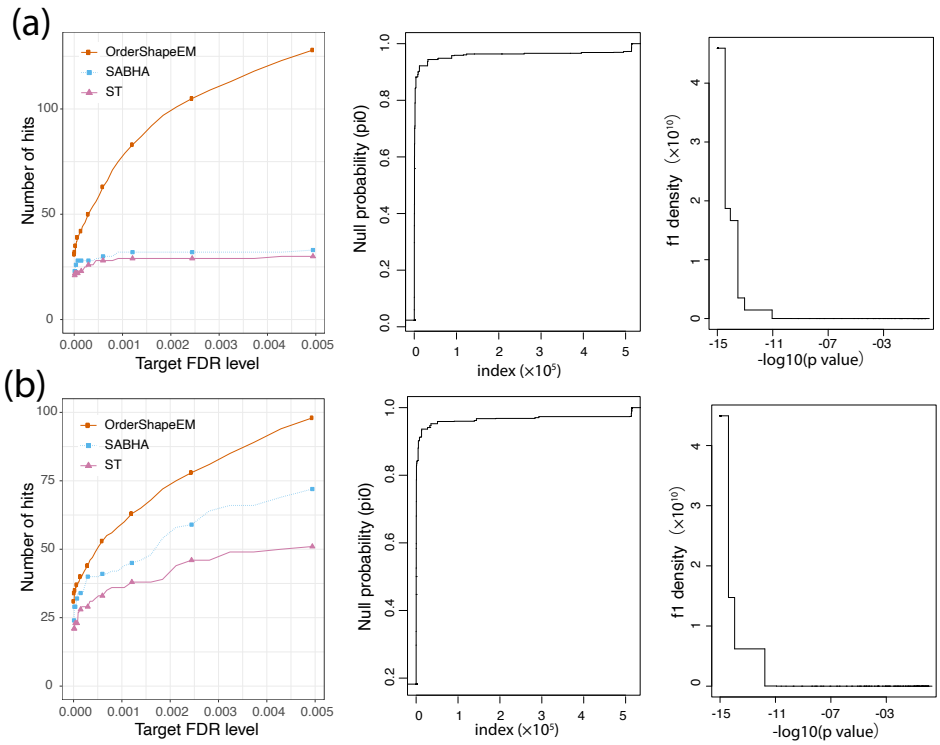
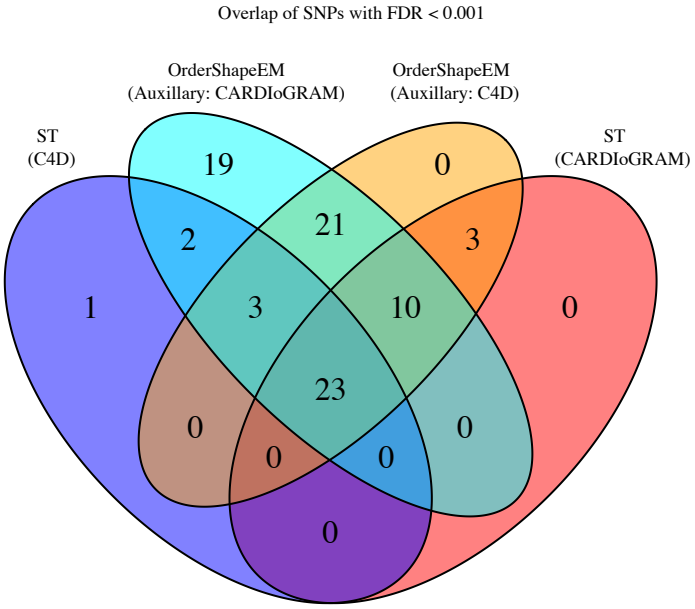


Fig 4: Venn diagram showing the overlap of significant SNPs (FDR < 0.001) between methods using or not using auxiliary information. Left to right: ST procedure on C4D data; OrderShape EM on C4D data with CARDIoGRAM data as auxiliary; OrderShapeEM on CARDIoGRAM data with C4D data as auxiliary; and ST procedure on CARDIoGRAM data.



4 shows the overlap of significant SNPs via the Venn diagram at FDR level 0.001. We observe that there was a significant overlap of associated SNPs between the two datasets, indicating a shared genetic architecture between the two populations. By using auxiliary information, OrderShapeEM recovered almost all the SNPs by ST procedure, in addition to many other SNPs that were missed by the ST procedure. Interestingly, for the $19 + 21 = 40$ SNPs that were identified by OrderShapeEM only, most of them were located in genes that had been reported being associated with phenotypes or diseases related to the cardiovascular or metabolic system. It is well known that metabolic disorders such as high blood cholesterol and triglyceride levels are risk factors for CAD.

7. Summary and discussions. We have developed a covariate-adjusted multiple testing procedure based on the Lfdr and shown that the oracle procedure is optimal in the sense of maximizing the ETP for a given value of mFDR. We propose an adaptive procedure to estimate the prior probabilities of being null that vary across different hypotheses and the distribution function of the p-values under the alternative hypothesis. Our estimation procedure is built on the isotonic regression which is tuning parameter free and computationally fast. We prove that the proposed method provides asymptotic FDR control when relevant consistent estimates are available. We obtain some consistency results for the estimates of the prior probabilities of being null and the alternative density under shape restrictions. In finite samples, the proposed method outperforms several existing approaches that exploit auxiliary information to boost power in multiple testing. The gain in efficiency of the proposed procedure is due to the fact that we incorporate both the auxiliary information and the information across p-values in an optimal way.

Our method has a competitive edge over competing methods when the signal is weak and the auxiliary information is moderate/strong, a practically important setting where power improvement is critical and possible with the availability of informative prior. However, when the auxiliary information is weak, our procedure could be less powerful than the BH/ST procedure. The power loss is more severe under strong and sparse signals. To remedy the power loss under these unfavorable conditions, we recommend testing the informativeness of the prior order information before the application of our method using, for example, the testing method from [24]. We could also examine the $\hat{\pi}_0$ plot after running our algorithm. If $\hat{\pi}_0$'s lack variability, which indicates the auxiliary information is very weak, our method could be less powerful than BH/ST and we advise against using it.

Our method is also robust across settings with a very moderate FDR inflation under small feature sizes. However, there are some special cases where our approach does not work well due to the violation of assumptions. In the varying alternative scenario, as suggested by one of the reviewers, we did observe some FDR inflation. We found this only happens when the order information has inconsistent effects on the π_0 and f_1 (i.e., the more likely the alternative hypothesis, the smaller the effect size). We did not find any FDR inflation if the order information has consistent effects (i.e., the more likely the alternative hypothesis, the larger the effect size). We believe such inconsistent effects may be uncommon in practice. In the varying null scenario, we observed a severe deterioration of the power of our method and it has virtually no power when the signal is sparse. This is somewhat expected since our approach assumes a uniformly distributed null p-value. Therefore, we should examine the p-value distribution before applying our method. We advise against using our method if we see a substantial deviation from the uniform assumption based on the right half of the p-value distribution.

There are several future research directions. For example, it is desirable to extend our method to incorporate other forms of structural information such as group structure, spatial structure or tree/hierarchical structure. Also, the proposed method is marginal based and it may no longer be optimal in the presence of correlations. We leave these interesting topics for future research.

Acknowledgements. The authors would like to thank the Associate Editor and the reviewers for their constructive comments and helpful suggestions, which substantially improved the paper. Data on coronary artery disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.cardiogramplusc4d.org. Cao acknowledges partial support from NIH 2UL1TR001427-5, Zhang acknowledges partial support from NSF DMS-1830392 and NSF DMS-1811747 and Chen acknowledges support from Mayo Clinic Center for Individualized Medicine.

8. Appendix. We provide proofs of all mathematical claims and additional simulation results.

Proof of Proposition 2.2. The following lemma can be proved using similar arguments as in the proof of the (weak) Glivenko-Cantelli Theorem (see

e.g. [49]) and we omit the details here. Define

$$D_{m,0}(\lambda) := \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\mathrm{Lfdr}_i(x_i) \le \lambda\},$$
$$D_{m,1}(\lambda) := \frac{1}{m}\sum_{i=1}^{m} \mathrm{Lfdr}_i(x_i)\mathbf{1}\{\mathrm{Lfdr}_i(x_i) \le \lambda\},$$
$$R_m(\lambda) = D_{m,1}(\lambda)/D_{m,0}(\lambda).$$

LEMMA 8.1. *Under Condition (C1), we have*

$$\sup_{\lambda\in[0,1]} |D_{m,0}(\lambda) - D_0(\lambda)| \to^p 0,$$
$$\sup_{\lambda\in[0,1]} |D_{m,1}(\lambda) - D_1(\lambda)| \to^p 0,$$
$$\sup_{\lambda\in[0,1]} |V_m(\lambda)/m - D_1(\lambda)| \to^p 0.$$

LEMMA 8.2. *Under Conditions (C1)-(C2),*

$$\sup_{x\ge\lambda_\infty} |R_m(x) - R(x)| \to^p 0,$$

and

$$\sup_{x\ge\lambda_\infty} |V_m(x)/D_{m,0}(x) - R(x)| \to^p 0.$$

PROOF OF LEMMA 8.2. By the monotonicity of D_0, $\min_{x\ge\lambda_\infty} D_0(x) = D_0(\lambda_\infty) > 0$ as $D_1(\lambda_\infty)/D_0(\lambda_\infty) < \alpha$. Then we have

$$\begin{aligned}
&\left|\frac{D_{m,1}(x)}{D_{m,0}(x)} - \frac{D_1(x)}{D_0(x)}\right| \\
=&\left|\frac{(D_{m,1}(x) - D_1(x))D_0(x) - D_1(x)(D_{m,0}(x) - D_0(x))}{D_0(x)D_{m,0}(x)}\right| \\
\le&\frac{D_0(1)|D_{m,1}(x) - D_1(x)| + D_1(1)|D_{m,0}(x) - D_0(x)|}{D_0(\lambda_\infty)\{D_0(x) - \sup_{\lambda\ge\lambda_\infty}|D_{m,0}(\lambda) - D_0(\lambda)|\}} \\
\le&\frac{D_0(1)\sup_{\lambda\ge\lambda_\infty}|D_{m,1}(\lambda) - D_1(\lambda)| + D_1(1)\sup_{\lambda\ge\lambda_\infty}|D_{m,0}(\lambda) - D_0(\lambda)|}{D_0(\lambda_\infty)\{D_0(\lambda_\infty) - \sup_{\lambda\ge\lambda_\infty}|D_{m,0}(\lambda) - D_0(\lambda)|\}} \to^p 0
\end{aligned}$$

uniformly for any $x \ge \lambda_\infty$. Similar argument shows the other result. □

PROOF. Set $e = \alpha - R(\lambda_\infty)$. By Lemma 8.2,

$$P(|R_m(\lambda_\infty) - R(\lambda_\infty)| < e/2) \to 1,$$

which implies that $P(R_m(\lambda_\infty) < \alpha) \to 1$. Thus $P(\lambda_m \geq \lambda_\infty) \to 1$ by the definition of λ_m. Then we have

$$\begin{aligned}
&\{R_m(\lambda_m) - V_m(\lambda_m)/D_{m,0}(\lambda_m)\} \\
\geq &\inf_{\lambda \geq \lambda_\infty} \{R_m(\lambda) - V_m(\lambda)/D_{m,0}(\lambda)\} \\
= &\inf_{\lambda \geq \lambda_\infty} \{R_m(\lambda) - R(\lambda) + R(\lambda) - V_m(\lambda)/D_{m,0}(\lambda)\} = o_p(1)
\end{aligned}$$

by Lemma 8.2. As $R_m(\lambda_m) \leq \alpha$, this implies that

$$V_m(\lambda_m)/\{D_{m,0}(\lambda_m) \vee 1\} \leq V_m(\lambda_m)/D_{m,0}(\lambda_m) \leq \alpha + o_p(1).$$

As $V_m(\lambda_m)/\{D_{m,0}(\lambda_m) \vee 1\} \leq 1$, by Lemma 8.3 below, we obtain

$$\limsup_{m \to +\infty} \mathrm{FDR}_m(\lambda_m) = \limsup_{m \to +\infty} E[V_m(\lambda_m)/\{D_{m,0}(\lambda_m) \vee 1\}] \leq \alpha.$$

□

LEMMA 8.3. *Consider the random sequence* $\{(X_m, Y_m)\}_m$. *Suppose* $X_m \leq C_0$ *and* $X_m \leq \alpha + Y_m$, *where* $Y_m = o_p(1)$ *and* C_0 *is some constant. Then we have*

$$\limsup_m E[X_m] \leq \alpha.$$

PROOF OF LEMMA 8.3. Note that here exists a subsequence X_{m_k} such that $\limsup_m E[X_m] = \lim_k E[X_{m_k}]$. Along this subsequence, we can pick a further subsequence $Y_{m_{k_j}}$ such that $Y_{m_{k_j}} \to^{a.s.} 0$. Thus with probability one,

$$\limsup_j X_{m_{k_j}} \leq \limsup_j Y_{m_{k_j}} + \alpha = \alpha.$$

As $X_{m_{k_j}} \leq C_0$, by Fatou's Lemma,

$$\limsup_m E[X_m] = \limsup_j E[X_{m_{k_j}}] \leq E[\limsup_j X_{m_{k_j}}] \leq \alpha.$$

□

Proof of Theorem 2.3. Define

$$\begin{aligned}
\widehat{D}_{m,0}(\lambda) &:= \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i) \leq \lambda\}, \\
\widehat{D}_{m,1}(\lambda) &:= \frac{1}{m} \sum_{i=1}^m \widehat{\mathrm{Lfdr}}_i(x_i) \mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i) \leq \lambda\}, \\
\widehat{R}_m(\lambda) &= \widehat{D}_{m,1}(\lambda)/\widehat{D}_{m,0}(\lambda).
\end{aligned}$$

LEMMA 8.4. Under Conditions (C1) and (C3), we have

$$\sup_{\lambda\geq\lambda_\infty}\left|\widehat{D}_{m,0}(\lambda)-D_0(\lambda)\right|\to^p 0, \tag{26}$$

$$\sup_{\lambda\geq\lambda_\infty}\left|\widehat{D}_{m,1}(\lambda)-D_1(\lambda)\right|\to^p 0. \tag{27}$$

PROOF OF LEMMA 8.4. We only prove (27) as the proof for (26) is similar. In view of Lemma 8.1, we only need to show that

$$\sup_{\lambda\geq\lambda_\infty}\left|\widehat{D}_{m,1}(\lambda)-D_{m,1}(\lambda)\right|\to^p 0.$$

To this end, we note that

$$\begin{aligned}
&\sup_{\lambda\geq\lambda_\infty}\left|\widehat{D}_{m,1}(\lambda)-D_{m,1}(\lambda)\right| \\
\leq&\sup_{\lambda\geq\lambda_\infty}\left|\frac{1}{m}\sum_{i=1}^m\widehat{\mathrm{Lfdr}}_i(x_i)\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}-\frac{1}{m}\sum_{i=1}^m\mathrm{Lfdr}_i(x_i)\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}\right| \\
&+\sup_{\lambda\geq\lambda_\infty}\left|\frac{1}{m}\sum_{i=1}^m\mathrm{Lfdr}_i(x_i)\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}-\frac{1}{m}\sum_{i=1}^m\mathrm{Lfdr}_i(x_i)\mathbf{1}\{\mathrm{Lfdr}_i(x_i)\leq\lambda\}\right| \\
\leq&m^{-1}\sum_{i=1}^m|\widehat{\mathrm{Lfdr}}_i(x_i)-\mathrm{Lfdr}_i(x_i)|+\sup_{\lambda\geq\lambda_\infty}\frac{1}{m}\sum_{i=1}^m|\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}-\mathbf{1}\{\mathrm{Lfdr}_i(x_i)\leq\lambda\}|,
\end{aligned}$$

where the first term in the last line converges to zero in probability by Condition (C3). To deal with the second term, notice that for any $0<\epsilon<\lambda_\infty/2$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m|\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}-\mathbf{1}\{\mathrm{Lfdr}_i(x_i)\leq\lambda\}| \\
=&\frac{1}{m}\sum_{i=1}^m\left[\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda,\mathrm{Lfdr}_i(x_i)>\lambda\}+\mathbf{1}\{\mathrm{Lfdr}_i(x_i)\leq\lambda,\widehat{\mathrm{Lfdr}}_i(x_i)>\lambda\}\right] \\
=&\frac{1}{m}\sum_{i=1}^m\left[\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda,\lambda+\epsilon\geq\mathrm{Lfdr}_i(x_i)>\lambda\}+\mathbf{1}\{\lambda-\epsilon<\mathrm{Lfdr}_i(x_i)\leq\lambda,\widehat{\mathrm{Lfdr}}_i(x_i)>\lambda\}\right] \\
&+\frac{1}{m}\sum_{i=1}^m\left[\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda,\mathrm{Lfdr}_i(x_i)>\lambda+\epsilon\}+\mathbf{1}\{\mathrm{Lfdr}_i(x_i)\leq\lambda-\epsilon,\widehat{\mathrm{Lfdr}}_i(x_i)>\lambda\}\right] \\
\leq&\frac{1}{m}\sum_{i=1}^m\mathbf{1}\{\lambda-\epsilon<\mathrm{Lfdr}_i(x_i)\leq\lambda+\epsilon\}+\frac{1}{m\epsilon}\sum_{i=1}^m|\widehat{\mathrm{Lfdr}}_i(x_i)-\mathrm{Lfdr}_i(x_i)|.
\end{aligned}$$

Together with Lemma 8.1 and Condition (C3), we obtain for any $0 < \epsilon < \lambda_\infty/2$,

$$\begin{aligned}
J &:= \sup_{\lambda \geq \lambda_\infty} \frac{1}{m} \sum_{i=1}^{m} |\mathbf{1}\{\widehat{\mathrm{Lfdr}}_i(x_i) \leq \lambda\} - \mathbf{1}\{\mathrm{Lfdr}_i(x_i) \leq \lambda\}| \\
&\leq \sup_{\lambda \geq \lambda_\infty} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\lambda - \epsilon < \mathrm{Lfdr}_i(x_i) \leq \lambda + \epsilon\} + \frac{1}{m\epsilon} \sum_{i=1}^{m} |\widehat{\mathrm{Lfdr}}_i(x_i) - \mathrm{Lfdr}_i(x_i)| \\
&\leq \sup_{\lambda \geq \lambda_\infty} |D_0(\lambda + \epsilon) - D_0(\lambda - \epsilon)| + 2 \sup_{\lambda \in [0,1]} |D_{m,0}(\lambda) - D_0(\lambda)| \\
&\quad + \frac{1}{m\epsilon} \sum_{i=1}^{m} |\widehat{\mathrm{Lfdr}}_i(x_i) - \mathrm{Lfdr}_i(x_i)| \\
&\leq \sup_{\lambda \geq \lambda_\infty} |D_0(\lambda + \epsilon) - D_0(\lambda - \epsilon)| + o_p(1).
\end{aligned}$$

As ϵ can be arbitrarily small, $\sup_{\lambda \geq \lambda_\infty} |D_0(\lambda + \epsilon) - D_0(\lambda - \epsilon)|$ can be made small due to the (uniform) continuity of D_0. Therefore, $J = o_p(1)$ and thus (27) holds. □

Proof of Theorem 2.3. Using similar arguments as in the proof of Lemma 8.2, we have

$$\sup_{\lambda \geq \lambda_\infty} \left| \widehat{R}_m(\lambda) - R(\lambda) \right| \to^p 0. \tag{28}$$

Following the proof of Proposition 2.2, we set $e = a - R(\lambda_\infty)$. Then we have

$$|\widehat{R}_m(\lambda_\infty) - R(\lambda_\infty)| \leq \sup_{\lambda \geq \lambda_\infty} \left| \widehat{R}_m(\lambda) - R(\lambda) \right| < e/2,$$

with probability tending to one, which suggests that $P(\widehat{R}_m(\lambda_\infty) < \alpha) \to 1$. Thus $P(\hat{\lambda}_m \geq \lambda_\infty) \to 1$ by the definition of $\hat{\lambda}_m$. Then on the event $\{\hat{\lambda}_m \geq \lambda_\infty\}$, we have

$$\begin{aligned}
&|V_m(\hat{\lambda}_m)/D_{m,0}(\hat{\lambda}_m) - \widehat{R}_m(\hat{\lambda}_m)| \\
&\leq \sup_{\lambda \geq \lambda_\infty} \left| \widehat{R}_m(\lambda) - V_m(\lambda)/D_{m,0}(\lambda) \right| \\
&= \sup_{\lambda \geq \lambda_\infty} \left| \widehat{R}_m(\lambda) - R(\lambda) \right| + \sup_{\lambda \geq \lambda_\infty} |R(\lambda) - V_m(\lambda)/D_{m,0}(\lambda)| = o_p(1),
\end{aligned}$$

by Lemma 8.2 and (28). As $\widehat{R}_m(\hat{\lambda}_m) \leq \alpha$, this implies that

$$V_m(\hat{\lambda}_m)/\{D_{m,0}(\hat{\lambda}_m) \vee 1\} \leq \alpha + o_p(1).$$

As $V_m(\hat{\lambda}_m)/\{D_{m,0}(\hat{\lambda}_m) \vee 1\} \leq 1$, by Lemma 8.3, we obtain

$$\limsup_{m\to+\infty} \mathrm{FDR}_m(\hat{\lambda}_m) = \limsup_{m\to+\infty} E[V_m(\hat{\lambda}_m)/\{D_{m,0}(\hat{\lambda}_m) \vee 1\}] \leq \alpha.$$

□

Proof of Theorem 3.1.

PROOF. As $P(f_0(x_i) = f_1(x_i)) = 0$, without loss of generality, we shall assume that $f_0(x_i) \neq f_1(x_i)$ for all i. Recall that $\phi(x,a) = af_0(x) + (1-a)f_1(x)$ and define

$$\rho(x,a) = \frac{\partial \log \phi(x,a)}{\partial a} = \frac{f_0(x) - f_1(x)}{\phi(x,a)},$$

which is nonincreasing in a. As $f_0(x_i) \neq f_1(x_i)$ for all i, it is straightforward to see that for any $1 \leq k \leq l \leq m$, $\sum_{i=k}^{l} \log \phi(x_i,a)$ is a strictly concave function for $0 \leq a \leq 1$. Let

$$\hat{a}_{kl} = \operatorname*{argmax}_{a\in[0,1]} \sum_{i=k}^{l} \log \phi(x_i,a)$$

be the unique maximizer. According to Theorem 3.1 of [37], we have

$$\hat{\pi}_{0i} = \max_{1\leq k\leq i} \min_{i\leq l\leq m} \hat{a}_{kl}.$$

Our goal is to show that the event

$$\hat{\pi}_{0,i_0} = \max_{1\leq k\leq i_0} \min_{i_0\leq l\leq m} \hat{a}_{kl} = \min_{i_0\leq l\leq m} \max_{1\leq k\leq i_0} \hat{a}_{kl} < \pi_{0,i_0} + \epsilon$$

has probability tending to one.

To this end, we let $B = B(i_0, N)$ be the event that $\sum_{i=k}^{i_0+N} \rho(x_i,a) = 0$ has a unique root $0 \leq \hat{a}_{k,i_0+N} < 1$ for all $1 \leq k \leq i_0$, and note that

$$P\left(\max_{1\leq k\leq i_0} \hat{a}_{k,i_0+N} < \pi_{0,i_0} + \epsilon\right) \leq P\left(\hat{\pi}_{0,i_0} < \pi_{0,i_0} + \epsilon\right).$$

On B, we have $\max_{1\leq k\leq i_0} \hat{a}_{k,i_0+N} < \pi_{0,i_0} + \epsilon$ if and only if

$$\max_{1\leq k\leq i_0} \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i_0} + \epsilon) < 0. \tag{29}$$

We see that (29) is equivalent to

$$\sum_{i=k}^{i_0+N} \{\rho(x_i, \pi_{0,i_0}) - \rho(x_i, \pi_{0,i_0} + \epsilon)\} > \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i_0}), \tag{30}$$

for $1 \le k \le i_0$. Next we derive an upper bound for the RHS of (30). As $\rho(x, a)$ is nonincreasing in a and $\pi_{0,i} \le \pi_{0,i_0} + \epsilon/2$ under the assumption in the theorem, we have

$$\begin{aligned}
\sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i_0}) &\le \sum_{i=i_0+1}^{i_0+N} \rho(x_i, \pi_{0,i_0}) + \sum_{i=k}^{i_0} \rho(x_i, \pi_{0,i}) \\
&= \sum_{i=i_0+1}^{i_0+N} (\rho(x_i, \pi_{0,i_0}) - \rho(x_i, \pi_{0,i})) + \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}) \\
&\le \sum_{i=i_0+1}^{i_0+N} (\rho(x_i, \pi_{0,i_0}) - \rho(x_i, \pi_{0,i_0} + \epsilon/2)) + \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}) \\
&= \sum_{i=i_0+1}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{2\phi(x_i, \pi_{0,i_0})\phi(x_i, \pi_{0,i_0} + \epsilon/2)} + \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}).
\end{aligned}$$

Using this upper bound and the fact that

$$\begin{aligned}
\rho(x, \pi_{0,i_0}) - \rho(x, \pi_{0,i_0} + \epsilon) =& \frac{f_0(x) - f_1(x)}{\phi(x, \pi_{0,i_0})} - \frac{f_0(x) - f_1(x)}{\phi(x, \pi_{0,i_0} + \epsilon)} \\
=& \frac{\epsilon(f_1(x) - f_0(x))^2}{\phi(x, \pi_{0,i_0})\phi(x, \pi_{0,i_0} + \epsilon)},
\end{aligned}$$

we know (30) is implied by

$$\sum_{i=k}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{\phi(x_i, \pi_{0,i_0})\phi(x_i, \pi_{0,i_0} + \epsilon)} - \sum_{i=i_0+1}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{2\phi(x_i, \pi_{0,i_0})\phi(x_i, \pi_{0,i_0} + \epsilon/2)} > \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}). \tag{31}$$

Some algebra shows that the LHS of (31) is bounded from below by

$$\sum_{i=k}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{2\phi(x_i, \pi_{0,i_0} + \epsilon/2)\phi(x_i, \pi_{0,i_0} + \epsilon)} \ge \sum_{i=k}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{2(f_0(x_i) \vee f_1(x_i))^2}.$$

Combining the above arguments, we get

$$
\begin{aligned}
&P\left(\hat{\pi}_{0,i_0} < \pi_{0,i_0} + \epsilon, B\right) \\
\geq &P\left(\max_{1\leq k\leq i_0} \hat{a}_{k,i_0+N} < \pi_{0,i_0} + \epsilon, B\right) \\
= &P\left(\sum_{i=k}^{i_0+N} \{\rho(x_i, \pi_{0,i_0}) - \rho(x_i, \pi_{0,i_0} + \epsilon)\} > \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i_0}) \text{ for all } 1 \leq k \leq i_0, B\right) \\
\geq &P\left(\sum_{i=k}^{i_0+N} \frac{\epsilon(f_1(x_i) - f_0(x_i))^2}{2(f_0(x_i) \vee f_1(x_i))^2} > \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}) \text{ for all } 1 \leq k \leq i_0\right) - P(B^c) \\
:= &P(A) - P(B^c).
\end{aligned}
$$

We first deal with $P(A)$. Notice that $\rho(x_i, \pi_{0,i})$ is a sequence of independent mean zero random variables with the variance

$$
\begin{aligned}
\text{var}(\rho(x_i, \pi_{0i})) &= \int \frac{(f_0(x) - f_1(x))^2}{\pi_{0i} f_0(x) + (1 - \pi_{0i}) f_1(x)} dx \\
&\leq \int \frac{(f_0(x) - f_1(x))^2}{\{\pi_0(0) f_0(x)\} \vee \{(1 - \pi_0(1)) f_1(x)\}} dx \\
&= \int_{\pi_0(0) f_0(x) > (1-\pi_0(1)) f_1(x)} \frac{(f_0(x) - f_1(x))^2}{\{\pi_0(0) f_0(x)\} \vee \{(1 - \pi_0(1)) f_1(x)\}} dx \\
&\quad + \int_{\pi_0(0) f_0(x) \leq (1-\pi_0(1)) f_1(x)} \frac{(f_0(x) - f_1(x))^2}{\{\pi_0(0) f_0(x)\} \vee \{(1 - \pi_0(1)) f_1(x)\}} dx \\
&\leq \int_{\pi_0(0) f_0(x) > (1-\pi_0(1)) f_1(x)} \frac{(f_0(x) - f_1(x))^2}{\pi_0(0) f_0(x)} dx \\
&\quad + \int_{\pi_0(0) f_0(x) \leq (1-\pi_0(1)) f_1(x)} \frac{(f_0(x) - f_1(x))^2}{(1 - \pi_0(1)) f_1(x)} dx \\
&\leq C_1 \int f_0(x) dx + C_2 \int f_1(x) dx < \infty,
\end{aligned}
$$

for some constants $C_1, C_2 > 0$. By Lemma 3.1 of [1], for any $\eta > 0$, there exists a large enough N such that,

$$
P\left(\max_{1\leq k\leq i_0} \left| \frac{1}{i_0 + N - k + 1} \sum_{i=k}^{i_0+N} \rho(x_i, \pi_{0,i}) \right| < \epsilon b \right) \geq 1 - O\left(\frac{1}{\epsilon^2 N}\right),
$$

for some constant

$$
b \leq E \frac{(f_1(x_i) - f_0(x_i))^2}{4(f_0(x_i) \vee f_1(x_i))^2}.
$$

Set $X_i = \frac{(f_1(x_i)-f_0(x_i))^2}{(f_0(x_i)\vee f_1(x_i))^2}$ which is a bounded random variable, and $\tilde{X}_i = EX_i - X_i$. Again by Lemma 3.1 of [1],

$$\begin{aligned}
&P\left(\min_{1\le k\le i_0}\frac{1}{i_0+N-k+1}\sum_{i=k}^{i_0+N}X_i > \frac{1}{2}EX_1\right)\\
=&P\left(\max_{1\le k\le i_0}\frac{1}{i_0+N-k+1}\sum_{i=k}^{i_0+N}\tilde{X}_i < \frac{1}{2}EX_1\right)\\
\ge&P\left(\max_{1\le k\le i_0}\left|\frac{1}{i_0+N-k+1}\sum_{i=k}^{i_0+N}\tilde{X}_i\right| < \frac{1}{2}EX_1\right)\\
>&1-O\left(\frac{1}{N}\right),
\end{aligned}$$

for large enough N. The above arguments thus imply that

$$P(A)\ge 1-O\left(\frac{1}{\epsilon^2 N}\right).$$

We next deal with B^c i.e., there exists a $1\le k\le i_0$ such that $\hat{a}_{k,i_0+N}=1$. Clearly, we only need to consider the case where $\pi_{0,i_0}+\epsilon<1$. In this case, we have $\pi_{0,i_0+N}\le \pi_0(t'') < \pi_{0,i_0}+\epsilon/2<1$. If $\hat{a}_{k,i_0+N}=1$, as the maximizer is unique, we have

$$\sum_{i=k}^{i_0+N}\log\phi(x_i,1) > \sum_{i=k}^{i_0+N}\log\phi(x_i,a) \tag{32}$$

for any $0\le a<1$. Under the assumption that $\int(\log f_i(x))^2 f_j(x)dx<\infty$ for $i,j=0,1$, we have $E[(\log\phi(x_i,a))^2]<\infty$ uniformly over i and $a\in[0,1]$. Note that for $a\ge\pi_{0i}$,

$$\begin{aligned}
(E\log\phi(x_i,a))' =&E\frac{f_0(x_i)-f_1(x_i)}{\phi(x_i,a)}\\
=&\int\frac{f_0(x)-f_1(x)}{\phi(x,a)}\phi(x,\pi_{0i})dx\\
=&\int\frac{f_0(x)-f_1(x)}{\phi(x,a)}\phi(x,\pi_{0i})-\frac{f_0(x)-f_1(x)}{\phi(x,\pi_{0i})}\phi(x,\pi_{0i})dx\\
=&\int\frac{(f_0(x)-f_1(x))^2(\pi_{0i}-a)}{\phi(x,a)}dx\\
\le&(\pi_{0i}-a)\int\frac{(f_0(x)-f_1(x))^2}{f_0(x)\vee f_1(x)}dx := C_0(\pi_{0i}-a),
\end{aligned}$$

where we have used the fact that $\int f_0(z)dz = \int f_1(z)dz = 1$. It is clear that as a function of a, $-E\log\phi(x_i,a)$ is convex. Thus we get

$$\begin{aligned} C_0(a-\pi_{0i})(1-a) - E\log\phi(x_i,a) \leq & -(E\log\phi(x_i,a))'(1-a) - E\log\phi(x_i,a) \\ \leq & -E\log\phi(x_i,1), \end{aligned}$$

that is

$$E\log\phi(x_i,a) - E\log\phi(x_i,1) \geq C_0(a-\pi_{0i})(1-a).$$

Now setting $\pi_{0,i_0} + \epsilon < a^* < 1$ and using the fact that $a^* - \pi_{0i} \geq \epsilon/2$ for $i \leq i_0 + N$, we obtain,

$$\begin{aligned} \sum_{i=k}^{i_0+N} (E\log\phi(x_i,a^*) - E\log\phi(x_i,1)) \geq & C_0(1-a^*)\sum_{i=k}^{i_0+N}(a^*-\pi_{0i}) \\ \geq & C_0(1-a^*)(i_0+N-k+1)\epsilon/2. \end{aligned}$$

For $\epsilon_0 < C_0(1-a^*)\epsilon/4$, let

$$B(a) := \max_{1\leq k\leq i_0} \left| \frac{1}{i_0+N-k+1} \sum_{i=k}^{i_0+N} \{\log\phi(x_i,a) - E\log\phi(x_i,a)\} \right| < \epsilon_0.$$

By Lemma 3.1 of [1], we have for large enough N,

$$P\left(B(a^*)\cap B(1)\right) > 1 - O\left(\frac{1}{\epsilon^2 N}\right).$$

Therefore on $B(a^*)\cap B(1)$, we have

$$\begin{aligned} & \{\text{there exists a } 1\leq k\leq i_0 \text{ such that } \hat{a}_{k,i_0+N} = 1\} \\ \subset & \cup_{k=1}^{i_0} \left\{ \sum_{i=k}^{i_0+N} \log\phi(x_i,1) > \sum_{i=k}^{i_0+N} \log\phi(x_i,a^*) \right\} \\ \subset & \cup_{k=1}^{i_0} \left\{ \frac{1}{i_0+N-k+1} \sum_{i=k}^{i_0+N} E\log\phi(x_i,1) + 2\epsilon_0 \right. \\ & \left. > \frac{1}{i_0+N-k+1} \sum_{i=k}^{i_0+N} E\log\phi(x_i,a^*) \right\} \\ \subset & \cup_{k=1}^{i_0} \{2\epsilon_0 > C_0(1-a^*)\epsilon/2\} = \emptyset. \end{aligned}$$

Then we get $P(B^c) \leq O\left(\frac{1}{\epsilon^2 N}\right)$ and thus

$$P\left(\hat{\pi}_{0,i_0} < \pi_{0,i_0} + \epsilon\right) \geq 1 - O\left(\frac{1}{\epsilon^2 N}\right).$$

Using similar arguments, we can prove that

$$P\left(\hat{\pi}_{0,i_0} > \pi_{0,i_0} - \epsilon\right) \geq 1 - O\left(\frac{1}{\epsilon^2 N}\right).$$

Therefore, we obtain

$$P\left(|\hat{\pi}_{0,i_0} - \pi_{0,i_0}| < \epsilon\right) \geq 1 - O\left(\frac{1}{\epsilon^2 N}\right).$$

□

Proof of Corollary 3.2.

Proof. For any $i_1 \leq i \leq i_l$, there exists a $2 \leq k \leq l$ such that $i_{k-1} \leq i \leq i_k$. Using the monotonicity of $\hat{\pi}_{0,i}$ and $\pi_{0,i}$, we get

$$\max_{i_1 \leq i \leq i_l} |\hat{\pi}_{0,i} - \pi_{0,i}| \leq \max_{1 \leq k \leq l} |\hat{\pi}_{0,i_k} - \pi_{0,i_k}| + \epsilon.$$

Thus by Theorem 3.1, we have

$$P\left(\max_{i_1 \leq i \leq i_l} |\hat{\pi}_{0,i} - \pi_{0,i}| < 2\epsilon\right) \geq P\left(\max_{1 \leq i \leq l} |\hat{\pi}_{0,i_k} - \pi_{0,i_{k-1}}| < \epsilon\right) \geq 1 - O\left(\frac{l}{\epsilon^2 N}\right).$$

□

Proof of Theorem 3.3. We provide some useful results from [48] and the high-level idea before presenting the detailed proof.

Some useful results. We present some results from [48], which will play an important role in the proof.

Recall that $\mathcal{F}$ denotes the class of densities on $[0,1]$. Let $\mathbf{G}_m = \{\mathbf{g} = (g_1, \ldots, g_m) : g_i \in \mathcal{F}\}$. Below we shall drop the subscript m for notational simplicity. Let ν be the Lebesgue measure (on $[0,1]$) and $L_r(\nu) = \{g : [0,1] \to \mathbb{R} : \int_0^1 |g|^r d\nu < \infty\}$. For $g \in L_r(\nu)$, write $\|g\|_{r,\nu}^r = \int_0^1 |g|^r d\nu$. We now define the entropy with bracketing. Consider $\mathbf{G}' \subseteq \mathbf{G}$. Let $N_B(\delta, \mathbf{G}', L_r(\nu))$ be the smallest value of N such that there exists a collection of functions $\{[\mathbf{g}_j^L, \mathbf{g}_j^U]\}_{j=1}^N$ with $\mathbf{g}_j^L = (g_j^{L,1}, \ldots, g_j^{L,m})$ and $\mathbf{g}_j^U = (g_j^{U,1}, \ldots, g_j^{U,m})$ such that for any $\mathbf{g} = (g^1, \ldots, g^m) \in \mathbf{G}'$, there exists a $1 \leq j \leq N$ satisfying that $g_j^{L,i} \leq g^i \leq g_j^{U,i}$ for all $1 \leq i \leq m$ and $\|\mathbf{g}_j^L - \mathbf{g}_j^U\|_{r,\nu,m}^r := m^{-1} \sum_{i=1}^m \|g_j^{L,i} - g_j^{U,i}\|_{r,\nu}^r \leq \delta$. Set $N_B(\delta, \mathbf{G}', L_r(\nu)) = +\infty$ if no finite set of such brackets exists. Let $\mathbb{H}_B(\delta, \mathbf{G}', L_r(\nu)) = \log N_B(\delta, \mathbf{G}', L_r(\nu))$. Write $d\mathbf{P} = (dP_1, \ldots, dP_m) = (f^1 d\nu, \ldots, f^m d\nu)$ and let A_i be some constant. We define $\mathbb{H}_B(\delta, \mathbf{G}', L_r(\mathbf{P}))$ in a similar way as $\mathbb{H}_B(\delta, \mathbf{G}', L_r(\nu))$ but with the

norm $\|\cdot\|_{r,\mathbf{P},m}^r = m^{-1}\sum_{i=1}^m \|\cdot\|_{r,P_i}^r$ to characterize the distance between $\mathbf{g}_j^L$ and $\mathbf{g}_j^U$. Operation on vector-valued function should be interpreted as applying the operation to each component of the vector-valued function.

LEMMA 8.5 (Lemma 7.11 of [48]). *Let*

$$\mathfrak{F} = \{f : [0,+\infty) \to [0,+\infty), f \text{ is decreasing }, f \le F\},$$

with F decreasing, $F \ge 1$ and $\int F^{2(1+a)}d\nu < \infty$ for some $a > 0$. Then for some $A > 0$,

$$\mathbb{H}_B(\delta, \mathfrak{F}, L_2(\nu)) \le A\delta^{-1}, \quad \text{for all } \delta > 0.$$

Below we present a modified version of Theorem 8.14 of [48], which is sufficient for our application. Note that the result in Theorem 8.14 of [48] is capable of dealing with dependent variables. However, to avoid unnecessary complication, we shall present a result that is specialized to the case of independent but non-identically distributed variables. We also mention that the entropy condition is on the convex class (34), which is different from the one in Theorem 8.14 of [48]. However, this change only requires a slightly modification (see the arguments of Theorem 8.6 below and the proof of Theorem 7.6 of [48]) of the proof in [48].

To state the result, let p_{i,θ_i} be a density indexed by a parameter θ_i for $1 \le i \le m$. Suppose we observe a set of random variables $x_i \sim p_{i,\theta_{0,i}}$ independently for $1 \le i \le m$ and $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,m}) \in \Theta$ for a given parameter space Θ. Write $\mathbf{p}_\theta = (p_{1,\theta_1}, \ldots, p_{m,\theta_m})$ with $\theta = (\theta_1, \ldots, \theta_m)$. Let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m) \in \Theta$ be an estimator of θ_0 such that

$$\sum_{i=1}^m \log p_{i,\hat{\theta}_i}(x_i) \ge \sum_{i=1}^m \log\left(\frac{p_{i,\hat{\theta}_i}(x_i) + p_{i,\theta_{0,i}}(x_i)}{2}\right). \tag{33}$$

Note that the maximum likelihood estimator of θ_0 automatically satisfies the above condition. Define $H_m^2(\mathbf{p}_\theta, \mathbf{p}_{\theta'}) = m^{-1}\sum_{i=1}^m H^2(p_{i,\theta_i}, p_{i,\theta_i'})$ and

$$\mathfrak{G}^{\text{conv}} = \left\{\frac{2\mathbf{p}_\theta}{\mathbf{p}_\theta + \mathbf{p}_{\theta_0}} : \theta \in \Theta\right\}. \tag{34}$$

Let

$$J_B(\delta) := \int_{\delta^2/c}^{\delta} \mathbb{H}_B^{1/2}(u, \mathfrak{G}^{\text{conv}}, L_2(\mathbf{p}_{\theta_0}))du \vee \delta \tag{35}$$

for some large enough c.

THEOREM 8.6. *Suppose that* $\{\mathbf{p}_\theta : \theta \in \Theta\}$ *is convex. Take* $\Psi(\delta) \geq J_B(\delta)$ *in such a way that* $\Psi(\delta)/\delta^2$ *is a non-increasing function of* δ. *Then for a universal constant* $\tilde{c}$ *and for*

$$\sqrt{m}\delta_m^2 \geq \tilde{c}\Psi(\delta_m), \tag{36}$$

we have for all $\delta \geq \delta_m$ *that*

$$P(H_m(\mathbf{p}_{\hat{\theta}}, \mathbf{p}_{\theta_0}) > \delta) \leq \tilde{c}\exp(-m\delta^2/\tilde{c}^2).$$

PROOF OF THEOREM 8.6. Let

$$Z_i(\theta) = \frac{2p_{i,\theta_i}(x_i)}{p_{i,\theta_i}(x_i) + p_{i,\theta_{0,i}}(x_i)}, \quad 1 \leq i \leq m.$$

We first claim the following basic inequality

$$H_m^2(\mathbf{p}_{\hat{\theta}}, \mathbf{p}_{\theta_0}) \leq \frac{1}{m}\sum_{i=1}^m \left(Z_i(\hat{\theta}) - P_{0,i}Z_i(\hat{\theta})\right), \tag{37}$$

$P_{0,i}Z_i(\theta) = \int \frac{2p_{i,\theta_i}}{p_{i,\theta_i}+p_{i,\theta_{0,i}}} p_{i,\theta_{0,i}} d\nu$. Note that

$$\begin{aligned} 0 \leq \sum_{i=1}^m \log Z_i(\hat{\theta}) &\leq \sum_{i=1}^m \left(Z_i(\hat{\theta}) - 1\right) \\ &= \sum_{i=1}^m \left(Z_i(\hat{\theta}) - P_{0,i}Z_i(\hat{\theta})\right) - \sum_{i=1}^m (1 - P_{0,i}Z_i(\hat{\theta})), \end{aligned}$$

where the first inequality follows from (33) and the second inequality follows from the fact that $\log(x) \leq x - 1$ for $x > 0$. On the other hand, we have

$$\begin{aligned} \sum_{i=1}^m (1 - P_{0,i}Z_i(\hat{\theta})) &= \sum_{i=1}^m \int \frac{p_{i,\theta_{0,i}} - p_{i,\hat{\theta}}}{p_{i,\theta_{0,i}} + p_{i,\hat{\theta}_i}} p_{i,\theta_{0,i}} d\nu \\ &= \sum_{i=1}^m \frac{1}{2}\int \frac{(p_{i,\theta_{0,i}} - p_{i,\hat{\theta}})^2}{p_{i,\theta_{0,i}} + p_{i,\hat{\theta}_i}} d\nu \geq H_m^2(\mathbf{p}_{\hat{\theta}}, \mathbf{p}_{\theta_0}), \end{aligned}$$

which gives (37). Applying the basic inequality and the peeling device, we have

$$\begin{aligned} &P(H_m(\mathbf{p}_{\hat{\theta}}, \mathbf{p}_{\theta_0}) > \delta) \\ \leq &P\left(\sup_{\theta\in\Theta: H_m(\mathbf{p}_\theta, \mathbf{p}_{\theta_0})>\delta} \frac{1}{m}\sum_{i=1}^m \{Z_i(\theta) - P_{0,i}Z_i(\theta)\} - H_m^2(\mathbf{p}_\theta, \mathbf{p}_{\theta_0}) > 0\right) \\ \leq &\sum_{s=0}^S P\left(\sup_{\theta\in\Theta: H_m(\mathbf{p}_\theta, \mathbf{p}_{\theta_0})\leq 2^{s+1}\delta} \frac{1}{m}\sum_{i=1}^m \{Z_i(\theta) - P_{0,i}Z_i(\theta)\} > 2^{2s}\delta^2\right) \end{aligned}$$

with $S = \min\{s : 2^{s+1}\delta > 1\}$. Observe the connection between $Z_i(\theta)$ and $\mathfrak{G}^{\text{conv}}$. The entropy condition can be used to control the upper bound above. The rest of the argument is similar to those in [48]. □

The high-level idea. To apply the above result, we shall take $\Theta = \Xi \times \mathcal{H}$ and $\theta = (\pi, f) \in \Theta$ in the above theorem. Most parts of our proof is devoted to showing the entropy condition (45). This is achieved in several steps. (1) We first apply Lemma 8.5 to the class of functions defined in (40), which implies a bound on the entropy of the class of functions $\mathfrak{G}_{k,i}^{U,\text{conv}}$ in (43). (2) We then argue that one can construct the delta-bracketing set for $\mathfrak{G}_{k,i}^{U,\text{conv}}$ with $2 \le i \le m$ based on the one for $\mathfrak{G}_{k,1}^{U,\text{conv}}$. Thus the entropy of $\mathfrak{G}_{k}^{U,\text{conv}}$ in (42) is of the same order as that of $\mathfrak{G}_{k,i}^{U,\text{conv}}$, which leads to (45) after some algebra. (3) Under (45), (36) is satisfied with $\delta = Mm^{-1/3}$ for some large enough M. The result thus follows from Theorem 8.6.

Proof. Consider the collection of mixture densities $\mathbf{F} = \{\tilde{\mathbf{f}} = (\tilde{f}^1, \ldots, \tilde{f}^m) : \tilde{f}^i = \tilde{\pi}(i/m)f_0 + (1 - \tilde{\pi}(i/m))\tilde{f}_1, \tilde{\pi} \in \Xi, \tilde{f}_1 \in \mathcal{H}\}$. It is known that

$$\mathbb{H}_B(\delta, \Xi, L_2(\mu_m)) \le A_1/\delta, \tag{38}$$

where μ_m denotes the discrete probability measure with equal mass $1/m$ on the grid $\{1/m, 2/m, \ldots, 1\}$, see e.g. [49]. Let $\{[\pi_k^L, \pi_k^U]\}_{k=1}^N$ be the delta-bracketing set for Ξ. For any $\tilde{\pi} \in \Xi$ and $\tilde{f}_1 \in \mathcal{H}$, there exists a $1 \le k \le N$ such that

(39)
$$\begin{aligned}\tilde{f}_k^{L,i} := \pi_k^L(i/m)f_0 + (1 - \pi_k^U(i/m))\tilde{f}_1 \le& \tilde{\pi}(i/m)f_0 + (1 - \tilde{\pi}(i/m))\tilde{f}_1 \\ \le& \pi_k^U(i/m)f_0 + (1 - \pi_k^L(i/m))\tilde{f}_1 := \tilde{f}_k^{U,i}.\end{aligned}$$

We focus on the upper bound in the following analysis. For $1 \le k \le N$, let

$$\begin{aligned}\mathbf{F}_k^L &= \{(\tilde{f}_k^{L,1}, \ldots, \tilde{f}_k^{L,m}) : \tilde{f}_1 \in \mathcal{H}\}, \\ \mathbf{F}_k^U &= \{(\tilde{f}_k^{U,1}, \ldots, \tilde{f}_k^{U,m}) : \tilde{f}_1 \in \mathcal{H}\},\end{aligned}$$

where $\tilde{f}_k^{L,i}$ and $\tilde{f}_k^{U,i}$ are defined in (39). Further define

$$\mathfrak{G}_{k,i}^U = \left\{ \left(\frac{\tilde{f}^i f^i}{\tilde{f}^i + f^i} \right)^{1/2} : \tilde{f}^i \text{ is the } i\text{th component of } \tilde{\mathbf{f}} \in \mathbf{F}_k^U \right\}. \tag{40}$$

Note that $\left(\frac{\tilde{f}^i f^i}{\tilde{f}^i+f^i}\right)^{1/2} \le (f^i)^{1/2} \vee 1$. Under the assumption that $\int_0^1 f_1^{1+a} d\nu < \infty$, we have $\sup_{1\le i\le m} \int (f^i)^{1+a} \vee 1 d\nu < \infty$. Applying Lemma 8.5 with $F = (f^i)^{1/2} \vee 1$, we know that

$$\mathbb{H}_B(\delta, \mathfrak{G}_{k,i}^U, L_2(\nu)) \le A_3/\delta.$$

Next we define

$$\mathfrak{G}^{\text{conv}} = \left\{ \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}}+\mathbf{f}} : \tilde{\mathbf{f}} \in \mathbf{F} \right\}, \tag{41}$$

$$\mathfrak{G}_k^{U,\text{conv}} = \left\{ \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}}+\mathbf{f}} : \tilde{\mathbf{f}} \in \mathbf{F}_k^U \right\}, \tag{42}$$

$$\mathfrak{G}_{k,i}^{U,\text{conv}} = \left\{ \frac{\tilde{f}^i}{\tilde{f}^i+f^i} : \tilde{f}^i \text{ is the } i\text{th component of } \tilde{\mathbf{f}} \in \mathbf{F}_k^U \right\}. \tag{43}$$

Our goal is to derive an upper bound for the entropy with bracketing of $\mathfrak{G}^{\text{conv}}$, and then apply Theorem 8.6 to obtain the desired result. To this end, we shall first derive the entropy with bracketing for the classes $\mathfrak{G}_{k,i}^{U,\text{conv}}$ and $\mathfrak{G}_k^{U,\text{conv}}$.

For $\tilde{f}^i$ and $\tilde{g}^i$ being the ith components of $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ in $\mathbf{F}_k^U$, we have

$$\begin{aligned}
&\int \left(\frac{\tilde{f}^i}{\tilde{f}^i+f^i} - \frac{\tilde{g}^i}{\tilde{g}^i+f^i} \right)^2 dP_i \\
=&\int \left\{ \left(\frac{\tilde{f}^i}{\tilde{f}^i+f^i} \right)^{1/2} - \left(\frac{\tilde{g}^i}{\tilde{g}^i+f^i} \right)^{1/2} \right\}^2 \left\{ \left(\frac{\tilde{f}^i}{\tilde{f}^i+f^i} \right)^{1/2} + \left(\frac{\tilde{g}^i}{\tilde{g}^i+f^i} \right)^{1/2} \right\}^2 dP_i \\
\le& 4 \int \left\{ \left(\frac{\tilde{f}^i}{\tilde{f}^i+f^i} \right)^{1/2} - \left(\frac{\tilde{g}^i}{\tilde{g}^i+f^i} \right)^{1/2} \right\}^2 dP_i \\
=& 4 \int \left\{ \left(\frac{\tilde{f}^i f^i}{\tilde{f}^i+f^i} \right)^{1/2} - \left(\frac{\tilde{g}^i f^i}{\tilde{g}^i+f^i} \right)^{1/2} \right\}^2 d\nu.
\end{aligned}$$

Hence we get

$$\mathbb{H}_B(2\delta, \mathfrak{G}_{k,i}^{U,\text{conv}}, L_2(P_i)) \le \mathbb{H}_B(\delta, \mathfrak{G}_{k,i}^U, L_2(\nu)) \le A_3/\delta.$$

Below we argue that one can construct the delta-bracketing set for $\mathfrak{G}_{k,i}^{U,\text{conv}}$ with $2 \le i \le m$ based on the one for $\mathfrak{G}_{k,1}^{U,\text{conv}}$. Consider $\tilde{f}^1$ which is the ith

component of $\tilde{\mathbf{f}} \in \mathbf{F}_k^U$ and a pair of functions (ζ^L, ζ^U) such that

(44)
$$\begin{aligned}\frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L}{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L + f^i} \leq & \frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\tilde{f}_1}{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\tilde{f}_1 + f^i} \\ \leq & \frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U}{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U + f^i}\end{aligned}$$

and

$$\begin{aligned}&\int \left(\frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L}{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L + f^i} - \frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U}{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U + f^i} \right)^2 dP_i \\ =&\int \left\{ \frac{(1-\pi_k^L(1/m))(\zeta^L - \zeta^U)f^i}{(\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L + f^i)(\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U + f^i)} \right\}^2 dP_i \\ \leq&\delta^2\end{aligned}$$

Clearly, (44) implies that $\zeta^L \leq \tilde{f}_1 \leq \zeta^U$. Moreover, (44) still holds if we replace $(\pi_k^L(1/m), \pi_k^U(1/m))$ by $(\pi_k^L(i/m), \pi_k^U(i/m))$ for any $2 \leq i \leq m$. Using the following bounds (which hold as $\varepsilon \leq \pi_k^L, \pi_k^U \leq 1-\varepsilon$)

$$\begin{aligned}&\frac{(1-\pi_k^L(i/m))}{(1-\pi_k^L(1/m))} \leq 1, \\ &\frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^U + f^i}{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^U + f^i} \leq \frac{\pi_k^U(1/m)}{\pi_k^U(i/m)} + \frac{1-\pi_k^U(1/m)}{1-\pi_k^U(i/m)} + 1 \leq C_1, \\ &\frac{\pi_k^U(1/m)f_0 + (1-\pi_k^L(1/m))\zeta^L + f^i}{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^L + f^i} \leq \frac{\pi_k^U(1/m)}{\pi_k^U(i/m)} + \frac{1-\pi_k^U(1/m)}{1-\pi_k^U(i/m)} + 1 \leq C_1,\end{aligned}$$

for some constant $C_1 > 0$, we can show that

$$\begin{aligned}&\int \left(\frac{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^L}{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^L + f^i} - \frac{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^U}{\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^U + f^i} \right)^2 dP_i \\ =&\int \left\{ \frac{(1-\pi_k^L(i/m))(\zeta^L - \zeta^U)f^i}{(\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^L + f^i)(\pi_k^U(i/m)f_0 + (1-\pi_k^L(i/m))\zeta^U + f^i)} \right\}^2 dP_i \\ \leq&C_1^4\delta^2.\end{aligned}$$

The above arguments suggest that we can construct the delta-bracketing set for $\mathfrak{G}_{k,i}^{U,\text{conv}}$ with $2 \leq i \leq m$ based on the one for $\mathfrak{G}_{k,1}^{U,\text{conv}}$. Therefore, we have

$$\mathbb{H}_B(\delta, \mathfrak{G}_k^{U,\text{conv}}, L_2(\mathbf{P})) \leq A_4/\delta.$$

Similarly, we can get

$$\mathbb{H}_B(\delta, \mathfrak{G}_k^{L,\text{conv}}, L_2(\mathbf{P})) \leq A_5/\delta,$$

where $\mathfrak{G}_k^{L,\text{conv}}$ is defined in a similar way as $\mathfrak{G}_k^{U,\text{conv}}$ but with $\tilde{\mathbf{f}} \in \mathbf{F}_k^L$. For any $\tilde{\mathbf{f}} \in \mathbf{F}$, there exists a $1 \leq k \leq N$ and $\tilde{\mathbf{f}}^L \in \mathbf{F}_k^L$ and $\tilde{\mathbf{f}}^U \in \mathbf{F}_k^U$ such that

$$\frac{\tilde{\mathbf{f}}^L}{\tilde{\mathbf{f}}^L + \mathbf{f}} \leq \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} \leq \frac{\tilde{\mathbf{f}}^U}{\tilde{\mathbf{f}}^U + \mathbf{f}}.$$

Let $\{[\mathbf{b}_i^L, \mathbf{c}_i^L]\}_{i=1}^{N_k^L}$ and $\{[\mathbf{b}_i^U, \mathbf{c}_i^U]\}_{i=1}^{N_k^U}$ be the delta-bracketing sets for $\mathfrak{G}_k^{L,\text{conv}}$ and $\mathfrak{G}_k^{U,\text{conv}}$ respectively. Then there exists a (i, j) such that

$$\mathbf{b}_i^L \leq \frac{\tilde{\mathbf{f}}^L}{\tilde{\mathbf{f}}^L + \mathbf{f}} \leq \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} \leq \frac{\tilde{\mathbf{f}}^U}{\tilde{\mathbf{f}}^U + \mathbf{f}} \leq \mathbf{c}_j^U.$$

By the triangle inequality,

$$\begin{aligned}
\|\mathbf{c}_j^U - \mathbf{b}_i^L\|_{2,\mathbf{P},m} \leq & \left\| \mathbf{c}_j^U - \frac{\tilde{\mathbf{f}}^U}{\tilde{\mathbf{f}}^U + \mathbf{f}} \right\|_{2,\mathbf{P},m} + \left\| \frac{\tilde{\mathbf{f}}^U}{\tilde{\mathbf{f}}^U + \mathbf{f}} - \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} \right\|_{2,\mathbf{P},m} \\
& + \left\| \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} - \frac{\tilde{\mathbf{f}}^L}{\tilde{\mathbf{f}}^L + \mathbf{f}} \right\|_{2,\mathbf{P},m} + \left\| \frac{\tilde{\mathbf{f}}^L}{\tilde{\mathbf{f}}^L + \mathbf{f}} - \mathbf{b}_i^L \right\|_{2,\mathbf{P},m} \\
\leq & \left\| \frac{\tilde{\mathbf{f}}^U}{\tilde{\mathbf{f}}^U + \mathbf{f}} - \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} \right\|_{2,\mathbf{P},m} + \left\| \frac{\tilde{\mathbf{f}}}{\tilde{\mathbf{f}} + \mathbf{f}} - \frac{\tilde{\mathbf{f}}^L}{\tilde{\mathbf{f}}^L + \mathbf{f}} \right\|_{2,\mathbf{P},m} + 2\delta.
\end{aligned}$$

We focus on the first component of the first term. Note that

$$\begin{aligned}
& \int \left(\frac{\pi_k^U(1/m) f_0 + (1 - \pi_k^L(1/m)) \tilde{f}_i}{\pi_k^U(1/m) f_0 + (1 - \pi_k^L(1/m)) \tilde{f}_i + f^i} - \frac{\pi(1/m) f_0 + (1 - \pi(1/m)) \tilde{f}_i}{\pi(1/m) f_0 + (1 - \pi(1/m)) \tilde{f}_i + f^i} \right)^2 dP_i \\
= & \int \left\{ \frac{(\pi_k^U(1/m) - \pi(1/m)) f_0 f^i + (\pi(1/m) - \pi_k^L(1/m)) \tilde{f}^i f^i}{(\pi_k^U(1/m) f_0 + (1 - \pi_k^L(1/m)) \tilde{f}_i + f^i)(\pi(1/m) f_0 + (1 - \pi(1/m)) \tilde{f}_i + f^i)} \right\}^2 dP_i \\
\leq & C_2 \{ (\pi_k^U(1/m) - \pi(1/m))^2 + (\pi(1/m) - \pi_k^L(1/m))^2 \},
\end{aligned}$$

for some constant $C_2 > 0$. Hence we obtain

$$\mathbb{H}_B(\delta, \mathfrak{G}^{\text{conv}}, L_2(\mathbf{P})) \leq A_6/\delta.$$

Note that

$$\int_{\delta^2/c}^{\delta} \mathbb{H}_B^{1/2}(u, \mathfrak{G}^{\text{conv}}, L_2(\mathbf{P})) du \leq A_7 \sqrt{\delta}. \tag{45}$$

Finally, we apply Theorem 8.6 (also see Theorem 7.6 of [48]). Consider $\Theta = \Xi \times \mathcal{H}$ and $\theta = (\pi, f) \in \Theta$. In view of (45), (36) is satisfied with $\delta = Mm^{-1/3}$ for some large enough M. Thus by Theorem 8.6, we have

$$P\left(H_m((\pi_0, f_1), (\hat{\pi}_0, \hat{f}_1)) > Mm^{-1/3}\right) \leq M_1 \exp(-M_2 m^{1/3}),$$

for some $M_1, M_2 > 0$. □

Proof of Corollary 3.4.

PROOF. Using (23), we obtain

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\int_{\mathrm{Lfdr}_i(x)>d_m} |\widehat{\mathrm{Lfdr}}_i(x) - \mathrm{Lfdr}_i(x)| f^i(x)dx \\
=&\frac{1}{m}\sum_{i=1}^{m}\int_{\mathrm{Lfdr}_i(x)>d_m} |\widehat{\mathrm{Lfdr}}_i(x) - \mathrm{Lfdr}_i(x)| \frac{\pi_0(i/m) f_0(x)}{\mathrm{Lfdr}_i(x)}dx \\
\leq&\frac{C}{md_m}\sum_{i=1}^{m}\int_0^1 |\widehat{\mathrm{Lfdr}}_i(x) - \mathrm{Lfdr}_i(x)|dx = o_p(1),
\end{aligned}$$

for some constant $C > 0$ and a sequence d_m with $d_m = o(1)$ and $m^{-1/3}/d_m = o(1)$. As D_0 in Condition (C1) is continuous at 0,

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\int_{\mathrm{Lfdr}_i(x)\leq d_m} |\widehat{\mathrm{Lfdr}}_i(x) - \mathrm{Lfdr}_i(x)| f^i(x)dx \leq&\frac{1}{m}\sum_{i=1}^{m} P(\mathrm{Lfdr}_i(x_i) \leq d_m) \\
=&D_0(d_m) + o_p(1) = o_p(1).
\end{aligned}$$

Thus we have

$$\frac{1}{m}\sum_{i=1}^{m}\int_0^1 |\widehat{\mathrm{Lfdr}}_i(x) - \mathrm{Lfdr}_i(x)| f^i(x)dx = o_p(1). \tag{46}$$

In view of (46), to justify Condition (C3), it suffices to show the following uniform law of large numbers,

$$\sup_{\pi\in\Xi, f\in\mathcal{H}} \left|\frac{1}{m}\sum_{i=1}^{m}(g_i(x_i) - E[g_i(x_i)])\right| = o_p(1), \tag{47}$$

where

$$g_i(x_i) = \left|\frac{\pi(i/m) f_0(x_i)}{\pi(i/m) f_0(x_i) + (1 - \pi(i/m)) f(x_i)} - \mathrm{Lfdr}_i(x_i)\right|.$$

We justify this claim in Lemma 8.7 below. By (46) and (47), we must have

$$\frac{1}{m}\sum_{i=1}^{m}|\widehat{\mathrm{Lfdr}}_i(x_i)-\mathrm{Lfdr}_i(x_i)|=o_p(1), \tag{48}$$

which verifies Condition (C3). □

LEMMA 8.7. *For Ξ and $\mathcal{H}$ as defined in Section 3.3, we have*

$$\sup_{\pi\in\Xi,f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}(g_i(x_i)-E[g_i(x_i)])\right|=O_p(m^{-1/3}).$$

PROOF OF LEMMA 8.7. Let

$$g_{+,i}(x_i)=\left(\frac{\pi(i/m)f_0(x_i)}{\pi(i/m)f_0(x_i)+(1-\pi(i/m))f(x_i)}-\mathrm{Lfdr}_i(x_i)\right)_+,$$
$$g_{-,i}(x_i)=\left(\mathrm{Lfdr}_i(x)-\frac{\pi(i/m)f_0(x_i)}{\pi(i/m)f_0(x_i)+(1-\pi(i/m))f(x_i)}\right)_+,$$

where $(a)_+=a\vee 0$. Note that $g_i(x_i)=g_{+,i}(x_i)+g_{-,i}(x_i)$. Thus we just need to show that

$$\sup_{\pi\in\Xi,f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}(g_{+,i}(x_i)-E[g_{+,i}(x_i)])\right|=O_p(m^{-1/3}), \tag{49}$$

$$\sup_{\pi\in\Xi,f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}(g_{-,i}(x_i)-E[g_{-,i}(x_i)])\right|=O_p(m^{-1/3}). \tag{50}$$

We only prove (49) as the arguments for (50) is essentially the same. Below we shall adopt the notation defined in the proof of Theorem 3.3. Note that $g_{+,i}(x_i)$ is a decreasing function of $f(x_i)$ and increasing function of $\pi(i/m)$. Recall from (38) that

$$\mathbb{H}_B(\delta,\Xi,L_1(\mu_m))\le A_1/\delta \tag{51}$$

for some $A_1>0$, where μ_m denotes the discrete probability measure with equal mass $1/m$ at the grids $\{1/m,2/m,\ldots,1\}$ Let $\{[\pi_k^L,\pi_k^U]\}_{k=1}^{N_1}$ be a δ-bracketing set for Ξ such that $m^{-1}\sum_{i=1}^{m}|\pi_k^L(i/m)-\pi_k^U(i/m)|\le\delta$. Suppose $\pi\in[\pi_k^L,\pi_k^U]$. Note that

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}(g_{+,i}(x_i)-E[g_{+,i}(x_i)])\\ \le&\frac{1}{m}\sum_{i=1}^{m}\left\{\left(f_k^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+-E\left(f_k^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\},\end{aligned}$$

where for any given $f \in \mathcal{H}$, we define

(52)
$$f_k^{U,i} := \frac{\pi_k^U(i/m) f_0}{\pi_k^U(i/m) f_0 + (1 - \pi_k^U(i/m)) f}, \quad f_k^{L,i} := \frac{\pi_k^L(i/m) f_0}{\pi_k^L(i/m) f_0 + (1 - \pi_k^L(i/m)) f}.$$

This observation motivates us to consider the following classes of vector-valued functions

$$\mathcal{F}_k^U := \{\mathbf{f}_k^U = (f_k^{U,1}, \ldots, f_k^{U,m}) : f \in \mathcal{H}\},$$
$$\mathcal{F}_k^L := \{\mathbf{f}_k^L = (f_k^{L,1}, \ldots, f_k^{L,m}) : f \in \mathcal{H}\},$$

where $f_k^{U,i}$ and $f_k^{L,i}$ are defined in (52). Note that $\mathcal{F}_{k,i}^U = \{f_k^{U,i} : f \in \mathcal{H}\}$ is a class of increasing functions that are bounded from below and above. Thus $\mathbb{H}_B(\delta, \mathcal{F}_{k,i}^U, L_1(P_i)) \leq A_2/\delta$. Using similar arguments as in the proof of Theorem 3.3, we can construct the delta-bracketing sets for $\mathcal{F}_{k,i}^U$ with $2 \leq i \leq m$ based on the one for $\mathcal{F}_{k,1}^U$. Thus we have $\mathbb{H}_B(\delta, \mathcal{F}_k^U, L_1(\mathbf{P})) \leq A_3/\delta$ and similarly $\mathbb{H}_B(\delta, \mathcal{F}_k^L, L_1(\mathbf{P})) \leq A_4/\delta$. Let $\{[\zeta_{k,j}^L, \zeta_{k,j}^U]\}_{j=1}^{N_2}$ and $\{[\xi_{k,j}^L, \xi_{k,j}^U]\}_{j=1}^{N_3}$ be the δ-bracketing sets for $\mathcal{F}_k^U$ and $\mathcal{F}_k^L$ respectively. For $f_k^U \in \mathcal{F}_k^U$ and $f_k^L \in \mathcal{F}_k^L$, there exists (j, l) such that $\zeta_{k,j}^L \leq \mathbf{f}_k^U \leq \zeta_{k,j}^U$ and $\xi_{k,l}^L \leq \mathbf{f}_k^L \leq \xi_{k,l}^U$. Thus we get

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m (g_{+,i}(x_i) - E[g_{+,i}(x_i)]) \\
\leq &\frac{1}{m}\sum_{i=1}^m \left\{ \left(f_k^{U,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ - E\left(f_k^{L,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ \right\} \\
\leq &\frac{1}{m}\sum_{i=1}^m \left\{ \left(\zeta_{k,j}^{U,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\xi_{k,l}^{L,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ \right\} \\
\leq &\frac{1}{m}\sum_{i=1}^m \left\{ \left(\zeta_{k,j}^{U,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\zeta_{k,j}^{U,i}(x_i) - \mathrm{Lfdr}_i(x_i)\right)_+ \right\} + C_1\delta,
\end{aligned}$$

for some $C_1 > 0$. Here we have used the fact that

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\left\{E\left(\zeta_{k,j}^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\xi_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}\\
\leq&\frac{1}{m}\sum_{i=1}^{m}E\left|\zeta_{k,j}^{U,i}(x_i)-\xi_{k,l}^{L,i}(x_i)\right|\\
\leq&\frac{1}{m}\sum_{i=1}^{m}E\left|\zeta_{k,j}^{U,i}(x_i)-f_k^{U,i}(x_i)+f_k^{U,i}(x_i)-f_k^{L,i}(x_i)+f_k^{L,i}(x_i)-\xi_{k,l}^{L,i}(x_i)\right|\\
\leq&\frac{1}{m}\sum_{i=1}^{m}E\left|f_k^{U,i}(x_i)-f_k^{L,i}(x_i)\right|+2\delta\\
\leq&\frac{C}{m}\sum_{i=1}^{m}\left|\pi_k^U(i/m)-\pi_k^L(i/m)\right|+2\delta=(C+2)\delta,
\end{aligned}$$

for some $C > 0$. Similarly,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\left(g_{+,i}(x_i)-E[g_{+,i}(x_i)]\right)\\
\geq&\frac{1}{m}\sum_{i=1}^{m}\left\{\left(f_k^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+ - E\left(f_k^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}\\
\geq&\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\xi_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\xi_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}-C_2\delta.
\end{aligned}$$

By the Hoeffding's inequality, we have for any $1 \leq k \leq N_1$, $1 \leq j \leq N_2$ and $1 \leq l \leq N_3$,

$$\begin{aligned}
&P\left(\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\zeta_{k,j}^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\zeta_{k,j}^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}\geq\epsilon\right)\leq\exp(-C_3m\epsilon^2),\\
&P\left(\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\xi_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+ - E\left(\xi_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}\geq\epsilon\right)\leq\exp(-C_3m\epsilon^2),
\end{aligned}$$

for some $C_3 > 0$. Hence we get

$$
\begin{aligned}
&P\left(\sup_{\pi\in\Xi, f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\left(g_{+,i}(x_i)-E[g_{+,i}(x_i)]\right)\right|>\epsilon\right)\\
\leq& P\left(\max_{1\leq k\leq N_1, 1\leq j\leq N_2}\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\zeta_{k,j}^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+-E\left(\zeta_{k,j}^{U,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}>\epsilon-C_1\delta\right)\\
&+P\left(\max_{1\leq k\leq N_1, 1\leq l\leq N_3}\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\zeta_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+-E\left(\zeta_{k,l}^{L,i}(x_i)-\mathrm{Lfdr}_i(x_i)\right)_+\right\}<C_2\delta-\epsilon\right)\\
\leq& 2\exp\left\{-C_3 m(\epsilon-\delta C_1\vee C_2)^2+(A_1+A_3\vee A_4)/\delta\right\},
\end{aligned}
$$

where we have used the union bound and the Hoeffding's inequality to obtain the second inequality. The result follows by choosing $\epsilon = C_4 m^{-1/3}$ and $\delta = m^{-1/3}$ for some large enough C_4. □

Proof of Theorem 3.5.

Proof. We first show that $\hat{\lambda}_m \to^p \lambda_0$. Recall from (28) that

$$\sup_{\lambda\geq\lambda_\infty}\left|\widehat{R}_m(\lambda)-R(\lambda)\right|\to^p 0.$$

For any small enough $\epsilon > 0$, by the definition of λ_0, we have $\inf_{\lambda_0+\epsilon\leq\lambda\leq 1} R(\lambda) > \alpha$. Therefore,

$$P(\inf_{\lambda_0+\epsilon\leq\lambda\leq 1}\widehat{R}_m(\lambda)>\alpha)\leq P(\hat{\lambda}_m<\lambda_0+\epsilon)\to 1.$$

On the other hand, as $R(\lambda_0-\epsilon)<\alpha$, we have

$$P(\widehat{R}_m(\lambda_0-\epsilon)<\alpha)\leq P(\hat{\lambda}_m\geq\lambda_0-\epsilon)\to 1.$$

Combing the above arguments, we get $\hat{\lambda}_m \to^p \lambda_0$. Next, following the arguments in the proof of Lemma 8.4, we have

$$\sup_{\lambda\geq\lambda_\infty/2}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{\theta_i=1,\widehat{\mathrm{Lfdr}}_i(x_i)\leq\lambda\}-D_2(\lambda)\right|\to^p 0.$$

As $\lambda_0\geq\lambda_\infty$, $P(\hat{\lambda}_m>\lambda_\infty/2)\to 1$. Thus we get

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{\theta_i=1,\widehat{\mathrm{Lfdr}}_i(x_i)\leq\hat{\lambda}_m\}-D_2(\hat{\lambda}_m)\to^p 0.$$

By the continuity of D_2, we have $D_2(\hat{\lambda}_m)\to^p D_2(\lambda_0)$. The conclusion thus follows. □

Derivation of the EM-algorithm from the full data likelihood. The EM algorithm can be motivated by the full data likelihood that has access to hidden/latent variables. To see this, we note that the full log-likelihood of $\{(x_i, \theta_i) : i = 1, 2 \ldots, m\}$ is given by

$$\begin{aligned}\log p(\mathbf{x}, \boldsymbol{\theta}) =& \sum_{i=1}^{m} \log\{(1-\theta_i) f_0(x_i) + \theta_i f_1(x_i)\} \\ &+ \sum_{i=1}^{m} \{(1-\theta_i)\log(\pi_{0i}) + \theta_i \log(1-\pi_{0i})\},\end{aligned}$$

where $\mathbf{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Let $\Pi^{(t)} = (\hat{\pi}_{01}^{(t)}, \ldots, \hat{\pi}_{0m}^{(t)})$. We note that the posterior distribution of θ_i given $\mathbf{x}, f_1$ and Π is equal to Bernoulli$(1 - Q_i^{(t)})$, where $Q_i^{(t)} = \hat{\pi}_{0i}^{(t)} f_0(x_i) / \{\hat{\pi}_{0i}^{(t)} f_0(x_i) + (1 - \hat{\pi}_{0i}^{(t)}) f_1(x_i)\}$. The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:
E-step: Define

$$\begin{aligned}D(f_1, \Pi | f_1^{(t)}, \Pi^{(t)}) =& E_{\boldsymbol{\theta}|f_1^{(t)}, \Pi^{(t)}}[\log p(\mathbf{x}, \boldsymbol{\theta})] \\ =& \sum_{i=1}^{m} \left\{ Q_i^{(t)} \log f_0(x_i) + (1 - Q_i^{(t)}) \log f_1(x_i) \right\} \\ &+ \sum_{i=1}^{m} \left\{ Q_i^{(t)} \log \pi_i + (1 - Q_i^{(t)}) \log(1 - \pi_i) \right\}\end{aligned}$$

as the expected value of the log-likelihood function with resepect to the current conditional distribution of $\boldsymbol{\theta}$ given the current estimates $f_1^{(t)}$ and $\Pi^{(t)}$.
M-step: Find the parameters that maximize $D(f_1, \Pi | f_1^{(t)}, \Pi^{(t)})$. Equivalently, we have

$$\begin{aligned}\hat{\Pi} =& \arg\max_{\Pi \in \mathcal{M}} \sum_{i=1}^{m} \left\{ Q_i^{(t)} \log \pi_i + (1 - Q_i^{(t)}) \log(1 - \pi_i) \right\} \\ =& \arg\min_{\Pi \in \mathcal{M}} \sum_{i=1}^{m} \left(Q_i^{(t)} - \pi_i \right)^2, \\ f_1^{(t+1)} =& \arg\max_{\tilde{f}_1 \in \mathcal{H}} \sum_{i=1}^{m} (1 - Q_i^{(t)}) \log \tilde{f}_1(x_i).\end{aligned}$$

Competing methods. A classic procedure for multiple testing is the BH procedure proposed in [5]. We now briefly describe the BH procedure. Let

$x_{(1)} \leq \ldots \leq x_{(m)}$ be the order statistics of the p-values $x_1, \ldots, x_m$. Given a control level $\alpha \in (0,1)$, let

$$k = \max \left\{ i \in \{0, 1, \ldots, m+1\} : x_{(i)} \leq \alpha \frac{i}{m} \right\},$$

where $x_0 = 0$ and $x_{(m+1)} = 1$. The BH procedure rejects all hypotheses for which $x_i \leq x_{(k)}$. If $k = 0$, then no hypotheses will be rejected. It has been shown that the BH procedure controls the FDR at the level $\alpha\pi_0$, where π_0 is the proportion of null hypothesis. R function *p.adjust* in the base stats package is used to obtain results based on the BH procedure. To improve power, [43] (ST) estimates the proportion of null hypothesis

$$\hat{\pi}(\lambda) = \min \left\{ 1, \frac{\#\{x_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)} \right\},$$

where λ is a tuning parameter. Let

$$k = \max \left\{ i \in \{0, 1, \ldots, m+1\} : x_{(i)} \leq \frac{\alpha}{\hat{\pi}(\lambda)} \frac{i}{m} \right\}.$$

The ST procedure rejects all hypotheses for which $x_i \leq x_{(k)}$. If $k = 0$, no hypotheses will be rejected. The bioconductor R package *qvalue* is used to obtain results based on the ST procedure.

To incorporate auxiliary information in a data-adaptive way, [34] proposed the structure adaptive BH algorithm (SABHA). Specifically, given a target FDR level α, a threshold $\tau \in [0,1]$, and values $\hat{\pi}_{01}, \ldots, \hat{\pi}_{0m} \in [0,1]$, where $\hat{\pi}_{0i}$ represents an estimated probability that the ith test corresponds to a null, define

$$k = \max \left\{ i \in \{1, \ldots, m\}, x_i \leq \left(\frac{\alpha}{\hat{\pi}_{0i}} \frac{i}{m} \right) \wedge \tau \right\}.$$

Reject hypotheses with corresponding p-value x_i satisfying

$$x_i \leq \left(\frac{\alpha}{\hat{\pi}_{0i}} \frac{k}{m} \right) \wedge \tau.$$

We use the code provided in [34] to implement SABHA. [31] proposed to use two parameters to estimate proportion of null hypothesis and number of rejections (Adaptive SeqStep). Specifically, let $A(\lambda, k) = \sum_{i=1}^{k} I(x_i > \lambda)$ count p-values exceeding the threshold λ within the first k ordered hypotheses and $R(s,k) = \sum_{i=1}^{k} I(x_i \leq s)$ count number of rejections within the

first k ordered hypotheses. Then the proportion of null hypotheses can be estimated by

$$\hat{\pi}(\lambda, k) = \frac{1 + A(\lambda, k)}{n(1-\lambda)}.$$

The Adaptive SeqStep procedure thus works as follows: for some $0 \leq s \leq \lambda \leq 1$, reject all hypotheses with $x_i \leq s$ and $H_{(i)}, i = 1, \ldots, \hat{k}_{AS}$, where $H_{(i)}$ are ordered hypotheses based on p-values, and

$$\hat{k}_{AS} = \max \{k : \mathrm{FDP}_{AS}(k; s, \lambda) \leq \alpha\},$$

where

$$\mathrm{FDP}_{AS}(k; s, \lambda) = \frac{s}{1-\lambda} \frac{1 + A(\lambda, k)}{R(s, k) \vee 1}.$$

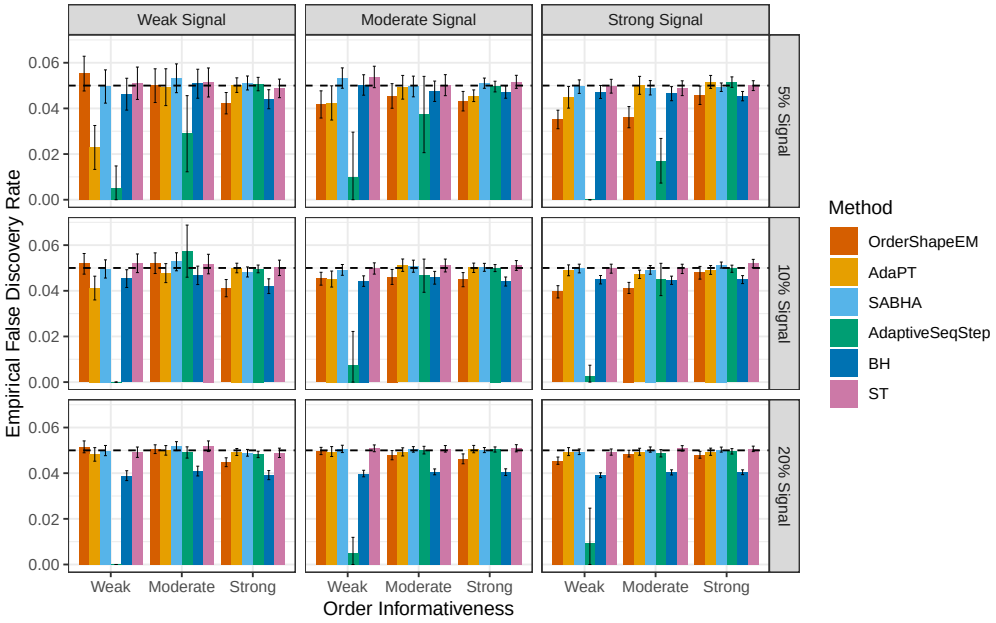
We use the code provided in [31] for implementation. We also compare to the adaptive p-value thresholding procedure (AdaPT) [30] and use the "adapt_glm" function in R package "adaptMT" (v0.2.1.9000) with natural splines of 6 d.f. as covariates for both the null probability and the alternative distribution. OrderShapeEM, AdaPT, SABHA and Adaptive SeqStep are multiple testing procedures that incorporate auxiliary information.

We evaluate the performance based on FDR control (empirical FDR) and power (true positive rate, i.e., number of true positives divided by number of alternatives) with the target FDR level $\alpha = 0.05$. Results are averaged over 100 replications (except for the global null where the number of replications is 2,000) and the 95% confidence interval are reported.

Additional simulation results. Figure 5 shows the numerical results when z-values under the alternative hypothesis are from the non-central gamma distribution. Figure 7 shows the numerical results when there is noise in the auxiliary information. Figure 8 compares OrderShapeEM to SABHA+, which uses the SABHA rejection rule and the mixing probabilities estimated by OrderShapeEM. The setting is the same as Figure 2. Figure 9 and Figure 10 show the numerical results under a lower signal density and under a global null, respectively. Figure 11 and Figure 12 show the numerical results under varying f_1 and varying f_0 respectively. Figure 13 shows the performance with $m = 500, 100, 2000$. Figure 14 shows the FDR control for AdaPT without the correction term (AdaPT+).

Fig 5: Performance under skewed alternative distribution.

(a) FDR control



(b) power comparison

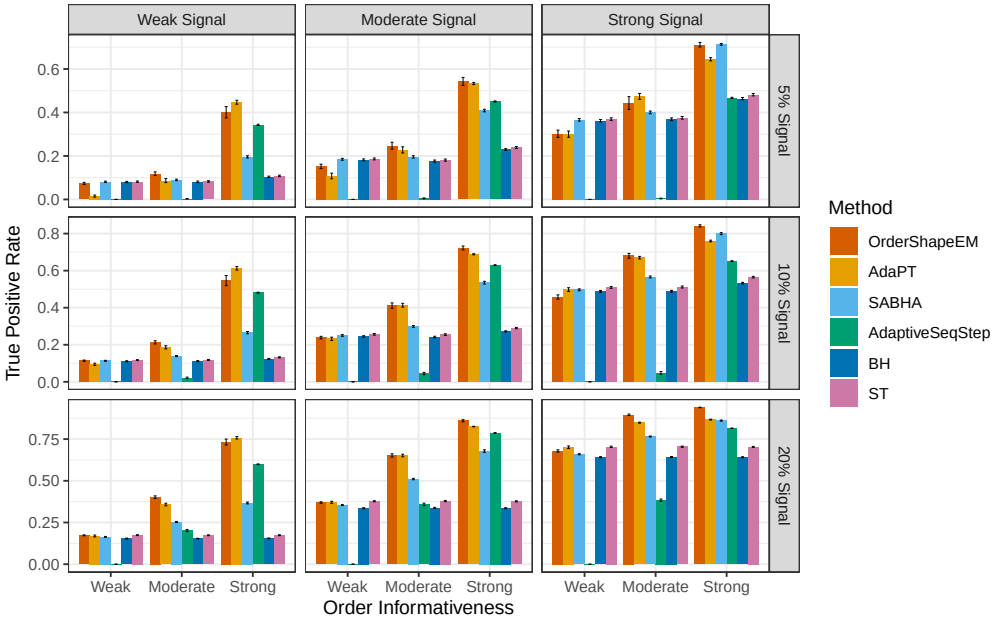
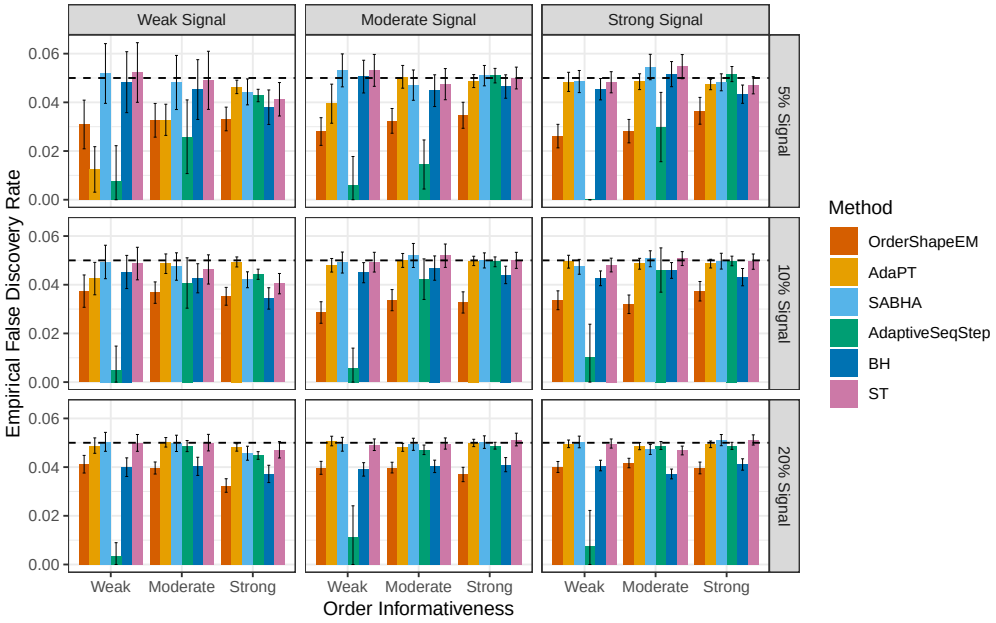


Fig 6: Performance under correlated hypotheses.

(a) FDR control



(b) power comparison

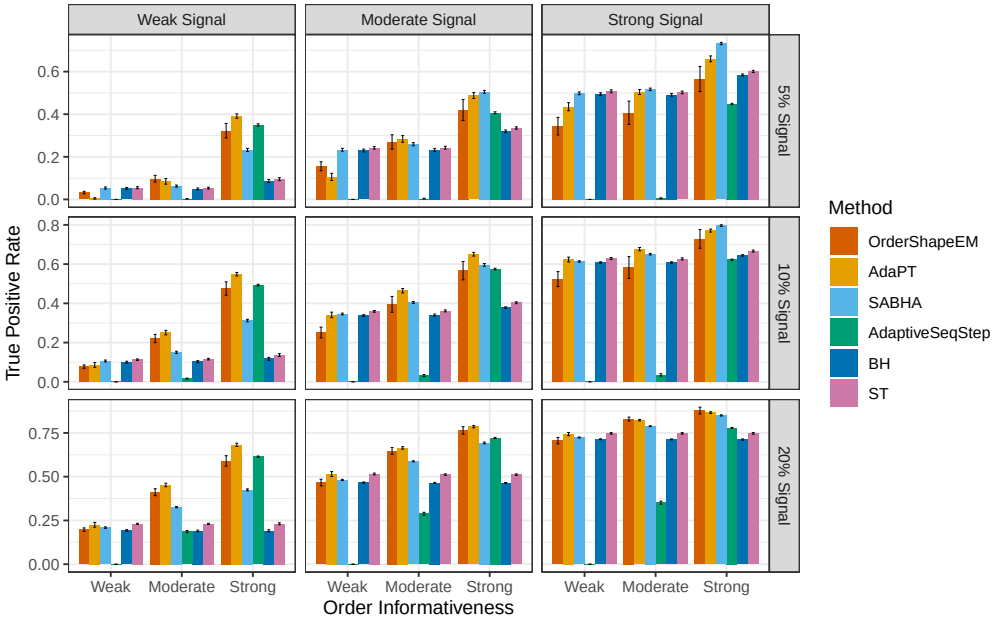
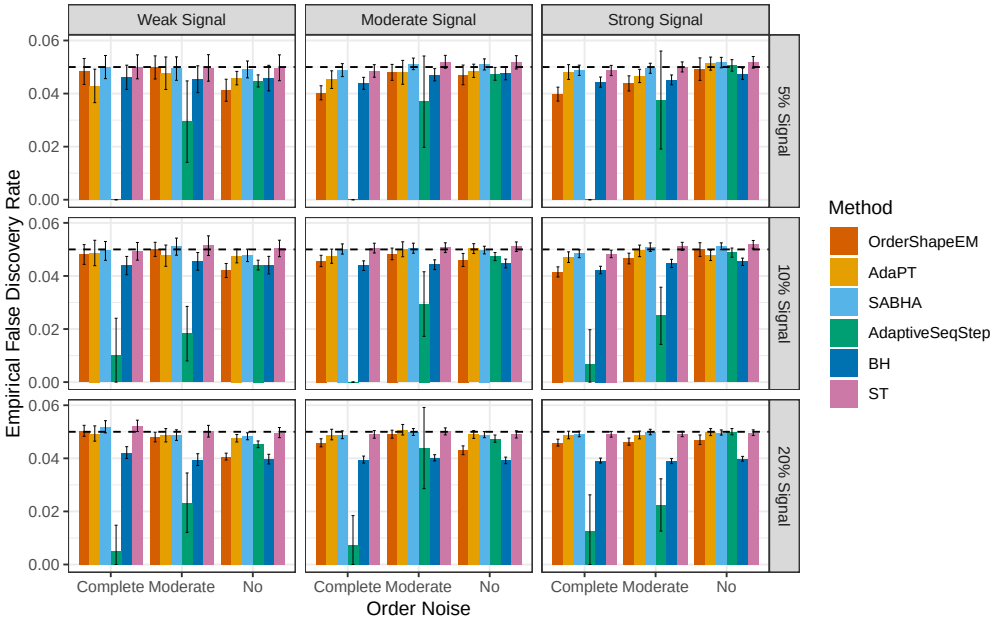


Fig 7: Performance under noisy auxiliary information.

(a) FDR control



(b) power comparison

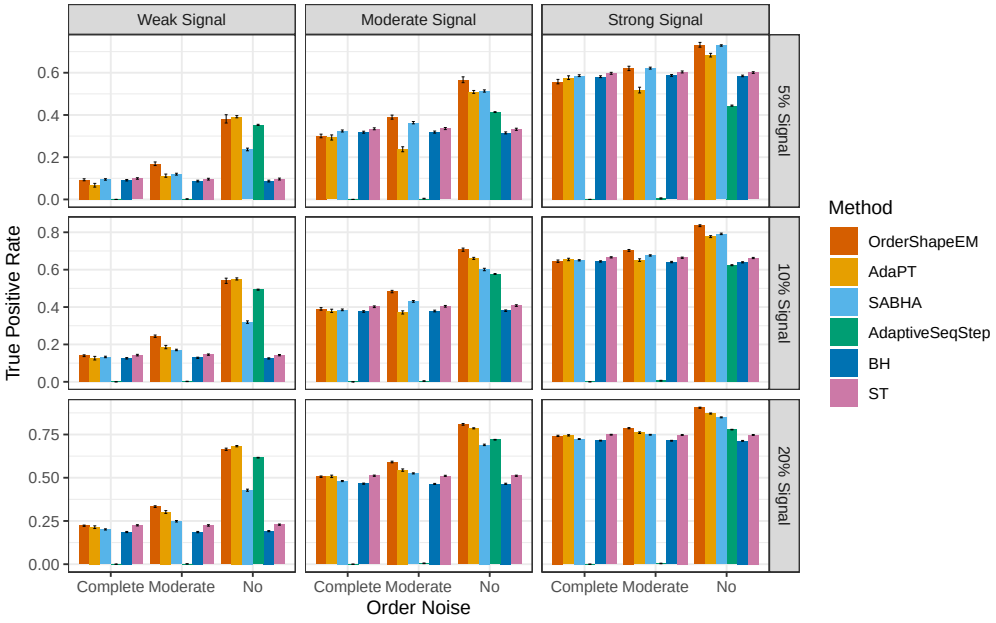
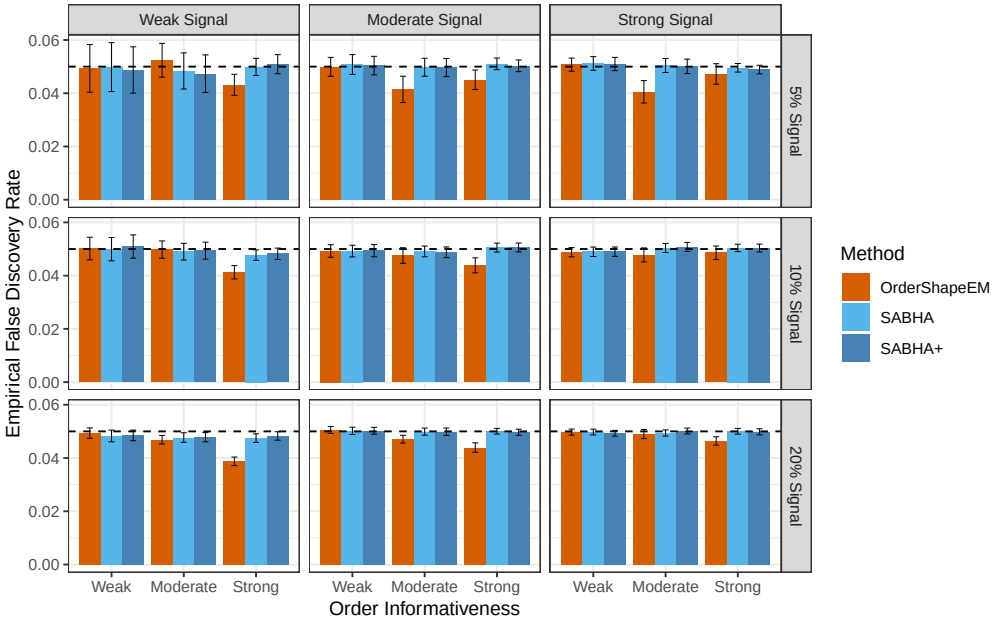


Fig 8: The effect of the optimal rejection rule.

(a) FDR control



(b) power comparison

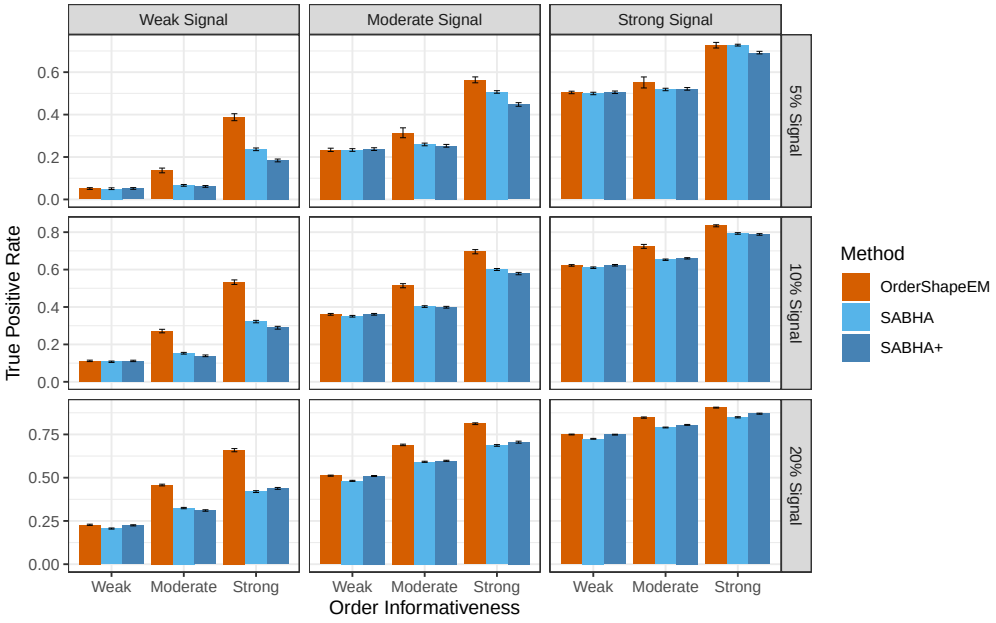
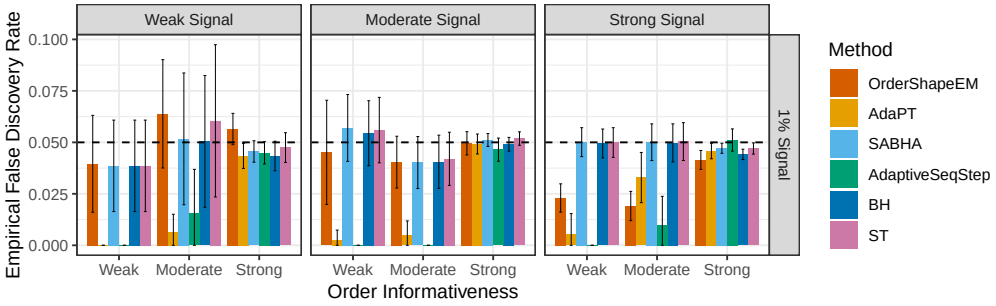


Fig 9: Performance under a lower signal density (1%).

(a) FDR control



(b) power comparison

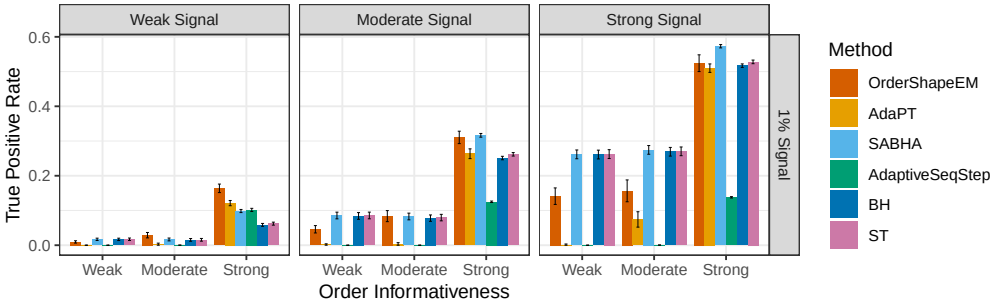


Fig 10: FDR control under the global null.

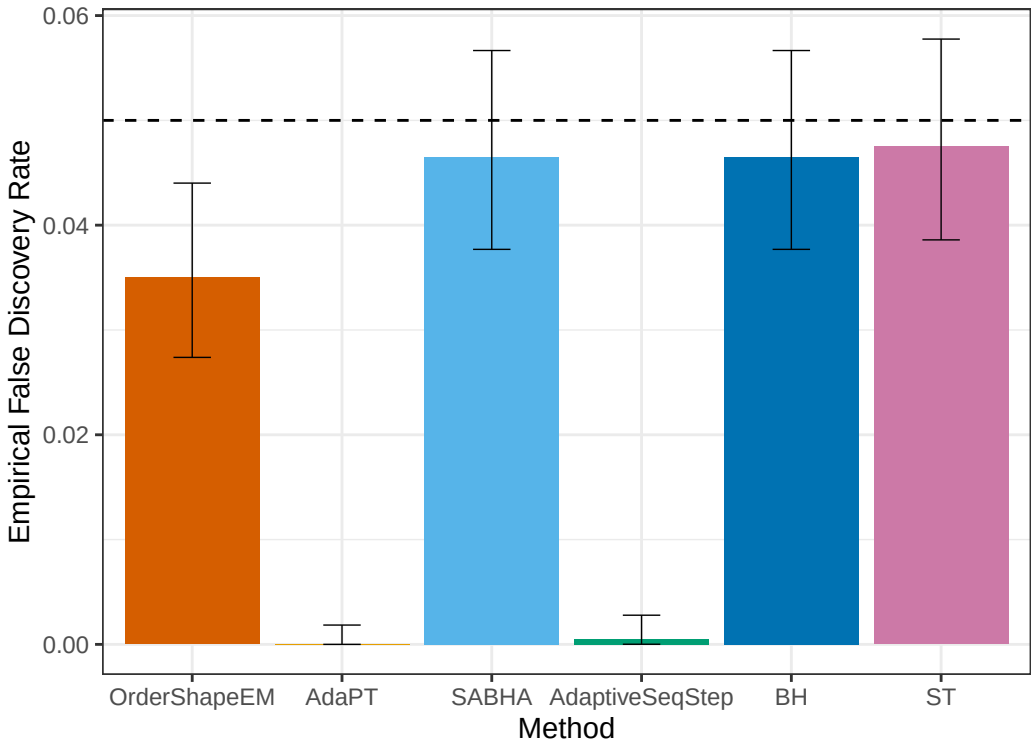
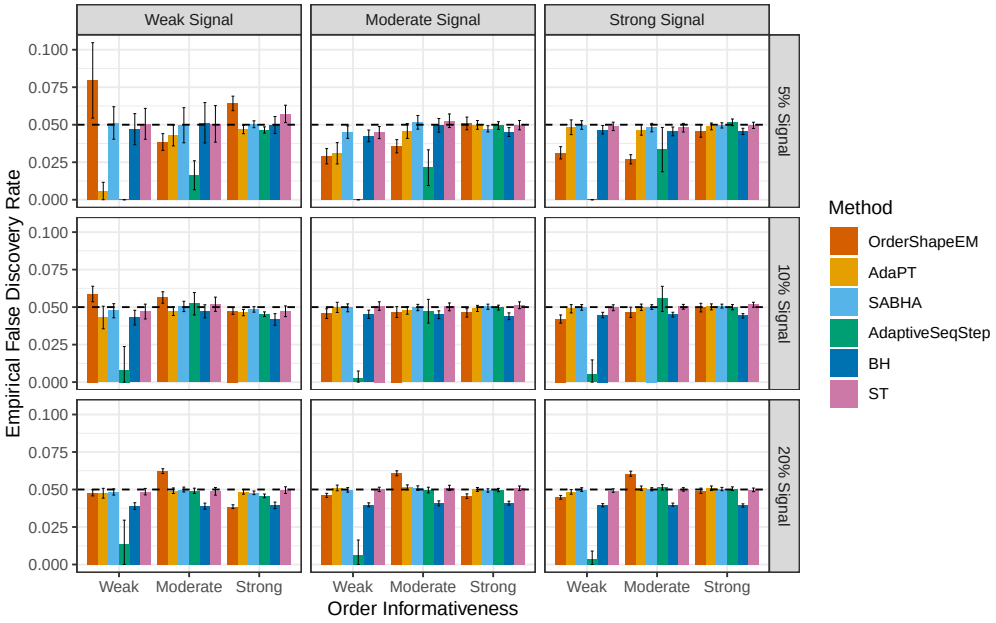


Fig 11: Performance under varying f_1.

(a) FDR control



(b) power comparison

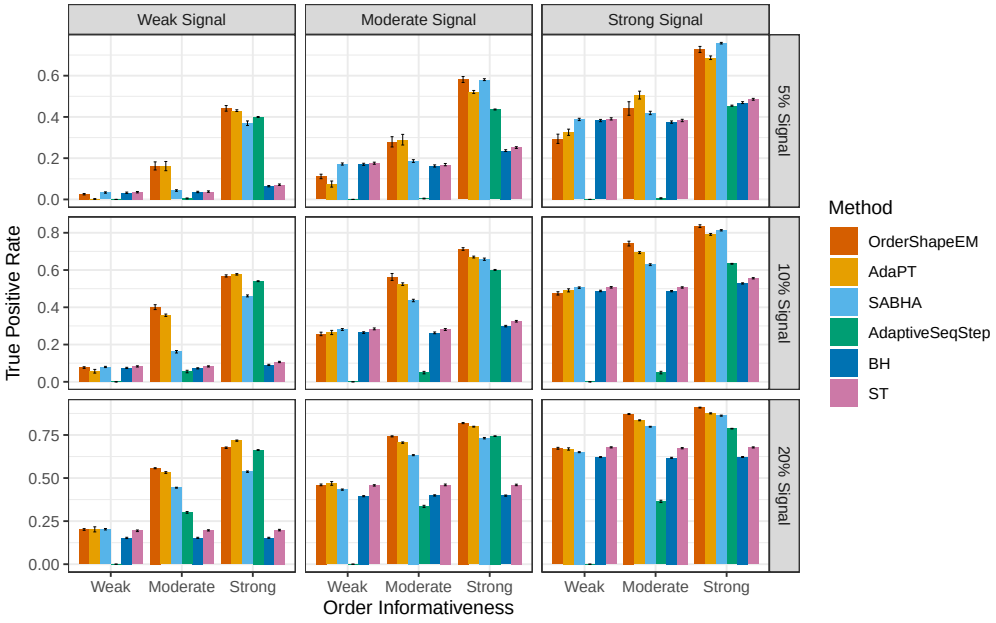
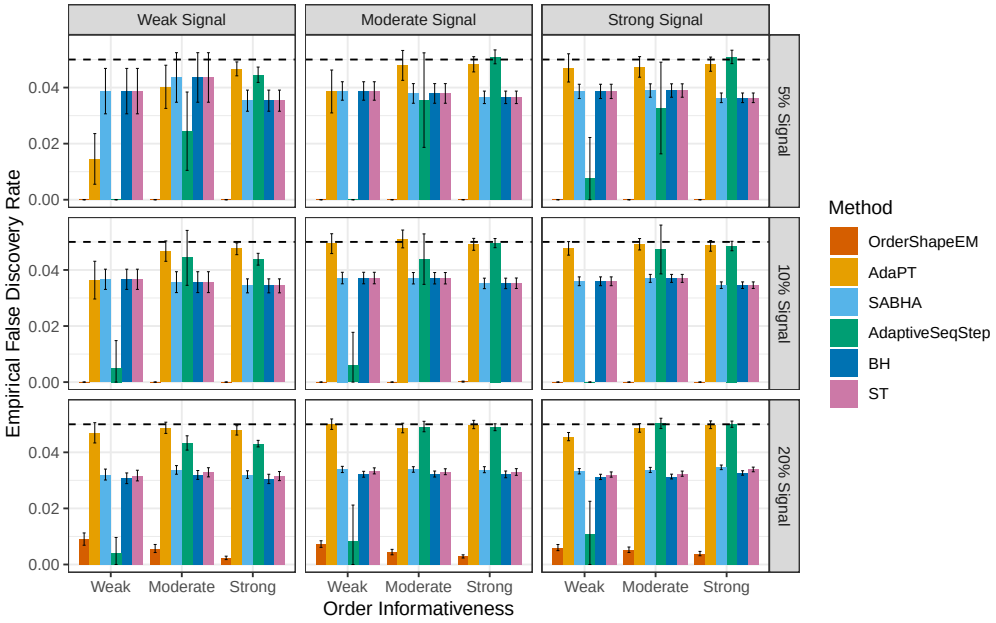


Fig 12: Performance under varying f_0.

(a) FDR control



(b) power comparison

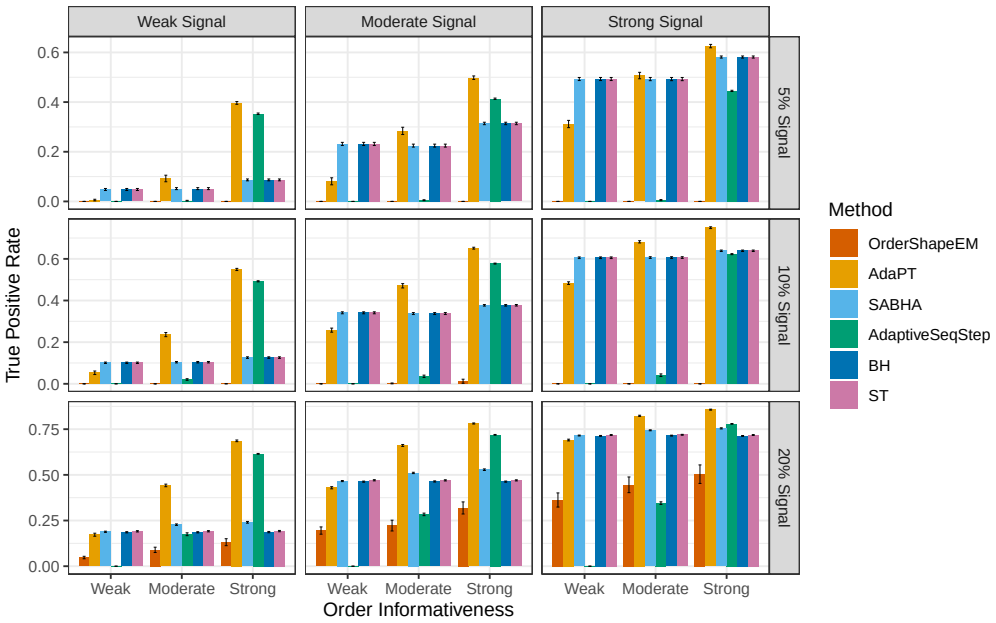
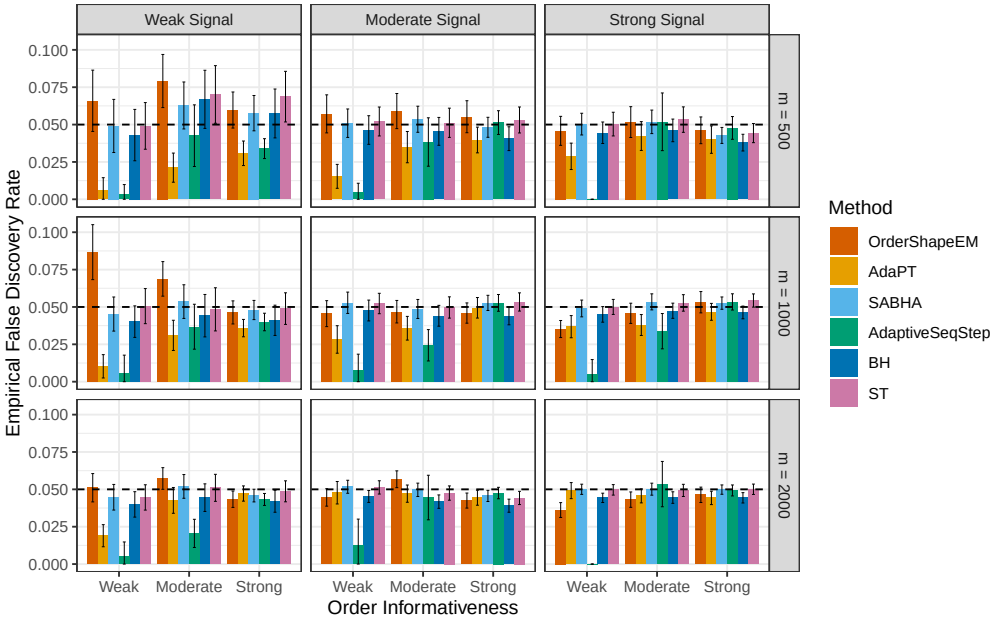


Fig 13: Performance under normal alternative distribution with $m = 500, 100, 2000$.

(a) FDR control



(b) power comparison

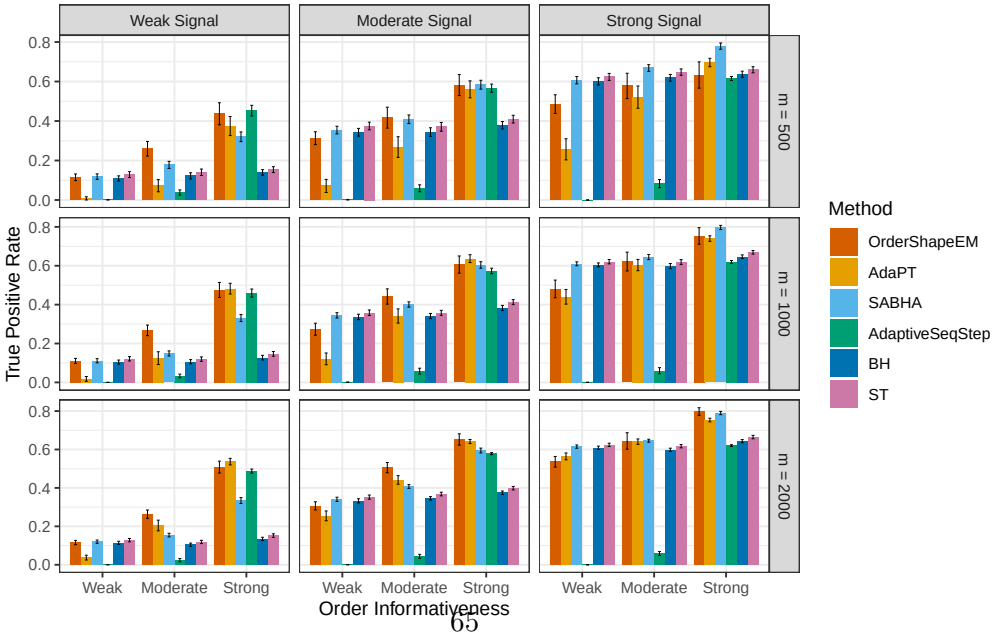
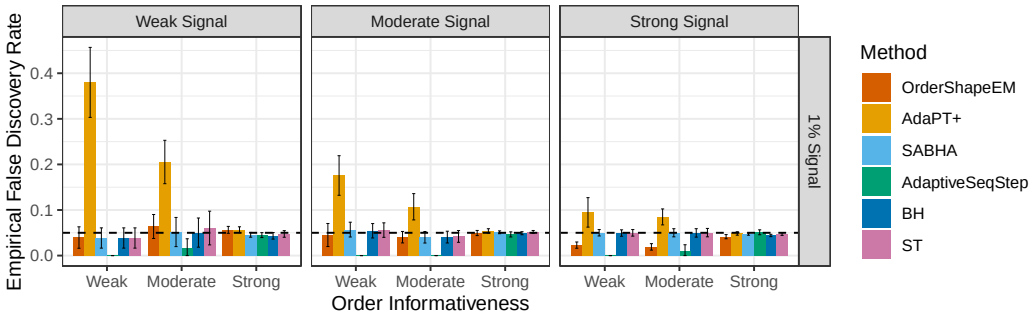


Fig 14: Inadequate FDR control for AdaPT without the correction term (AdaPT+).



References.

[1] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, **26**, 641-647.

[2] Barber, R. F., and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, **43**, 2055-2085.

[3] 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**, 1172–1183.

[4] Barlow, R. E., and Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, **67**, 140-147.

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

[6] Birgé, L. (1987). Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, **15**, 995-1012.

[7] Boca, S. M., and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, **6**, e6035.

[8] Coronary Artery Disease (C4D) Genetics Consortium. (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature Genetics*, **43**, 339-344.

[9] Cai, T. T., and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *Journal of the American Statistical Association*, **104**, 1467–1481.

[10] Cao, H., Sun, W., and Kosorok, M. R. (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika*, **100**, 495–502.

[11] Deb, N., Saha, S., Guntuboyina, A., and Sen, B. (2019). Two-component mixture model in the presence of covariates. arXiv preprint arXiv:1810.07897.

[12] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1-22.

[13] Dobriban, E. (2017). Weighted mining of massive collections of p-values by convex optimization. *Information and Inference: A Journal of the IMA*, **7**, 251-275.

[14] Durot, C., Kulikov, V. N., and Lopuhaä, H. P. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *Annals of Statistics*, **40**, 1578-1608.

[15] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical science*, **23**, 1-22.

[16] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

[17] Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

[18] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008). Unsupervised empirical Bayesian testing with external covariates. *Annals of Applied Statistics*, **2**, 714–735.

[19] Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.

[20] Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scandinavian*

Actuarial Journal, **1956**, 125-153.

[21] G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society, Series B*, **78**, 423-444.

[22] Henzi, A., M'osching, A. and Dümbgen, L. (2020). Accelerating the Pool-Adjacent-Violators Algorithm for isotonic distributional regression. arXiv:2006.05527

[23] Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, **105**, 1215–1227.

[24] Huang, J.Y., Bai, L., Cui, B.W., Wu, L., Wang, L.W., An, Z.Y,, Ruan, S.L. Yu, Y., Zhang, X.Y., and Chen, J. (2020). Leveraging biological and statistical covariates improves the detection power in epigenome-wide association testing. *Genome biology*, **21**: 1-19.

[25] Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, **13**, 577–580.

[26] Ignatiadis, N., and Huber, W. (2017). Covariate-powered weighted multiple testing with false discovery rate control. arXiv preprint arXiv:1701.05179.

[27] Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, **41**, 200–209.

[28] Kristensen, V. N., Lingjarde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Borresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Review Cancer*, **14**, 299–313.

[29] Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B*, **67**, 555–572.

[30] Lei, L., and Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society, Series B*, **80**, 649–679.

[31] Lei, L., and Fithian, W. (2016). Power of Ordered Hypothesis Testing. arXiv preprint arXiv:1606.01969.

[32] Lei, L., Ramdas, A., and Fithian, W. (2020). STAR: A general interactive framework for FDR control under structural constraints. *Biometrika*, to appear.

[33] Li, A., and Barber, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association*, **112**, 837–849.

[34] Li, A., and Barber, R. F. (2019). Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *Journal of the Royal Statistical Society, Series B*, **81**, 45–74.

[35] Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

[36] Lynch, G., Guo, W., Sarkar, S., and Finner, H. (2017). The control of the false discovery rate in fixed sequence multiple testing. *Electronic Journal of Statistics*, **11**, 4649–4673.

[37] Robertson, T., and Waltman, P. (1968). On estimating monotone parameters. *Annals of Mathematical Statistics*, **39**, 1030–1039.

[38] Robertson, T., Wright, F. T., and Dykstra, R. (1988). *Order restricted statistical inference*, Wiley.

[39] Schunkert, H., Konig, IR., Kathiresan, S., Reilly, MP., Assimes, TL., Holm, H., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, **43**, 333–338.

[40] Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False

discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, **110**, 459–471.

[41] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479-498.

[42] Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society, Series B*, **66**, 187–205.

[43] Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.

[44] Sun, W., Reich, B. J., Cai, T. T., Guindani, M., and Schwartzman, A. (2015). False discovery control in large-scale multiple testing. *Journal of the Royal Statistical Society, Series B*, **77**, 59–83.

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

[46] Tang, W., and Zhang, C. (2005). Bayes and empirical bayes approaches to controlling the false discovery rate. *Technical report*, Dept. Statistics and Biostatistics, Rutgers Univ.

[47] Tansey, W., Koyejo, O., Poldrack, R. A., and Scott, J. G. (2018). False discovery rate smoothing. *Journal of the American Statistical Association*, **13**, 1156–1171.

[48] van de geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[49] van der vaart, A., and Wellner, J. (2000). *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics, New York.

[50] Xiao, J., Cao, H., and Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome wide multiple testing. *Bioinformatics*, **33**, 2873-2881

[51] Zhang, X., and Chen, J. (2020). Covariate adaptive false discovery rate control with applications to Omics-Wide multiple testing. *Journal of the American Statistical Association*, to appear.