ZAP: Z-VALUE ADAPTIVE PROCEDURES FOR FALSE DISCOVERY RATE CONTROL WITH SIDE INFORMATION

DENNIS LEUNG AND WENGUANG SUN

ABSTRACT. Adaptive multiple testing with covariates is an important research direction that has gained major attention in recent years. It has been widely recognized that leveraging side information provided by auxiliary covariates can improve the power of false discovery rate (FDR) procedures. Currently, most such procedures are devised with p-values as their main statistics. However, for two-sided hypotheses, the usual data processing step that transforms the primary statistics, known as z-values, into p-values not only leads to a loss of information carried by the main statistics, but can also undermine the ability of the covariates to assist with the FDR inference. We develop a z-value based covariate-adaptive (ZAP) methodology that operates on the intact structural information encoded jointly by the z-values and covariates. It seeks to emulate the oracle z-value procedure via a working model, and its rejection regions significantly depart from those of the p-value adaptive testing approaches. The key strength of ZAP is that the FDR control is guaranteed with minimal assumptions, even when the working model is misspecified. We demonstrate the state-of-the-art performance of ZAP using both simulated and real data, which shows that the efficiency gain can be substantial in comparison with p-value based methods. Our methodology is implemented in the R package `zap`.

1. INTRODUCTION

In modern scientific studies, a ubiquitous task is to test a multitude of two-sided hypotheses regarding the presence of nonzero effects. The problem of multiple testing with covariates has received much recent attention, as leveraging contextual information beyond what is offered by the main statistics can enhance both the power and interpretability of existing false discovery rate (FDR; Benjamini and Hochberg, 1995) methods. This has marked a gradual paradigm shift from the Benjamini-Hochberg (BH) procedure and its immediate variants (e.g. Benjamini and Hochberg, 2000, Storey, 2002) that are based solely on the p-values. For instance, in the differential analysis of RNA-sequencing data, the average read depths across samples can provide useful side information alongside individual p-values, and incorporating such information promises to improve the efficiency of existing methods. The importance of this direction has been reflected by its intense research activities; see Boca and Leek (2018), Chen et al. (2017), Ignatiadis et al. (2016), Lei and Fithian (2018), Li and Barber (2019), Yurko et al. (2020), Zhang and Chen

2000 *Mathematics Subject Classification.* 62H05.
Key words and phrases. multiple testing, false discovery rate, z-value, beta mixture, side information.

(2020) for an incomplete list of related works. In contrast with BH and its variants that apply a universal threshold to all p-values, these methods boil down to setting varied p-value thresholds that are adaptive to the covariate information.

This seemingly natural modus operandi, which involves using p-values as the basic building blocks, however, is suboptimal. For the most commonly tested two-sided hypotheses, the p-values are typically formed via a data reduction step, which applies a non-bijective transformation to "primary" test statistics such as the z-values, t-statistics (Ritchie et al., 2015) or Wald statistics (Love et al., 2014). Sun and Cai (2007) and Storey et al. (2007) argued that reducing z-values to p-values may lead to substantial loss of information. A main thrust of this article is to reveal a new source of information loss in the context of covariate-adaptive multiple testing, and to develop a z-value covariate-adaptive (ZAP) methodology that bypasses the data reduction step. As illustrated in Section 2.2, the interactive relationship between the z-values and the covariates can capture structural information that can be exploited for more testing power. However, this interactive information may be undercut, and in some scenarios, completely forgone when converting z-values to p-values. Hence, the data reduction step not only leads to a loss of information carried by the main statistics, but also undermines the ability of the covariates in assisting with the FDR inference.

Few works on covariate-adaptive testing have pursued the z-value direction since combining the z-values and covariates poses an additional layer of challenges. Existing z-value based procedures either make strong assumptions on the underlying model (Scott et al., 2015), or are not robust for handling multi-dimensional covariate data (Cai et al., 2019). By contrast, ZAP retains the merits of z-value based methods and avoids information loss, neither relying on strong assumptions nor forgoing robustness. It faithfully preserves the interactive structure and effectively incorporates both the primary statistics and covariates into inference. ZAP is deployed with a working model, whose potential misspecification will not invalidate the FDR control. The proposed methodology fundamentally departs from p-values based methods by sidestepping the information loss occurred in forming the p-values.

Our contribution is twofold. First, ZAP represents a z-value based, covariate-adaptive testing framework that attains state-of-the-art power performance under minimal assumptions, filling an important gap in the literature. Particularly, we propose the first z-value based procedure with finite-sample guarantee on FDR control. Second, in light of a plethora of p-value based covariate-adaptive methods that have emerged in recent years, our study explicates new sources of information loss in data processing, which provides new insights and gives caveats for conducting covariate-adaptive inference in practical settings.

The rest of the paper is structured as follows. Section 2 states the problem formulation and describes the high-level ideas of ZAP. Section 3 formally introduces our two data-driven methods of ZAP and their implementation details. Numerical results based on both simulated and real data are presented in Section 4. Section 5 concludes the article with a discussion of extensions and open issues.

2. Problem Formulation and Basic Framework

2.1. **The problem statement.** Suppose we are interested in making inference of m real-valued parameters μ_i, $i = 1, \ldots, m$, and for each i, we observe a primary statistic $Z_i \in \mathbb{R}$ ("z-value") and an auxiliary covariate $X_i \in \mathbb{R}^p$ that can be multivariate. We consider a multiple testing problem where the goal is to identify nonzero effects or determine the values of the indicators

$$H_i \equiv I(\mu_i \neq 0) = \begin{cases} 1 & \text{if } \mu_i \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.1}$$

Assume that the triples $\{H_i, Z_i, X_i\}_{i=1}^m$ are independent and identically distributed, and the data are described by the following mixture model:

$$Z_i | X_i = x \sim f_x(z) \equiv f(z|x) = (1 - w_x) f_0(z) + w_x f_{1,x}(z), \tag{2.2}$$

where $w_x \equiv P(H_i = 1 | X_i = x)$ is the conditional probability of having a non-zero effect given $X_i = x$ and $f_{1,x}(z) \equiv f(z|H_i = 1, X_i = x)$ is the conditional density under the alternative. f_0 denotes the null density, which is invariant to the covariate value. In this article we assume $f_0(z) \equiv \phi(z)$, the density of a $\mathcal{N}(0,1)$ variable[1]. In contrast with the model in Scott et al. (2015) that assumes a fixed alternative density, i.e. $f_{1,x} \equiv f_1$, Model (2.2) provides a more general framework for multiple testing with covariates by allowing both w_x and $f_{1,x}(z)$ to vary in x.

Let $\mathcal{R} \subset \{1, \ldots, m\}$ be the set of hypotheses rejected by a multiple testing procedure. In large-scale testing problems, the widely used FDR is defined as

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right],$$

where $V = \sum_{i=1}^m (1 - H_i) I(i \in \mathcal{R})$ and $R = \sum_{i=1}^m I(i \in \mathcal{R})$ are respectively the number of false positives and the number of rejections. Throughout, $\mathbb{E}[\cdot]$ denotes an expectation operator with respect to the joint distribution of $\{H_i, Z_i, X_i\}_{i=1}^m$. The ratio $V/(R \vee 1)$ is known as the false discovery proportion (FDP). The power of a testing procedure can be evaluated using the expected number of true discoveries $\text{ETD} = \mathbb{E}[R - V]$ or the true positive rate

$$\text{TPR} = \mathbb{E}\left[\frac{R - V}{(\sum_{i=1}^m H_i) \vee 1}\right].$$

Our goal is to devise a powerful procedure that can control the FDR under a pre-specified level $\alpha \in (0, 1)$.

2.2. **Information loss in covariate-adaptive testing.** A two-sided p-value is formed by the non-bijective transformation $P_i = 2\Phi(-|Z_i|)$, where Φ is the cumulative distribution function of a $\mathcal{N}(0,1)$ variable. We call a testing procedure *z-value based* if it makes rejection decisions based on the full dataset $\{Z_i, X_i\}_{i=1}^m$, and *p-value based* if it does so only based on the reduced dataset $\{P_i, X_i\}_{i=1}^m$. This section presents examples to illustrate that the interactive structure between Z_i and

[1]This can be easily achieved via the composite transformation $\Phi^{-1} \circ G_0(\cdot)$ if the primary test statistic has a known null distribution function $G_0(\cdot)$, e.g. a t-distribution, where $\Phi(\cdot)$ is the standard normal distribution function.

X_i may not be preserved by transforming into p-values. The associated information loss leads to decreased power in the FDR inference.

Consider Model (2.2), and suppose $X_i \sim \text{Unif}(-1, 1)$. Our study examines three situations:

Example 2.1 Asymmetric alternatives: $f(z|x) = \frac{8-x}{10} f_0(z) + \frac{x+2}{10}\phi(z - 1.5)$.
Example 2.2 Unbalanced covariate effects on the non-null proportions:

$$f(z|x) = 0.8f_0(z) + \frac{1-x}{10}\phi(z + 1.5) + \frac{1+x}{10}\phi(z - 1.5).$$

Example 2.3 Unbalanced covariate effects on the alternative means:

$$f(z|x) = 0.9f_0(z) + 0.1\phi(z - 1.5\ \text{sgn}(x)),\ \text{where sgn}(x) = I(x \geqslant 0) - I(x < 0).$$

We investigate two approaches to FDR analysis for these examples that respectively reject hypotheses with suitably small posterior probabilities $\{P(H_i = 0|P_i, X_i)\}_{i=1}^m$ and $\{P(H_i = 0|Z_i, X_i)\}_{i=1}^m$. The latter probabilities are assumed to be known by an oracle[2]. In the literature the z-value based quantity $P(H_i = 0|Z_i, X_i)$ is also called the conditional local false discovery rate (CLfdr, Cai and Sun, 2009, Efron, 2008). It is known that the *optimal* p-value and z-value based procedures, which maximize true discoveries subject to false discovery constraints, have the respective forms

$$\boldsymbol{\delta}^{\mathcal{P}} = \big\{I[P(H_i = 0|P_i, X_i) \leqslant t_{\mathcal{P}}]\big\}_{i=1}^m \text{ and } \boldsymbol{\delta}^{\mathcal{Z}} = \big\{I[P(H_i = 0|Z_i, X_i) \leqslant t_{\mathcal{Z}}]\big\}_{i=1}^m,$$

where the rejection decisions are expressed by indicators, and the thresholds $t_{\mathcal{P}}$ and $t_{\mathcal{Z}}$ are calibrated such that the nominal FDR level is exactly α; see Appendix A.1 for a review. In our comparisons we choose suitable thresholds such that the FDR of both methods is exactly 0.1, and their powers are reported as TPR empirically computed by 150 repeated experiments for $m = 1000$:

Example 2.1: $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{P}}} = 4.4\%$; $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{Z}}} = 11.7\%$.
Example 2.2: $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{P}}} = 3.4\%$; $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{Z}}} = 5.5\%$.
Example 2.3: $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{P}}} = 0.6\%$; $\text{TPR}_{\boldsymbol{\delta}^{\mathcal{Z}}} = 2.6\%$.

Apparently, $\boldsymbol{\delta}^{\mathcal{Z}}$ is more powerful than $\boldsymbol{\delta}^{\mathcal{P}}$.

To understand the differences in power, we first remark that either oracle procedure essentially amounts to one by which i is rejected if and only if

$$Z_i \in \mathcal{S}(X_i) \subset \mathbb{R} \tag{2.3}$$

for some rejection region $\mathcal{S}(\cdot)$ on the z-value scale that is a function of the covariate value; the theoretical derivation is sketched in Appendix A.2. Let $\mathcal{S}^{\mathcal{P}}(x)$ and $\mathcal{S}^{\mathcal{Z}}(x)$ denote the respective rejection regions of $\boldsymbol{\delta}^{\mathcal{P}}$ and $\boldsymbol{\delta}^{\mathcal{Z}}$ on the z-value scale for a given covariate value x, which are plotted for the three examples in Figure 2.1. On the left panel, both $\mathcal{S}^{\mathcal{P}}(x)$ and $\mathcal{S}^{\mathcal{Z}}(x)$ enlarge as the covariate value increases, suggesting that the covariates are informative for both methods. The information loss leading to the lesser power of $\boldsymbol{\delta}^{\mathcal{P}}$ in Example 2.1 is intrinsically within the main

[2]In practice these posterior probabilities are unknown.

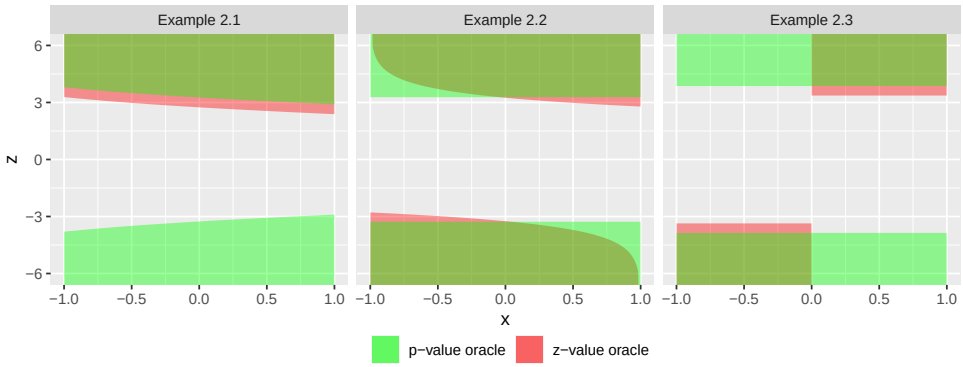

FIGURE 2.1. Comparisons of rejection regions. For each covariate value $x \in (-1, 1)$, the rejection regions $\mathcal{S}^{\mathcal{P}}(x)$ and $\mathcal{S}^{\mathcal{Z}}(x)$ are respectively marked in green and red on the z-value scale. The overlapped region is depicted in yellow.

statistics when converting z-values to p-values (Storey et al., 2007, Sun and Cai, 2007). By contrast, the middle and right panels show that $\mathcal{S}^{\mathcal{Z}}(x)$ changes with x, while $\mathcal{S}^{\mathcal{P}}(x)$ is completely insensitive to the changes in x; see Appendix A.2 for the relevant calculations. Hence, the covariates are only informative for $\boldsymbol{\delta}^{\mathcal{Z}}$. This fundamental phenomenon reveals that upon reduction to p-values, the information loss not only can occur internally within the main statistics, but also externally due to the failure of $\boldsymbol{\delta}^{\mathcal{P}}$ in fully capturing the original interactive information between Z_i and X_i. When the latter interactive structure represents the bulk of the information provided by the covariates for testing, reduction to p-values can substantially undermine the covariates' ability to assist with inference.

2.3. **The ZAP framework and a preview of contributions.** The previous examples motivate us to focus on z-value adaptive procedures to avoid information loss. This naturally boils down to pursuing the oracle procedure $\boldsymbol{\delta}^{\mathcal{Z}}$ in some shape or form, which presents unique challenges. Existing z-value based works such as Scott et al. (2015) and Cai et al. (2019) are built directly upon $\{P(H_i = 0|Z_i, X_i)\}_{i=1}^m$, the CLfdr statistics, which unfortunately involves unknown quantities that can be difficult to estimate in the presence of covariates. Commonly used algorithms may not produce desired estimates, and even lead to invalid FDR procedures if the modelling assumptions are violated. That the theory on FDR control critically depends on the quality of these estimates has greatly limited the scope and applicability of these works.

We aim to develop a new class of z-value adaptive (ZAP) procedures that are assumption-lean, robust and capable of effectively exploiting the interactive information between Z_i and X_i. The key idea is to emulate the oracle procedure $\boldsymbol{\delta}^{\mathcal{Z}}$ while circumventing the direct estimation of $P(H_i = 0|Z_i, X_i)$. Next we first outline the key steps (ranking and thresholding) of our framework and then provide a preview of its contributions.

In the first *ranking* step, we introduce the new concept of *assessor function*, which can be estimated from the data based on a working model, to construct a new sequence of significance indices $\{T_i\}_{i=1}^m$ as proxies for $\{P(H_i = 0|Z_i, X_i)\}_{i=1}^m$. While many potential working models can be used, in this work we focus on a class of beta-mixture models that are carefully defined on a bijective transformation of the z-values and particularly suitable for two-sided testing (Section 3.2). In the second *thresholding* step, ZAP calibrates a threshold along the ranking produced by $\{T_i\}_{i=1}^m$. The essential idea is to count the number of false rejections by any candidate threshold value with a "mirroring" sequence of the rejected significance indices, which can be created via either simulation (Algorithm 1) or partial data masking (Algorithm 2). The key strength of ZAP over the methods in Scott et al. (2015) and Cai et al. (2019) is that it seeks to emulate the oracle z-value procedure while avoiding a direct substitution of $P(H_i|Z_i, X_i)$ with its estimate. ZAP is assumption-lean and robust in the sense that it is provably valid for FDR control under model misspecifications, and the choice of the working model only affects the power. We stress that the resulting rejection regions of ZAP significantly depart from those of p-value adaptive methods, including the closely related CAMT (Zhang and Chen, 2020) and AdaPT (Lei and Fithian, 2018). Our simulation and real data studies show that the efficiency gain can be substantial.

3. Data-Driven ZAP Procedures

This section develops the framework of ZAP and its data-driven algorithms for covariate-adaptive FDR inference. Section 3.1 introduces the concept of assessor function and a prototype procedure inspired by the oracle z-value procedure. The assessor function can be constructed based on a working beta-mixture model, which is proposed in Section 3.2. Sections 3.3 and 3.4 lay out two variants of data-driven ZAP procedures and establish their theoretical properties. Further implementation details are discussed in Section 3.5.

3.1. Preliminaries: oracle z-value procedure, assessor function and a prototype ZAP algorithm.

To facilitate the development of a working model, we consider the following lossless transformation: $U_i = \Phi(Z_i)$. The transformed statistic U_i is referred to as a *u-value*[3], which, according to (2.2), obeys the induced mixture model

$$U_i|X_i = x \sim h_x(u) \equiv h(u|x) = (1 - w_x)h_0(u) + w_x h_{1,x}(u), \tag{3.1}$$

with $h_0(u)$ and $h_{1,x}(u) \equiv h(u|H_i = 1, X_i = x)$ respectively being the null Unif$(0, 1)$ and conditional alternative densities. An optimal FDR procedure (Cai and Sun, 2009, Heller and Rosset, 2021, Sun and Cai, 2007) is a thresholding rule based on the conditional local false discovery rates (CLfdr)

$$\text{CLfdr}_i \equiv P(H_i = 0|Z_i, X_i) = P(H_i = 0|U_i, X_i) = \frac{(1 - w_{X_i})h_0(U_i)}{h_{X_i}(U_i)}, i = 1, \ldots, m. \tag{3.2}$$

[3]Despite the similarity in their constructions, the u-values should *not* be treated as p-values for one-sided tests, which are not the subject matter of this work.

Since each CLfdr_i is a function of U_i conditional on X_i, we let $\text{CLfdr}_x(u) : (0,1) \to (0,1)$ be the corresponding function defined on the u-value scale for a given realized covariate value x. Related data-driven CLfdr procedures involve first estimating the CLfdr statistics, and second determining a threshold for them using, for example, step-wise algorithms (Sun and Cai, 2007), randomized rules (Basu et al., 2018) or linear programming (Heller and Rosset, 2021). However, the first estimation step poses significant challenges as it boils down to a hard *density regression* problem (Dunson et al., 2007). For example, to estimate $f_x(\cdot)$ (or equivalently $h_x(\cdot)$), a line of works (Deb et al., 2021, Scott et al., 2015, Tansey et al., 2018) proceeds by assuming a fixed alternative density, i.e.

$$f_{1,x}(z) \equiv f_1(z), \tag{3.3}$$

to make way for the application of an EM algorithm. If the assumption fails to hold, the CLfdr statistics can be poorly estimated and lead to both invalid FDR control and adversely affected power. The non-parametric CARS procedure developed in Cai et al. (2019) does not require the assumption in (3.3). However, it still employs the CLfdr statistics as its basic building blocks, which are estimated with kernel density methods. Due to the curse of dimensionality, the methodology becomes unstable in the presence of multivariate covariates, which has limited its applicability. For example, the real data analyses considered in Section 4.2 requires handling up to six (expanded) covariates.

By contrast, ZAP strives to sensibly emulate the oracle procedure without heavy reliance on the quality of the CLfdr estimates, which is its key strength. To motivate our data-driven procedures in the next sections, we shall first discuss a prototype ZAP procedure to illustrate two key steps of our testing framework: (a) how to combine Z_i (or equivalently U_i) and X_i for assessing the significance of hypotheses; and (b) how to threshold the new significance indices.

Step (a) involves the construction of an *assessor function*[4] $a_x(u) : (0,1) \to (0,1)$, which seeks to approximate the $\text{Clfdr}_x(u)$ function to integrate the information in both the u-value and covariate. For the present assume that $a_x(\cdot)$ is pre-determined. Let $T_i \equiv a_{X_i}(U_i)$ be the new significance index for i and $c_i(t) \equiv P(T_i \leqslant t | H_i = 0)$ be its null distribution. Assume that $c_i(\cdot)$ is continuous and strictly increasing[5], and denote its inverse by $c_i^{-1}(\cdot)$. All hypotheses will then be ordered according to the T_i's, with a smaller T_i indicating a more significant hypothesis.

In Step (b), we aim to determine a threshold for the T_i's to control the FDR. This involves the construction of a conservative FDP estimator for any candidate threshold t by the Barber-Candes (BC) method (Arias-Castro et al., 2017, Barber and Candès, 2015):

$$\widehat{\text{FDP}}(t) \equiv \frac{1 + \#\{i : S_i \geqslant 1 - c_i(t)\}}{\#\{i : T_i \leqslant t\} \vee 1} = \frac{1 + \#\{i : T_i^{\mathfrak{m}} \leqslant t\}}{\#\{i : T_i \leqslant t\} \vee 1}, \tag{3.4}$$

[4]Or simply known as an assessor.

[5]Both are true as the consequences of the way we will construct $a_x(\cdot)$; see the discussion after Lemma C.2 in Appendix C.

where, given that i is a true null, $S_i \equiv c_i(T_i)$ is the probability of realizing a smaller significance index and $T_i^{\mathfrak{m}} \equiv c_i^{-1}(1 - S_i)$ is the *mirror statistic* that "reflects" T_i's position in the distribution c_i. Define

$$\hat{t}(\alpha) \equiv \max\{t \in (0, t_{\max}] : \widehat{\mathrm{FDP}}(t) \leqslant \alpha\}, \tag{3.5}$$

where $t_{\max} \equiv \max\{t : c_i(t) \leqslant 0.5 \text{ for all } i\}$. It follows from Barber and Candès (2019, Lemma 1) that a procedure that rejects i whenever $T_i \leqslant \hat{t}(\alpha)$ controls the FDR at level α; see Appendix B. Importantly, the FDR is controlled under the desired level α whether $a_x(\cdot)$ is a good approximation of $\mathrm{CLfdr}_x(\cdot)$ or not.

However, the assessor $a_x(\cdot)$, which is taken as pre-determined thus far, is to be estimated from the observed data in practice. This leads to additional difficulties in both methodological and theoretical developments; for one thing, the theory in Barber and Candès (2019) cannot be directly applied to prove the FDR controlling property. Section 3.2 discusses a working beta-mixture model, whose parameters can be estimated from the observed data and subsequently used to construct a data-driven assessor $\hat{a}_x(\cdot)$. From there we can test the hypotheses in a data-driven manner by either implementing the prototype procedure directly using $\hat{a}_x(\cdot)$ as if it is pre-determined (Section 3.3), or mimic the prototype procedure in a more nuanced manner by leveraging the partial data masking technique in Lei and Fithian (2018) (Section 3.4). These two variants of ZAP entail different techniques to quantify the uncertainties in $\hat{a}_x(\cdot)$, with each having its relative strength and weakness: the direct approach offers asymptotic FDR control under suitable regularity conditions, and is both computationally and power efficient, while the data masking approach offers finite-sample FDR control but is computationally intensive and moderately less powerful in practice.

3.2. **A beta-mixture model.** We now develop a working model to approximate Model (3.1), which will be subsequently used to construct the assessor. We propose to capture the overall shape of $h_x(u)$ using a three-component mixture:

$$h_x(u) = (1 - \pi_{l,x} - \pi_{r,x})h_0(u) + \pi_{l,x}h_{l,x}(u) + \pi_{r,x}h_{r,x}(u), \tag{3.6}$$

where, given $X_i = x$, $\pi_{l,x}$ and $\pi_{r,x}$ respectively denote the mixing probabilities that $\mu_i < 0$ and $\mu_i > 0$ [6], and $h_{l,x}$ and $h_{r,x}$ respectively represent the densities of the negative and positive effects (on the left and right sides of the null). Our working model assumes that $\pi_{l,x}$ and $\pi_{r,x}$ are multinomial probabilities with regression parameter vectors θ_l and θ_r:

$$\pi_{l,x} = \frac{\exp(\tilde{x}^T\theta_l)}{1 + \exp(\tilde{x}^T\theta_r) + \exp(\tilde{x}^T\theta_l)},\ \pi_{r,x} = \frac{\exp(\tilde{x}^T\theta_r)}{1 + \exp(\tilde{x}^T\theta_r) + \exp(\tilde{x}^T\theta_l)},$$

where $\tilde{x} = (1, x^T)^T$ is the intercept-augmented covariate vector. Further, $h_{l,x}$ and $h_{r,x}$ are chosen to be beta densities with regression parameters β_l and β_r:

$$h_{l,x}(u) = \frac{1}{B(k_{l,x}, \gamma_l)}u^{k_{l,x}-1}(1-u)^{\gamma_l-1}, h_{r,x}(u) = \frac{1}{B(\gamma_r, k_{r,x})}u^{\gamma_r-1}(1-u)^{k_{r,x}-1},$$

[6]The different symbols $\pi_{l,x}$ and $\pi_{r,x}$ are used in the working model. In the true data generating model (3.1), the mixing probability is denoted w_x.

where $k_{l,x} = \{1 + \exp(-\tilde{x}^T\beta_l)\}^{-1}$ and $k_{r,x} = \{1 + \exp(-\tilde{x}^T\beta_r)\}^{-1}$, for two *fixed* shape parameters γ_l and γ_r. $h_{l,x}$ and $h_{r,x}$ are respectively left-leaning (right-skewed) and right-leaning (left-skewed) functions. We require that $\gamma_l > 2$ and $\gamma_r > 2$ to ensure that both are strictly monotone and *convex*, and thus provide a reasonable approximation to the underlying true density in practice; see Lemma C.1 in Appendix C for a precise result. The exact choices for $\{\gamma_l, \gamma_r\}$ will be further discussed in Section 3.5. The working model may be generalized to capture non-linearity in x using, say, spline functions.

Beta mixtures have long been identified as a flexible modeling tool for variables taking values in the unit interval; see Ferrari and Cribari-Neto (2004), Ji et al. (2005), Markitsis and Lai (2010), Migliorati et al. (2018), Parker and Rothenberg (1988), Pounds and Morris (2003) for related works. In the context of covariate-adaptive multiple testing, Lei and Fithian (2018) and Zhang and Chen (2020) employ a two-component beta-mixture model for the p-values that consists of a uniform and another left-leaning beta component. Our working model defined on the u-value scale can be viewed as a natural extension of these works to capture important patterns in the u-value distribution associated with two-sided covariate-adaptive testing.

The assessor can be constructed as the $\mathrm{CLfdr}_x(\cdot)$ function with respect to our working model (3.6). Since $h_0 \equiv 1$, it follows that

$$a_x(u) \equiv \frac{1 - \pi_{l,x} - \pi_{r,x}}{1 - \pi_{l,x} - \pi_{r,x} + \pi_{l,x}h_{l,x}(u) + \pi_{r,x}h_{r,x}(u)}, \quad 0 < u < 1. \tag{3.7}$$

The corresponding data-driven assessor is denoted by $\hat{a}_x(u)$ if the parameters $\{\theta_l, \theta_r, \beta_l, \beta_r\}$ are estimated from the data for its construction.

3.3. **Asymptotic ZAP.** We now develop a direct data-driven version of the prototype algorithm in Section 3.1. To construct $\hat{a}_x(u)$, we first obtain the maximum likelihood estimates (MLE) of the unknown regression parameters $\{\theta_l, \theta_r, \beta_l, \beta_r\}$ with the data $\{U_i, X_i\}_{i=1}^m$; the EM algorithm for their computations are provided in Appendix F.1. Denote $\hat{T}_i \equiv \hat{a}_{X_i}(U_i)$, and let $\hat{c}_i(\cdot)$ be its null distribution by treating $\hat{a}_{X_i}(\cdot)$ as if it is pre-determined. With $\hat{S}_i \equiv \hat{c}_i(\hat{T}_i)$, the estimated mirror statistics are correspondingly defined as $\hat{T}_i^{\mathfrak{m}} \equiv \hat{c}_i^{-1}(1 - \hat{S}_i)$, which can be computed numerically by performing quantile estimation. The FDP for a candidate threshold t can be estimated as

$$\widehat{\mathrm{FDP}}_{asymp}(t) \equiv \frac{1 + \#\{i : \hat{T}_i^{\mathfrak{m}} \leqslant t\}}{\#\{i : \hat{T}_i \leqslant t\} \vee 1}. \tag{3.8}$$

Define $\hat{t}_{asymp}(\alpha) \equiv \sup\{0 \leqslant t \leqslant 1 : \widehat{\mathrm{FDP}}_{asymp}(t) \leqslant \alpha\}$, and reject i whenever $\hat{T}_i \leqslant \hat{t}_{asymp}(\alpha)$. In practice, it suffices to consider only the values of $\hat{T}_1, \ldots, \hat{T}_m$ as candidate thresholds. This procedure is summarized in Algorithm 1.

The main theory requires the following classical assumption from the literature on misspecified models (White, 1981, 1982):

Assumption 1 (Existence of a unique maximizer)**.** *The expected log-likelihood*

$$\mathbb{E}\log[(1 - \pi_{l,X_i} - \pi_{r,X_i}) + \pi_{l,X_i}h_{l,X_i}(U_i) + \pi_{r,X_i}h_{r,X_i}(U_i)]$$

Algorithm 1: Asymptotic ZAP

1 Construct $\hat{a}_{X_i}(\cdot)$'s using the MLEs obtained via the EM algorithm in Appendix F.1 and compute $\hat{T}_i = \hat{a}_{X_i}(U_i)$ for each i.
2 Compute the mirror statistics $\{\hat{T}_i^{\mathtt{m}}\}_{i=1}^m$:
(i) Generate i.i.d. realizations $u_1, \ldots, u_N$ from Unif$(0,1)$ for a large N.
(ii) For each i, evaluate $\hat{a}_{X_i}(u_1), \ldots, \hat{a}_{X_i}(u_N)$ to simulate the null distribution $\hat{c}_i(\cdot)$. Compute $\hat{T}_i^{\mathtt{m}}$ via e.g. `quantile()` in `R`.
3 Order $\{\hat{T}_i\}_{i=1}^m$ as $\hat{T}_{(1)} \leqslant \cdots \leqslant \hat{T}_{(m)}$. Reject i if $T_i \leqslant \hat{T}_{(k)}$, where $k = \max\left\{l \in \{1, \ldots, m\} : \frac{1+\#\{i:\hat{T}_i^{\mathtt{m}} \leqslant \hat{T}_{(l)}\}}{l \vee 1} \leqslant \alpha\right\}$.

of the beta-mixture model (3.6) *has a unique maximum at* $\{\theta^*, \beta^*\}$ *over* $\theta \in \Theta$ *and* $\beta \in \boldsymbol{B}$ *for compact spaces* Θ *and* $\boldsymbol{B}$*, where* $\theta \equiv (\theta_l^T, \theta_r^T)^T$ *and* $\beta \equiv (\beta_l^T, \beta_r^T)^T$*. The expectation is taken with respect to the true joint distribution of* $\{H_i, Z_i, X_i\}$*.*

Together with Assumptions 2 - 3 in Appendix D.1, which are standard regularity and strong-law conditions, we can prove the following asymptotic FDR controlling property.

Theorem 3.1. *Let* $\hat{a}_x(\cdot)$ *be constructed with the MLE* $\{\hat{\theta}, \hat{\beta}\} = \underset{\theta,\beta}{\operatorname{argmax}} \sum_{i=1}^m \log h_{X_i}(U_i)$ *of the beta-mixture model* (3.6)*. Under Assumptions 1-3, the procedure that rejects* i *whenever* $\hat{T}_i \leqslant \hat{t}_{asymp}(\alpha)$ *controls the FDR asymptotically in the sense that* $\limsup_{m\to\infty} FDR \leqslant \alpha$*.*

We highlight two aspects of this result. First, it doesn't require the estimated assessor function to be a good proxy for $\mathrm{CLfdr}_x(\cdot)$. Hence, its theory is more attractive than that of Cai et al. (2019), which requires consistent CLfdr estimates to ensure asymptotic FDR control. Second, to establish Glivenko-Cantelli results (Lemma D.5) for the following three empirical processes

$$m^{-1}\sum_{i=1}^m I\left(\hat{T}_i \leqslant t\right), \quad m^{-1}\sum_{i=1}^m (1-H_i) I\left(\hat{T}_i \leqslant t\right), \quad m^{-1}\sum_{i=1}^m I\left\{\hat{S}_i \geqslant 1 - \hat{c}_i(t)\right\},$$

typical of similar asymptotic analyses (Storey et al., 2004, Zhang and Chen, 2020), we heavily utilize the concavity properties (Lemma C.2) of the functional form in (3.7) to uniformly control the deviations of the estimated assessors $\hat{a}_{X_i}(\cdot)$ from the assessors $a^*_{X_i}(\cdot)$ constructed with the population parameters $\{\theta^*, \beta^*\}$; the delicate techniques involved may be of independent interest.

3.4. **Finite-sample ZAP.** This section introduces an alternative ZAP procedure that offers finite-sample control of the FDR. The operation again involves approximating the CLfdr statistics via an assessor function. However, the thresholding step is based on a more nuanced approach to FDP estimation inspired by the p-value method AdaPT (Lei and Fithian, 2018). In this approach, multiple testing is conducted in an iterative manner, where data are initially partially masked and then gradually revealed at steps $t = 0, 1, \ldots$, with the thresholds sequentially updated based on the revealed data at each step. In what follows, if $g_1(\cdot)$ and $g_2(\cdot)$

are two functions defined on the same space, $g_1 \preceq g_2$ means $g_1(x) \leqslant g_2(x)$ for all x in that space. If C is a constant, $g_1 \preceq C$ means $g_1(x) \leqslant C$ for all x. Similarly we can define $g_1 \succeq g_2$ and $g_1 \succeq C$.

Since the concavity[7] of the assessor functional form in (3.7) suggests that rejecting hypotheses with small values of $T_i = a_{X_i}(U_i)$ amounts to rejecting extreme u-values near 0 or 1, our iterative algorithm emulates this essential operational characteristic of the prototype procedure. We first divide the covariate values into a left and right group based on the observed u-values:

$$\mathcal{X}_l = \{X_i : U_i \leqslant 0.5\} \text{ and } \mathcal{X}_r = \{X_i : U_i > 0.5\}.$$

At each step $t = 0, 1, \ldots$, let $s_{l,t} : \mathcal{X}_l \to [0, 0.25]$ and $s_{r,t} : \mathcal{X}_r \to [0.75, 1]$ denote two corresponding thresholding functions, and define the candidate rejection set $\mathcal{R}_t \equiv \mathcal{R}_{l,t} \cup \mathcal{R}_{r,t}$, where

$$\mathcal{R}_{l,t} \equiv \{i : U_i \leqslant s_{l,t}(X_i) \wedge 0.5\} \text{ and } \mathcal{R}_{r,t} \equiv \{i : U_i \geqslant s_{r,t}(X_i) \vee 0.5\}. \tag{3.9}$$

Let $\mathcal{A}_t \equiv \mathcal{A}_{l,t} \cup \mathcal{A}_{r,t}$ be the corresponding set of "accepted" hypotheses, where

$$\mathcal{A}_{l,t} \equiv \{i : 0.5 - s_{l,t}(X_i) \leqslant U_i \leqslant 0.5\} \text{ and } \mathcal{A}_{r,t} \equiv \{i : 0.5 < U_i \leqslant 1.5 - s_{r,t}(X_i)\}.$$

Intuitively, $|\mathcal{A}_{l,t}|$ estimates the number of false rejections in the left candidate rejection set $\mathcal{R}_{l,t}$: Given $H_i = 0$ and $U_i \leqslant 0.5$, the events $\{U_i < s_{l,t}(X_i)\}$ and $\{U_i > 0.5 - s_{l,t}(X_i)\}$ are equally likely. The FDP of possibly rejecting $\mathcal{R}_t$ at step t can then be estimated as

$$\widehat{\mathrm{FDP}}_{finite}(t) = \frac{1 + |\mathcal{A}_t|}{|\mathcal{R}_t| \vee 1}. \tag{3.10}$$

If $\widehat{\mathrm{FDP}}_{finite}(t) \leqslant \alpha$, the algorithm terminates and the hypotheses in $\mathcal{R}_t$ are rejected. Otherwise, the algorithm proceeds to the next step $t + 1$ and updates the two thresholding functions under two restrictions. First, it must be that $s_{l,t+1} \preceq s_{l,t}$ and $s_{r,t+1} \succeq s_{r,t}$; this ensures that $\mathcal{R}_t$ shrinks in size as t increases. Second, $s_{l,t+1}$ and $s_{r,t+1}$ must be updated based on the knowledge of $|\mathcal{R}_t|$, $|\mathcal{A}_t|$ and the partially masked data $\{\tilde{U}_{t,i}, X_i\}_{i=1}^m$ only, where

$$\tilde{U}_{t,i} \equiv \begin{cases} U_i & \text{if } U_i \notin \mathcal{A}_t \cup \mathcal{R}_t \\ \{\breve{U}_i, U_i\} & \text{if } U_i \in \mathcal{A}_t \cup \mathcal{R}_t \end{cases} \tag{3.11}$$

is a singleton or a two-element set depending on whether i is in the "masked" set $\mathcal{A}_t \cup \mathcal{R}_t$, and $\breve{U}_i$ is the "reflection" of U_i about the "middle" axis at $u = 0.25$ or $u = 0.75$, depending on which group (left or right) U_i belongs to:

$$\breve{U}_i \equiv (1.5 - U_i)I(U_i > 0.5) + (0.5 - U_i)I(U_i \leqslant 0.5).$$

For example, if the underlying U_i is 0.1 and i is masked at step t, the algorithm can only update for $s_{lt+1}$and s_{rt+1} with the partial knowledge that U_i is either 0.1 or its reflection value 0.4. Algorithm 4 in Appendix F.2 describes one such updating scheme which applies an EM algorithm (Appendix F.3) acting only on the partially masked data to estimate the beta-mixture model (3.6). Figure 3.1

[7]Refer to Lemma C.2

illustrates how the data $\{U_i, X_i\}_{i=1}^m$ are partitioned into $\mathcal{A}_t$, $\mathcal{R}_t$ and the unmasked set $\{1, \ldots, m\} \backslash \{\mathcal{A}_t \cup \mathcal{R}_t\}$ at a given step t, based on Example 2.2 in Section 2.2. In particular, we remark that the algorithm cannot tell the true data point from a given red-pink (blue-cyan) pair in the plot (b) where the reflection points $\{\breve{U}_i\}_{i \in \mathcal{A}_t \cup \mathcal{R}_t}$ are also shown.

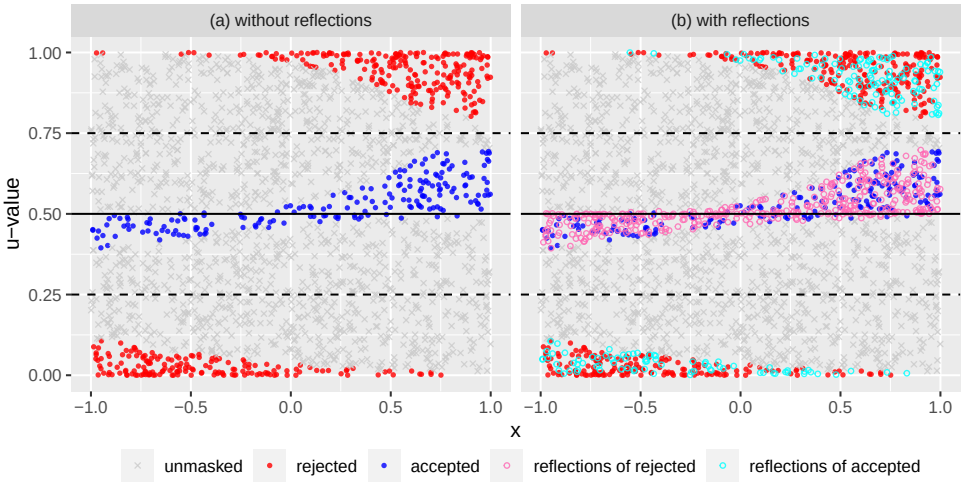

FIGURE 3.1. Illustration of Algorithm 2 at a step t based on Example 2.2, $m = 2000$. **(a)**: The red, blue and grey are points in the respective sets $\mathcal{R}_t$, $\mathcal{A}_t$ and $\{1, \ldots, m\} \backslash \{\mathcal{A}_t \cup \mathcal{R}_t\}$, where $\widehat{\mathrm{FDP}}_{finite}(t) = 193/389 \approx 0.5$. **(b)**: The reflections of the points in $\mathcal{R}_t$ and $\mathcal{A}_t$ are respectively shown in pink and cyan. A pink $\breve{U}_i$ below (above) 0.5 is the reflection of a red U_i with the same covariate value about the middle axis at $u = 0.25$ ($u = 0.75$) of the left (right) group; the cyan are the reflections of the blue.

The steps described above are summarized in Algorithm 2, whose finite-sample FDR controlling property is stated in Theorem 3.2.

Theorem 3.2 (Finite-sample FDR control)**.** *Under the conditions that (i) $s_{r,t+1} \geq s_{r,t}$ and $s_{l,t+1} \preceq s_{l,t}$ and (ii) $s_{r,t+1}$, and $s_{l,t+1}$ are updated based on $|\mathcal{R}_t|$, $|\mathcal{A}_t|$ and $\{\tilde{U}_{t,i}, X_i\}_{i=1}^m$ only, Algorithm 2 controls the FDR under α for finite samples. Specifically, we have*

$$\mathbb{E}\left[FDP \middle| \{H_i, X_i\}_{i=1}^m\right] \leqslant \alpha.$$

Lastly, we highlight a crucial difference between Algorithm 2 and AdaPT in the present context. Operating on the two-sided p-values, AdaPT proceeds iteratively with a single thresholding function s_t defined on $\{X_i\}_{i=1}^m$ such that $s_t \preceq 0.5$, and the ratio $\frac{1+|\{i:P_i \geqslant 1-s_t(X_i)\}|}{1 \vee |\{i:P_i \leqslant s_t(X_i)\}|}$ is used as an FDP estimator for the candidate rejection set $\{i : P_i \leqslant s_t(X_i)\}$. It is easy to see that

$$P_i \leqslant s_t(X_i) \iff U_i \leqslant s_t(X_i)/2 \text{ or } U_i \geqslant 1 - s_t(X_i)/2. \tag{3.12}$$

Algorithm 2: Finite-sample ZAP

Data: $\{U_i, X_i\}_{i=1}^m$
Input: FDR level α, initial thresholding functions $s_{l,0} \preceq 0.5$ and $s_{r,0} \succeq 0.5$;
for *t= 0,1 . . . ,* **do**
  Compute $\widehat{\mathrm{FDP}}_{finite}(t)$ in (3.10);
  if $\widehat{FDP}_{finite}(t) > \alpha$ **then**
    Update $s_{l,t+1}$ and $s_{r,t+1}$ while respecting the two conditions in Theorem 3.2. E.g. Apply Algorithm 4 in Appendix F.2;
  else
    Record $\mathcal{R}_t$; break;
  end
end
Output: Reject all hypotheses in $\mathcal{R}_t$.

Hence, on the u-value scale, AdaPT always adopts symmetric rejection regions about $u = 0.5$. By contrast, Algorithm 2 employs two different thresholding functions $s_{l,t}$ and $s_{r,t}$, which allow for asymmetric rejection regions, and therefore provides additional flexibility to fully capitialize on covariate information for two-sided tests. As seen in Figure 3.1, the pattern of the candidate rejection points in red agrees with the middle panel of Figure 2.1; as the covariate increases from -1 to 1, the algorithm's rejection priorities change from the u-values near 0 to those near 1.

3.5. **Implementation details.** An `R` package `zap` for our two data-driven methods is available on `https://github.com/dmhleung/zap`, and we shall discuss further details of their implementation.

For both data-driven procedures, the shape parameters $\{\gamma_l, \gamma_r\}$ of the working model need to be pre-specified before running the EM algorithms. While requiring $\gamma_l, \gamma_r > 2$ ensures a convex shape for the three-component beta-mixture density (Lemma C.1), we recommend choosing $(\gamma_l, \gamma_r) = (4, 4)$ as a default, which has yielded consistently good performance in our numerical studies.

To illustrate the effectiveness of our recommendation, we simulate 8000 i.i.d. z-values $Z_1, \ldots, Z_{8000}$ from the normal mixture model

$$0.78 f_0(z) + 0.15\phi(z + 1.5) + 0.07\phi(z - 2) \tag{3.13}$$

without any covariates. The histogram of the corresponding u-values is plotted in Figure 3.2(a), overlaid with the true underlying density function, as well as estimated densities of the beta mixture (3.6) fitted with regression intercepts only, where (γ_l, γ_r) is respectively fixed at $(1, 1)$ and $(4, 4)$. When modeling p-values with a two-component beta mixture, Lei and Fithian (2018) and Zhang and Chen (2020) set an analogous shape parameter to be 1, so $(\gamma_l, \gamma_r) = (1, 1)$ would be a seemingly natural choice to extend their model for two-sided tests. Both fitted densities visually coincide with the true density, attesting to the flexibility of beta mixtures for modeling data on the unit interval. However, the estimated component probabilities differ significantly for $(\gamma_l, \gamma_r) = (1, 1)$ vs $(\gamma_l, \gamma_r) = (4, 4)$. In Figure 3.2(b),

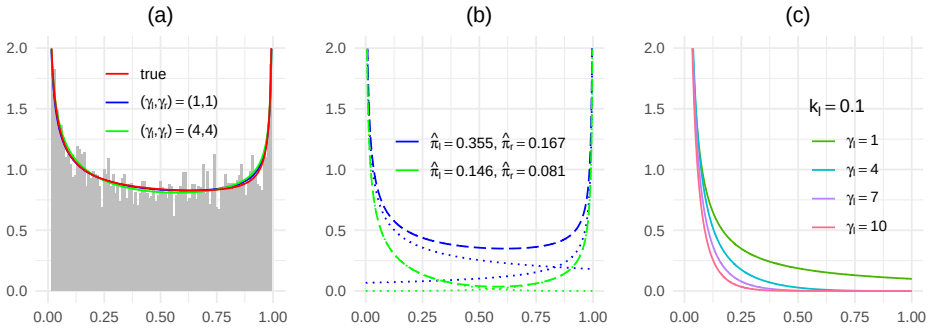

FIGURE 3.2. **(a)**: Histogram of $U_i = \Phi(Z_i)$ generated by (3.13). The red curve is the true density; while the blue and green curves respectively correspond to the estimated densities of our beta-mixture model with (γ_l, γ_r) set at $(1, 1)$ and $(4, 4)$. **(b)**: The long dashed blue curve is the entire non-null component of the solid blue estimated density in (a). The two dotted blue curves are the left-leaning and right-leaning non-null components that add up to the long dashed blue. The green curves are constructed analogously with respect to the solid green density in (a). The legend shows the estimated probabilities for the left and right-leaning components, where the subscript "x" is omitted from $\hat{\pi}_l$ and $\hat{\pi}_r$ as the fit uses intercepts only. **(c)**: Plot of the left-leaning beta density $B(k_l, \gamma_l)^{-1} u^{k_l - 1}(1 - u)^{\gamma_l - 1}$ for $k_l = 0.1$ and different values of γ_l.

we present the estimated quantities pertaining to the non-null components. We can see that setting $(\gamma_l, \gamma_r) = (1, 1)$ has drastically overestimated the left and right non-null probabilities, whereas setting $(\gamma_l, \gamma_r) = (4, 4)$ provides good approximations to the truths. To gain insight into why larger shape parameters are preferred, in Figure 3.2(c) we plot the density of a left-leaning beta density

$$B(k_l, \gamma_l)^{-1} u^{k_l - 1}(1 - u)^{\gamma_l - 1}$$

for different values of γ_l and a fixed $k_l = 0.1$, which supposedly captures the negative effects in two-sided tests. We can see that small values of γ_l tend to yield a density component that slants in the middle of the unit interval. As a result, when added to another right-leaning beta density for the positive effects with a similar but mirroring shape, it gives rise to an overall non-null density component with a large plateau in the middle of the interval $(0, 1)$ akin to the U-shaped blue curve in Figure 3.2(b). This inflates the non-null probability estimates. In contrast, larger values of γ_l and γ_r effectively mitigate the issue by rendering sharply convex non-null component densities like the purple in Figure 3.2(c), avoiding overestimation and leading to better approximation of the tails. More setups are experimented in Appendix G.1; the choice of $(\gamma_l, \gamma_r) = (4, 4)$ produces reasonable probability estimates throughout.

Other aspects of implementation are as follows. For the asymptotic method (Algorithm 1), since a large N allows us to compute the mirror statistics up to arbitrary precision, we evaluate at $N = 50000$ uniform realizations by default. For the iterative finite-sample method (Algorithm 2), we set the initial thresholding functions as $s_{l0} \equiv 0.2$ and $s_{r0} \equiv 0.8$, but other values close to 0.25 and 0.75 tend to be equally effective. We also update the thresholding functions every $\lceil m/100 \rceil$ steps. Ideally one would want to update at every step along the way to reveal the masked u-values sooner. However, it is more practical to carry out intermittent updates since the EM component involved in Algorithm 4 is computationally costly. Lastly, one can also perform feature selection at any step if X_i is multivariate, as long as it is done properly based on the masked data, akin to what was suggested by Lei and Fithian (2018, Section 4.2). We have not performed this step for simplicity.

4. Numerical studies

We conduct numerical studies to gauge the performance of ZAP alongside other methods on both simulated and real data. For expositional considerations, here we only limit the comparisons to a selection of representative FDR methods. This makes the ensuing graphs (Figures 4.1-4.2) less crowded by lines and easier to read. More methods in the literature are included to expand our studies in Appendix G.2, but the basic conclusions do not change. The methods being considered here are:

(a) ZAP (asymp): Algorithm 1 with specifications described in Section 3.5.
(b) ZAP (finite): Algorithm 2 with specifications described in Section 3.5.
(c) CAMT: the covariate-adaptive multiple testing method (Zhang and Chen, 2020).
(d) AdaPT: the adaptive p-value thresholding method (Lei and Fithian, 2018). Their working model is updated based on the EM algorithm for every $\lceil m/100 \rceil$ steps; other default specifications are chosen based on the R package `adaptMT`.
(e) IHW: independent hypothesis weighting method (Ignatiadis et al., 2016). We remark here that IHW can only handle univariate covariates.
(f) FDRreg: false discovery rate regression method (Scott et al., 2015). The theoretical null $\mathcal{N}(0, 1)$ has been used.

Among them, ZAP and FDRreg are z-valued based, while all other methods are p-value based.

4.1. **Simulation studies.** We simulate data to test $m = 5000$ hypotheses. Two-dimensional covariates $X_i = (X_{1i}, X_{2i})^T$, $i = 1, \ldots, m$, are independently generated from the bivariate normal distribution $\mathcal{N}\left\{ \binom{0}{0}, \left(\begin{smallmatrix} 1/2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right) \right\}$. Conditional on $X_i = x \equiv (x_1, x_2)^T$, Z_i is generated with a normal mixture density

$$(1 - w_{l,x} - w_{r,x}) f_0(z) + w_{l,x}\phi(z - \mu_{l,x}) + w_{r,x}\phi(z - \mu_{r,x}), \tag{4.1}$$

where $w_{l,x}$ and $w_{r,x}$ [8] are probabilities that control the sparsity levels of negative and positive effects, and $\mu_{l,x} < 0$ and $\mu_{r,x} > 0$ are negative and positive non-null

[8] These data generating probabilities $\{w_{l,x}, w_{r,x}\}$ should again be distinguished from $\{\pi_{l,x}, \pi_{r,x}\}$ in the working model (3.6).

normal means. The covariate-adjusted overall non-null density is then given by

$$f_{1,x}(z) = \frac{w_{l,x}\phi(z-\mu_{l,x}) + w_{r,x}\phi(z-\mu_{r,x})}{w_{l,x}+w_{r,x}}. \tag{4.2}$$

We shall allow $\{w_{l,x}, w_{r,x}, \mu_{l,x}, \mu_{r,x}\}$ to depend on x in different ways to induce the simulation setups below, which can be considered as more realistic versions of the stylized examples in Section 2.2. Note that the sum $X_{\bullet i} \equiv X_{1i} + X_{2i}$ of the covariate components is $\mathcal{N}(0,1)$-distributed, and $x_\bullet \equiv x_1 + x_2$ will denote a realized value of it in what follows.

Setup 1 (Asymmetric alternatives). The quantities in (4.1) are

$$w_{r,x} = \frac{1}{1+\exp(-\eta-\zeta x_\bullet)}, \quad \mu_{r,x} = \frac{2\varepsilon}{1+\exp(-\zeta x_\bullet)}, \quad w_{l,x} = 0, \quad \mu_{l,x} = 0,$$

with the simulation parameters ranging as

$$\zeta \in \{0, 0.5, 1\}, \quad \varepsilon \in \{1.3, 1.5, 1.7, 1.9, 2.1\} \text{ and } \eta = -2.$$

Since $w_{l,x} = 0$, all the non-null statistics come from the right centered alternative density $\phi(z-\mu_{r,x})$. We briefly explain the simulation parameters. ε is an effect size parameter. Generally, ζ controls the *informativeness* of the covariates in relation to both the non-null probabilities and alternative means: when $\zeta > 0$, a greater value of $x_\bullet$ makes the signals denser and stronger (i.e. $w_{r,x}$ and $\mu_{r,x}$ become larger). The value of η controls the sparsity levels. For example, when the covariates are non-informative at $\zeta = 0$, setting $\eta = -2$ yields a baseline signal proportion of roughly 12%, i.e. $w_{r,x} = w_{l,x} + w_{r,x} = 11.9\%$. Note that $f_{1,x}$ in (4.2) varies in x given the dependence of $\mu_{r,x}$ on x, so (3.3) is an invalid assumption.

Setup 2 (Unbalanced covariate effects on the non-null proportions). Let

$$w_{r,x} = \frac{\exp(\zeta x_\bullet)}{\exp(-\eta)+\exp(-\zeta x_\bullet)+\exp(\zeta x_\bullet)}, \; w_{l,x} = \frac{\exp(-\zeta x_\bullet)}{\exp(-\eta)+\exp(-\zeta x_\bullet)+\exp(\zeta x_\bullet)},$$

$\mu_{rx} = \varepsilon$ and $\mu_{lx} = -\varepsilon$. We fix $\eta = -2.5$ and vary other parameters in the range

$$\zeta \in \{0, 0.7, 1\}, \quad \varepsilon \in \{1.3, 1.5, 1.7, 1.9, 2.1\}.$$

Only $w_{l,x}$ and $w_{r,x}$ depend on the covariate value: for $\zeta > 0$, $w_{r,x}$ increases and $w_{l,x}$ decreases as $x_\bullet$ increases, and vice versa as $x_\bullet$ decreases. In consideration of (4.2), the conditional non-null density $f_{1,x}(z)$ will change sharply in shape from concentrating on negative z-values to concentrating on positive z-values as $x_\bullet$ increases from being negative to positive. This relationship provides important structural information which can be leveraged for enhancing the power. However, if one collapses the z-values into two-sided p-values, then the analogous conditional p-value density is less likely to capture drastic changes in $x_\bullet$, since both very negative and positive $x_\bullet$ can correspond to very small p-values, making the interactive relationship between the p-values and the covariates less pronounced. Intuitively, this would lead to power loss of p-value based methods. The choice of η corresponds to a baseline signal proportion of roughly 14% when $\zeta = 0$.

Setup 3 (Unbalanced covariate effects on the alternative means). Let

$$w_{r,x} = \frac{1/2}{1+\exp(-\eta)}, w_{l,x} = \frac{1/2}{1+\exp(-\eta)}, \mu_{r,x} = \frac{2\varepsilon}{1+\exp(-\zeta x_\bullet)}, \mu_{l,x} = \frac{-2\varepsilon}{1+\exp(\zeta x_\bullet)}.$$

The simulation parameters range as

$$\zeta \in \{0, 1.5, 3\}, \quad \varepsilon \in \{1.3, 1.5, 1.7, 1.9, 2.1\} \text{ and } \eta = -2.$$

Our choice of η corresponds to the a baseline signal proportion of roughly 12% when $\zeta = 0$. When the covariates are informative ($\zeta > 0$), $\mu_{r,x}$ and $\mu_{l,x}$ respectively become more positive and less negative as $x_\bullet$ increases. Such a directional relationship can be exploited by ZAP for improving the power. However, if one collapses the z-values into p-values, then under $H_i = 1$ both very positive and negative values of $X_{\bullet i}$ can imply a small P_i, and the interactive relationship between the main statistic P_i and auxiliary statistic X_i will be much weakened. Hence we expect ZAP to exhibit higher power than p-value based methods.

We apply the six methods at the nominal FDR level 0.05. Since IHW can only handle univariate covariates, it is applied with $X_{\bullet i}$, which is an effective summary covariate in all three setups. The simulation results are reported in Figure 4.1, where the empirical FDR and TPR levels of different methods are computed based on 150 repetitions. The following observations can be made:

(a) Asymptotic ZAP, depicted in blue, is in general more powerful than finite-sample ZAP, depicted in red. This is likely attributable to the latter's information loss from the "u-value masking" step. The advantage of the finite-sample ZAP is in its theoretical properties.

(b) Both asymptotic and finite-sample ZAP methods achieve state-of-the-art power performance in all three setups. The FDR levels are consistently controlled under the nominal level 0.05. Both ZAP methods demonstrate superior performances over the p-value based methods (CAMT, AdaPT and IHW). The gains in power become more substantial when the covariates become more informative.

(c) The covariate-adjusted non-null density (4.2) depends on x for all three setups, so FDRreg, which makes the conflicting assumption in (3.3), is possibly invalid for FDR control. Although FDRreg has comparable power to the ZAP methods in Setup 1, it overshoots the FDR bound of 0.05, and it can't match the power of ZAP in Setup 2 because the assumption $f_{1,x}(z) \equiv f_1(z)$ itself obstructs the interactive information between the z-values and the covariates to be utilized.

(d) In Setup 3, ZAP only has moderate power advantage over the other methods when the covariates are informative, but still, it has "salvaged" more power than others. In fact, it is shown in Appendix G.2 that Setup 3 poses a hard multiple testing problem, and admittedly the beta mixture may not be the most suitable working model for this data generating mechanism. In Section 5 we will discuss alternative working models to implement the ZAP methods.

4.2. **Real data.** This section investigates the performance of ZAP using several publicly available real datasets summarized in Table 1. Three data sets (`bottomly`,

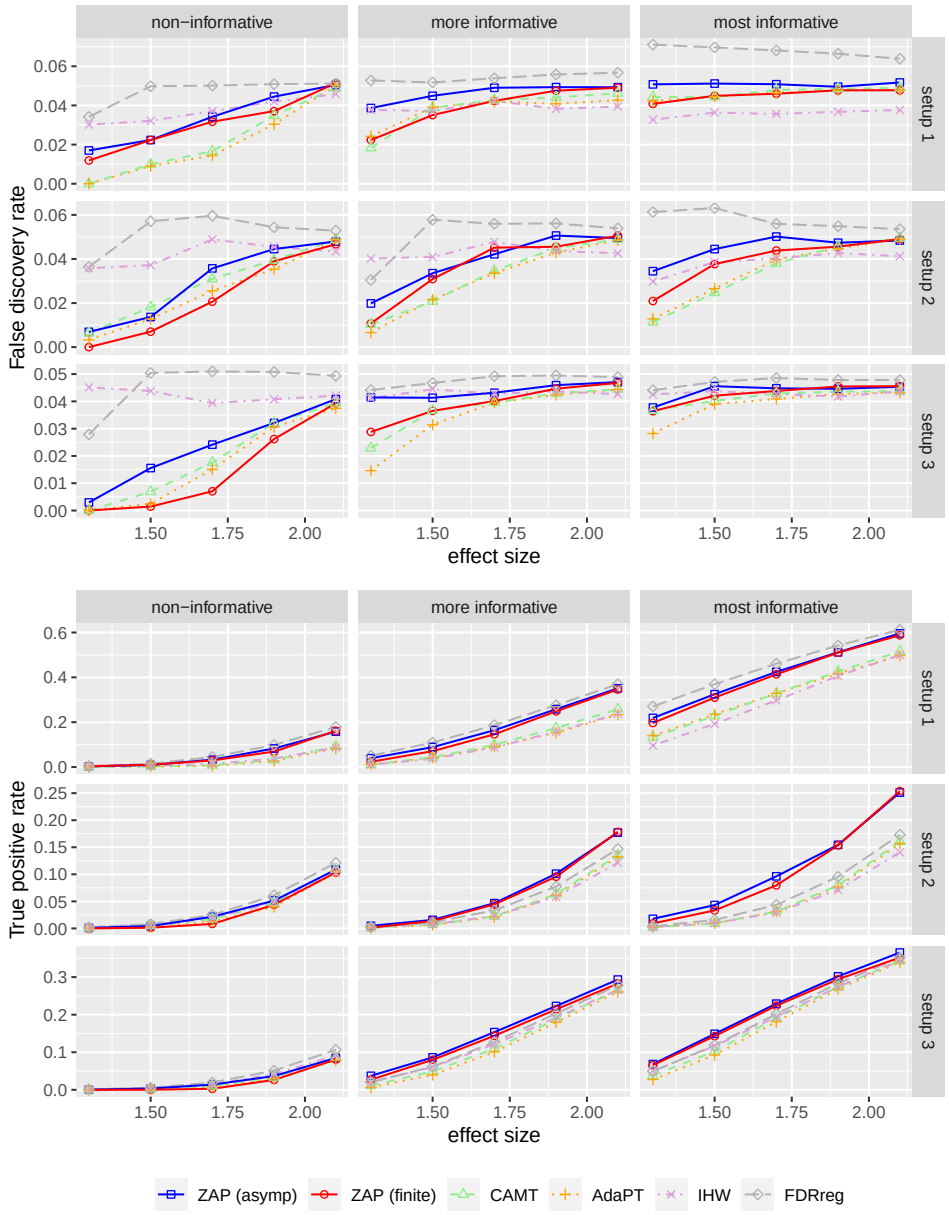

FIGURE 4.1. FDR and TPR performances of different methods under **Setup 1 - 3** . All methods are applied at a targeted FDR level of 0.05. The x-axes show the values of ε. `non-informative`, `more informative` and `most informative` correspond to different values of ζ from the smallest to the largest.

`airway`, `hippo`) are generated by RNA sequencing (RNA-Seq) experiments for detecting differential expressions in transcriptomes, where the primary statistic Z_i

Name	# tests	Brief description
`bottomly`	11484	DE in striatum for the two mouse strains C57BL/6J(B6) and DBA/2J(D2); bulk RNA-seq (Bottomly et al., 2011).
`airway`	20941	DE in human airway smooth muscle cell lines in response to dexamethasone; bulk RNA-seq (Himes et al., 2014).
`hippo`	15000	DE in mouse hippocampus in response to enzymatic dissociation in comparison to standard tissue homogenization; scRNA-seq (Harris et al., 2019).
`scott`	7004	Synchronous firing of pairs of neurons, based on neuron recordings in the primary visual cortex of an anesthetized monkey in response to visual stimuli (Scott et al., 2015).

TABLE 1. Description of four real datasets. "# tests" shows the number of tests for each dataset after any necessary data preprocessing. DE = Differential Expression.

measures the observed difference in the expression level of a gene under two experimental conditions. Meanwhile, an auxiliary covariate, the average normalized read count for each gene, is collected alongside the primary data. The datasets `bottomly` and `airway` have been analyzed by the works of Ignatiadis et al. (2016), Lei and Fithian (2018), Zhang and Chen (2020) with the methods IHW, AdaPT and CAMT respectively. The more recent data set `hippo` (Harris et al., 2019) is generated by the cutting-edge single-cell RNA (sc-RNA) sequencing technology to study differential expressions in mouse hippocampus. For all datasets above, we have adopted the standard data pre-processing step, which filters out genes with excessively low read counts across samples before further downstream analyses (Chen et al., 2016) such as model fitting and multiple testing. This is a common practice among bioinformaticians for a number of reasons; see Appendix G.3 for more discussion. The fourth dataset is based on the experiments in Smith and Kohn (2008) and Kelly et al. (2010), where each Z_i is a normalized test statistic that, for a given pair of neurons in the primary visual cortex, measures how synchronous their spike trains are, and Scott et al. (2015) has applied FDRreg to it for detecting neural interactions. Correspondingly, each such hypothesis has two covariates: the distance and the correlation of the "tuning curves" between the two activated neurons. We have named this dataset `scott` for short.

For the RNA-seq datasets, all the methods that accommodate multivariate covariates (CAMT, AdaPT, FDRreg and the two methods of ZAP) are applied with the log mean normalized read count expanded by a natural cubic spline basis with 4 interior knots, using the `ns` function in the `R` package `splines` (with its `df` argument set to 6), and IHW is applied with the original log mean normalized read count as it can only handle a univariate covariate. As there are two covariates for the neural dataset `scott`, IHW is not applied, and following what was done in Scott et al. (2015), the multivariate methods are applied with each of the two covariates expanded by a B-spline basis using the `bs` function in `R` with its argument `df` set to 3, which results in six expanded covariates in total.

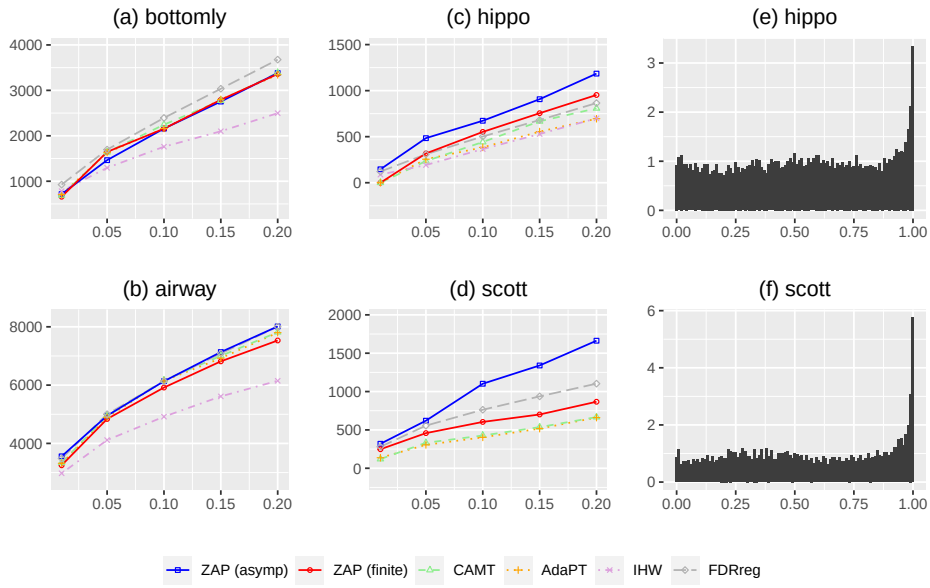

FIGURE 4.2. (**a**) – (**d**) plot the numbers of rejections for different methods across datasets, against targeted FDR level at $0.01, 0.05, 0.1, 0.15, 0.2$. (**e**) and (**f**) are respectively the histograms of the "u-values" for the `hippo` and `scott` datasets.

The number of rejections for the various methods are shown in Figure 4.2(a)-(d), and ZAP has attained top power performance. For the datasets `bottomly` and `airway`, we do not see substantial differences between the rejection numbers of ZAP and other methods such as CAMT, with the exception of FDRreg, which shows moderately more rejections than others for `bottomly`. However, as observed in our simulation studies, FDRreg can be invalid for FDR control and the power gain may be due to overflow in FDR. Simple histogram plots (Figure G.3 in Appendix G.3, for instance) show that the u-values are almost symmetrically distributed for these two datasets, which suggests that p-value and z-value based methods tend to have comparable power, unless reduction to p-values fails to capture the interactive information between the z-values and covariates. For the datasets `hippo` and `scott`, the histograms, which are shown in Figure 4.2(e) and (f), show that the u-value distribution is asymmetric. This explains why ZAP exhibits considerable power improvement over the p-value methods. Specifically, the patterns in Figure 4.2 (c) and (d) are in agreement with our intuition that ZAP is capable of exploiting the distributional asymmetry. Similar to what we observed in the simulation studies, the asymptotic ZAP tends to reject more hypotheses than the finite-sample ZAP, whose validity is based on fewer assumptions (Theorem 3.2). In summary, the real data analyses affirm that z-value based approaches to covariate-adaptive testing can better exploit the full data $\{Z_i, X_i\}_{i=1}^m$ to boost testing power.

5. Discussion

We have introduced ZAP, which is a z-value based covariate-adaptive testing framework that offers FDR controls under minimal assumptions. In particular, our method for finite-sample FDR control assumes no more than the knowledge of the z-value null distribution. The main thrust of our proposal is to avoid the common data reduction step of forming two-sided p-values used by most other covariate-adaptive methods in the recent literature, so as to preserve as much information as possible, based upon which more powerful procedures can be devised.

As presented, ZAP operates through a simple three-component beta mixture working model which is a careful extension of the two-component beta mixture for p-value based testing. While there is no "one-size-fits-all" solution to all FDR analysis problems (see the extensive simulation studies in the recent paper of Korthauer et al. (2019), for instance), we believe the current form of ZAP is widely applicable to many covariate-adaptive testing situations. Apparently, one can also extend our current approach by adopting other working models of choice while guarding against potential model misspecifications. For example, normal mixtures are another popular class of models used by researchers for FDR testing in different applications (McLachlan et al., 2006, Nguyen et al., 2018), and naturally, one can incorporate covariate information via a mixture of regressions (Leisch, 2008). Note that any such working model does not have to be defined on the u-value scale like our beta mixture, since it is just a means to arrive at a sensible assessor function via CLfdr consideration. In fact, one can even pursue machine-learning ideas to accommodate very flexible predictive functional forms for the regressions involved, such as the gradient boosted tree (Yurko et al., 2020). While the implementation of these potential extensions deserves much deeper investigation than intended for the present work, we shall briefly discuss the subtleties that may arise.

Conceptually, our asymptotic method can be easily extended, since as long as one has constructed the assessor function, presumably the $\mathrm{CLfdr}_x(\cdot)$ under an estimable working model of choice, rejection decisions can be based on computing the mirror statistics as in Algorithm 1. There are two caveats: One is that if the working model induces overly complex assessor functions, the determination of the mirror statistics can be time consuming, which counteracts the relative efficiency achieved by the current simplistic beta mixture. For example, if an assessor involves a kernel estimate which is typically a sum of m terms in the present context, the evaluation of each uniform realization in Algorithm 1 will become very expensive. Moreover, developing an asymptotic justification like Theorem 3.1 may be prohibitive, as nice properties of the working model may not be readily available to prove the requisite Glivenko-Cantelli results.

For finite-sample FDR control, we first note that at each step t in Algorithm 2, a "left" u-value U_i is masked depending on whether it is in the "left masking region"

$$\underbrace{(0, s_{l,t}(X_i)]}_{\text{"left" rejection region}} \bigcup \underbrace{[0.5 - s_{l,t}(X_i), 0.5]}_{\text{"left" acceptance region}},$$

and similarly for a "right" u-value. By requiring that $s_{l,t+1} \leq s_{l,t}$ and $s_{rt+1} \geq s_{rt}$, the masked u-values are gradually revealed. However, this is only one particular way of shrinking the rejection regions and their companying acceptance regions to reveal the u-values; so long as the rejection regions are shrunk based only on the partially masked information available at step t, the proof of Theorem 3.2 can be adapted to establish finite-sample FDR control. Hence, other working models, which may lead to different ways of shrinking the rejection regions based on their associated CLfdr calculations, can be deployed too. We leave these possibilities to future research that may be opportune in other instances.

References

Arias-Castro, E., Chen, S., et al. (2017). "Distribution-free multiple testing." *Electronic Journal of Statistics*, 11(1): 1983–2001.

Barber, R. F. and Candès, E. J. (2015). "Controlling the false discovery rate via knockoffs." *The Annals of Statistics*, 43(5): 2055–2085.

— (2019). "A knockoff filter for high-dimensional selective inference." *The Annals of Statistics*, 47(5): 2504–2537.

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

Benjamini, Y. 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 (Methodological)*, 57(1): 289–300.

— (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics." *Journal of educational and Behavioral Statistics*, 25(1): 60–83.

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

Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). "Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays." *PloS one*, 6(3): e17820.

Cai, T. T. and Sun, W. (2009). "Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks." *J. Amer. Statist. Assoc.*, 104: 1467–1481.

Cai, T. T., Sun, W., and Wang, W. (2019). "Covariate-assisted ranking and screening for large-scale two-sample inference." In *Royal Statistical Society*, volume 81.

Chen, X., Robinson, D. G., and Storey, J. D. (2017). "The functional false discovery rate with applications to genomics." *Biostatistics*.

Chen, Y., Lun, A. T., and Smyth, G. K. (2016). "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline." *F1000Research*, 5.

Deb, N., Saha, S., Guntuboyina, A., and Sen, B. (2021). "Two-component mixture model in the presence of covariates." *Journal of the American Statistical Association*, 1–35.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). "Bayesian density regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 163–183.

Efron, B. (2008). "Simultaneous inference: When should hypothesis testing problems be combined?" *Ann. Appl. Stat.*, **2**: 197–223.

Ferrari, S. and Cribari-Neto, F. (2004). "Beta regression for modelling rates and proportions." *Journal of applied statistics*, 31(7): 799–815.

Harris, R. M., Kao, H.-Y., Alarcon, J. M., Hofmann, H. A., and Fenton, A. A. (2019). "Hippocampal transcriptomic responses to enzyme-mediated cellular dissociation." *Hippocampus*, 29(9): 876–882.

Heller, R. and Rosset, S. (2021). "Optimal control of false discovery criteria in the two-group model." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1): 133–155.

Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderman, B., Whitaker, R. M., Duan, Q., Lasky-Su, J., Nikolos, C., et al. (2014). "RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells." *PloS one*, 9(6): e99625.

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(7): 577–580.

Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K. R. (2005). "Applications of beta-mixture models in bioinformatics." *Bioinformatics*, 21(9): 2118–2122.

Kelly, R. C., Smith, M. A., Kass, R. E., and Lee, T. S. (2010). "Local field potentials indicate network state and account for neuronal response variability." *Journal of computational neuroscience*, 29(3): 567–579.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). "A practical guide to methods controlling false discoveries in computational biology." *Genome biology*, 20(1): 1–21.

Lei, L. and Fithian, W. (2018). "AdaPT: an interactive procedure for multiple testing with side information." *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(4): 649–679.

Leisch, F. (2008). "FlexMix Version 2: Finite Mixtures with." *Journal of Statistical Software*, 28(4): 1–35.

Li, A. and Barber, R. F. (2019). "Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1): 45–74.

Love, M. I., Huber, W., and Anders, S. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome biology*, 15(12): 1–21.

Markitsis, A. and Lai, Y. (2010). "A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes." *Bioinformatics*, 26(5): 640–646.

McLachlan, G. J., Bean, R., and Jones, L. B.-T. (2006). "A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays." *Bioinformatics*, 22(13): 1608–1615.

Migliorati, S., Di Brisco, A. M., Ongaro, A., et al. (2018). "A new regression model for bounded responses." *Bayesian Analysis*, 13(3): 845–872.

Nguyen, H. D., Yee, Y., McLachlan, G. J., and Lerch, J. P. (2018). "False discovery rate control under reduced precision computation for analysis of neuroimaging data." *arXiv preprint arXiv:1805.04394*.

Parker, R. and Rothenberg, R. (1988). "Identifying important results from multiple statistical tests." *Statistics in medicine*, 7(10): 1031–1043.

Pounds, S. and Morris, S. W. (2003). "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values." *Bioinformatics*, 19(10): 1236–1242.

Resnick, S. (2019). *A probability path*. Springer.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic acids research*, 43(7): e47–e47.

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(510): 459–471.

Smith, M. A. and Kohn, A. (2008). "Spatial and temporal scales of neuronal correlation in primary visual cortex." *Journal of Neuroscience*, 28(48): 12591–12603.

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

Storey, J. D., Dai, J. Y., and Leek, J. T. (2007). "The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments." *Biostatistics*, 8(2): 414–432.

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 (Statistical Methodology)*, 66(1): 187–205.

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

Tansey, W., Koyejo, O., Poldrack, R. A., and Scott, J. G. (2018). "False discovery rate smoothing." *Journal of the American Statistical Association*, 113(523): 1156–1171.

White, H. (1981). "Consequences and detection of misspecified nonlinear regression models." *Journal of the American Statistical Association*, 76(374): 419–433.

— (1982). "Maximum likelihood estimation of misspecified models." *Econometrica: Journal of the Econometric Society*, 1–25.

Yurko, R., G?Sell, M., Roeder, K., and Devlin, B. (2020). "A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk." *Proceedings of the National Academy of Sciences*, 117(26): 15028–15035.

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*, (just-accepted): 1–31.

Appendix A. Oracle procedures

A.1. **Optimality of $\boldsymbol{\delta}^{\mathcal{P}}$ and $\boldsymbol{\delta}^{\mathcal{Z}}$.** In this section we briefly review the optimality properties of the oracle procedures $\boldsymbol{\delta}^{\mathcal{P}}$ and $\boldsymbol{\delta}^{\mathcal{Z}}$ in Section 2.2 and how their thresholds $t_{\mathcal{P}}$ and $t_{\mathcal{Z}}$ are determined. To streamline the discussion, we will let $\{M_i\}_{i=1}^m$ denote the set of *main statistics*, where it can either be that $M_i = P_i$ or $M_i = Z_i$ for all i, depending on whether p-value or z-value based methods are considered. Given the data $\{M_i, X_i\}_{i=1}^m$, it is well-known that *optimal* procedures, which aim to maximize true discoveries subject to false discovery constraints, should operate by rejecting i if its corresponding posterior probability $P(H_i = 0|M_i, X_i)$ falls below a data-dependent threshold $t_{\mathcal{M}}$. We will use $\boldsymbol{\delta}^{\mathcal{M}}$ (in a similar way as $\boldsymbol{\delta}^{\mathcal{P}}$ or $\boldsymbol{\delta}^{\mathcal{Z}}$) to denote the procedure that thresholds the quantities $\{P(H_i = 0|M_i, X_i)\}_{i=1}^m$ with $t_{\mathcal{M}}$.

There are subtly different ways to define "optimality", depending on the particular false discovery (e.g. FDR, mFDR, pFDR) and power (e.g. TPR, ETD, mFNR) measures used, which may lead to different ways of setting $t_{\mathcal{M}}$. A most recent result of Heller and Rosset (2021, Theorem 3.1) suggests that among all the testing procedures that are functions of $\{M_i, X_i\}_{i=1}^m$, the $t_{\mathcal{M}}$ that renders an ETD-maximing $\boldsymbol{\delta}^{\mathcal{M}}$ with $FDR \leqslant \alpha$ can be found by solving an integer optimization problem (Heller and Rosset, 2021, Theorem 3.1). For our purpose, by letting $L_{(1)} < \cdots < L_{(m)}$ be the order statistics of the posterior probabilities $\{P(H_i = 0|M_i, X_i)\}_{i=1}^m$, we have considered the computationally simpler optimal procedure first proposed in Sun and Cai (2007) which takes $t_{\mathcal{M}} = L_{(j)}$, where

$$j \equiv \max\left\{ i' \in \{1, \ldots, m\} : \frac{\sum_{i=1}^{i'} L_{(i)}}{i'} \leqslant \alpha \right\}. \tag{A.1}$$

This procedure has $FDR \leqslant \alpha$ because for any procedure that produces a rejection set $\mathcal{R}$ based on $\{M_i, X_i\}_{i=1}^m$, its FDR can be written as

$$FDR = \mathbb{E}\left[\underbrace{\mathbb{E}\left[\frac{\sum_{i\in\mathcal{R}}(1-H_i)}{|\mathcal{R}| \vee 1} \Big| \{M_i, X_i\}_{i=1}^m \right]}_{\mathbb{E}[FDP|\{M_i,X_i\}_{i=1}^m]} \right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{R}} P(H_i = 0|M_i, X_i)}{|\mathcal{R}| \vee 1} \right].$$

Conditional on any instance of the data $\{M_i, X_i\}_{i=1}^m$, $\boldsymbol{\delta}^{\mathcal{M}}$ prioritizes rejections of the hypotheses that are least likely to be true nulls, all the while controlling the conditional FDR $\mathbb{E}[FDP|\{M_i, X_i\}_{i=1}^m]$ below α by setting $t_{\mathcal{M}} = L_{(j)}$, as the ratio $\sum_{i=1}^{i'} L_{(i)}/i'$ in (A.1) is precisely the conditional FDR of rejecting the most promising i' hypotheses. As a result, its controls the FDR under α since the conditional version $\mathbb{E}[FDP|\{M_i, X_i\}_{i=1}^m]$ is always not larger than α.

We now give a more precise account of the optimal property of the prior procedure. Another popular measure of type 1 errors is the marginal FDR (mFDR), which for any rejection set $\mathcal{R}$ is the ratio

$$mFDR \equiv \frac{\mathbb{E}[V]}{\mathbb{E}[R]},$$

where V and R are defined as in the main text. For $\alpha \in (0,1)$, it is known that among all the procedures based on $\{M_i, X_i\}_{i=1}^m$ with $mFDR \leqslant \alpha$, the ETD-maximizing procedure is the one given by $\boldsymbol{\delta}^{\mathcal{M}}$ that sets $t_{\mathcal{M}} = \lambda^*_{\mathcal{M}}$, where

$$\lambda^*_{\mathcal{M}} \equiv \sup\left\{\lambda \in (0,1] : \frac{\sum_i \mathbb{E}[(1-H_i)I(P(H_i=0|M_i,X_i) \leqslant \lambda)]}{\sum_i \mathbb{E}[I(P(H_i=0|M_i,X_i) \leqslant \lambda)]} \leqslant \alpha\right\}; \tag{A.2}$$

see Sun and Cai (2007) and Heller and Rosset (2021). In practice, since $\lambda^*_{\mathcal{M}}$ could be tricky to obtain even with oracle knowledge, this optimal procedure for mFDR control is often approximated by our computationally handy version with $t_{\mathcal{M}} = L_{(j)}$ above when m is large, as is the case with most FDR analyses. Their asymptotic equivalence can be shown by standard arguments, such as those in Sun and Cai (2007, Section 4). The aforementioned references provide a more detailed exposition.

A.2. **Rejection regions for Examples 2.1-2.3.** Consider the p-value conditional mixture density

$$P_i|X_i = x \sim g_x(p) \equiv g(p|x) = (1-w_x)g_0(p) + w_x g_{1,x}(p), \tag{A.3}$$

induced by (2.2), where $g_0 \equiv 1$ is the uniform null density of P_i, and $g_{1,x}(p) \equiv g(p|H_i = 1, X_i = x)$ is the conditional alternative density of P_i. The rejection regions $\mathcal{S}^{\mathcal{P}}(x)$ and $\mathcal{S}^{\mathcal{Z}}(x)$ in Figure 2.1 are derived based on the threshold $\lambda^*_{\mathcal{M}}$ in (A.2), where

$$\mathcal{S}^{\mathcal{Z}}(x) \equiv \{z : P(H_i=0|Z_i=z, X_i=x) \leqslant \lambda^*_{\mathcal{Z}}\} = \left\{z : \frac{w_x f_{1,x}(z)}{(1-w_x)f_0(z)} \leqslant \frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}\right\}$$

and

$$\mathcal{S}^{\mathcal{P}}(x) \equiv \{z : P(H_i=0|P_i=p, X_i=x) \leqslant \lambda^*_{\mathcal{P}}\} = \left\{z : \frac{w_x g_{1,x}(2\Phi(-|z|))}{(1-w_x)g_0(2\Phi(-|z|))} \leqslant \frac{1-\lambda^*_{\mathcal{P}}}{\lambda^*_{\mathcal{P}}}\right\}.$$

$\lambda^*_{\mathcal{P}}$ and $\lambda^*_{\mathcal{Z}}$ are $\lambda^*_{\mathcal{M}}$ defined with $M_i = P_i$ and $M_i = Z_i$ for all i respectively. Since the examples are relatively simple, these regions can be found by numerical means, and we will derive $\mathcal{S}^{\mathcal{Z}}(x)$ in Example 2.2 as an illustration: It is the set

$$\left\{z : \frac{0.2\left[\frac{1-x}{2}\exp(-\frac{(z+\mu)^2}{2}) + \frac{1+x}{2}\exp(-\frac{(z-\mu)^2}{2})\right]}{0.8\exp(-\frac{z^2}{2})} \leqslant \frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}\right\},$$

where $\mu \equiv 1.5$. By setting $\frac{w_x f_{1,x}(z)}{(1-w_x)f_0(z)} = \frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}$, one arrives at the equation

$$\frac{1+x}{2}\exp(2\mu z) - 4\frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}\exp\left(\frac{\mu^2}{2}\right)\exp(\mu z) + \frac{1-x}{2} = 0.$$

in z. To solve for a solution z^*, we can apply the formula for the solutions of a quadratic equation to get

$$\exp(\mu z^*) = \frac{4\frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}\exp(\frac{\mu^2}{2}) \pm \sqrt{16\left(\frac{1-\lambda^*_{\mathcal{Z}}}{\lambda^*_{\mathcal{Z}}}\right)^2\exp(\mu^2) - (1-x)(1+x)}}{1+x},$$

which in turn implies the two boundary points

$$z^* = \frac{1}{\mu}\log\left\{\frac{4\frac{1-\lambda_Z^*}{\lambda_Z^*}\exp(\frac{\mu^2}{2}) \pm \sqrt{16\left(\frac{1-\lambda_Z^*}{\lambda_Z^*}\right)^2\exp(\mu^2) - (1-x)(1+x)}}{1+x}\right\}$$

for the red regions in the middle panel of Figure 2.1 as a function of x.

We now explain why $\mathcal{S}^{\mathcal{P}}(x)$ doesn't change with x in Examples 2.2 and 2.3. From the conditional p-value density (A.3), one can see that

$$P(H_i = 0|P_i, X_i) = \frac{(1-w_{X_i})g_0(P_i)}{g_{X_i}(P_i)} = \frac{(1-w_{X_i})g_0(P_i)}{(1-w_{X_i})g_0(P_i) + w_{X_i}g_{1,X_i}(P_i)}.$$

Hence, if w_x and $g_{1,x}$ do not depend on x, it is apparent that $\mathcal{S}_x^{\mathcal{P}}$ will not vary in x. Simple calculations can show that $w_x = 0.2$ and $w_x = 0.1$ for Examples 2.2 and 2.3 respectively, and

$$g_{1,x}(p) = \frac{\phi(\Phi^{-1}(p/2) - 1.5) + \phi(-\Phi^{-1}(p/2) - 1.5)}{2\phi(-\Phi^{-1}(p/2))} = \frac{\phi(-|z| - 1.5) + \phi(|z| - 1.5)}{2\phi(|z|)}$$

for both examples; all of these quantities do not depend on x.

Appendix B. Proof for the prototype method

We will prove the FDR validity of the prototype procedure in Section 3.1. The false discovery proportion of the prototype testing procedure, which thresholds the test statistics T_i's with the threshold $\hat{t}_\alpha \equiv \hat{t}(\alpha)$, can be written as

$$\begin{aligned} FDP &= \frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1 \vee \#\{T_i \leqslant \hat{t}_\alpha\}} \\ &= \frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1 + \#\{i \text{ null: } S_i \geqslant 1 - c_i(\hat{t}_\alpha)\}} \underbrace{\frac{1 + \#\{i \text{ null: } S_i \geqslant 1 - c_i(\hat{t}_\alpha)\}}{1 \vee \#\{T_i \leqslant \hat{t}_\alpha\}}}_{\leqslant \alpha \text{ by the defintion of our procedure}} \\ &\leqslant \alpha \frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1 + \#\{i \text{ null: } S_i \geqslant 1 - c_i(\hat{t}_\alpha)\}}. \end{aligned}$$

We only have to show that $\mathbb{E}\left[\frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1+\#\{i \text{ null: } S_i \geqslant 1-c_i(\hat{t}_\alpha)\}}\right]$ is bounded by 1 using the stopping time argument from (Barber and Candès, 2019).

Without loss of generality we will assume the true nulls are the first m_0 hypotheses. For each $i \in \{1, \ldots, m_0\}$, define

$$\breve{S}_i = \begin{cases} S_i & \text{when } S_i \leqslant 0.5 \\ 1 - S_i & \text{when } S_i > 0.5 \end{cases}$$

and $\breve{T}_i \equiv c_i^{-1}(\breve{S}_i)$. Using the order statistics $\breve{T}_{(1)} \leqslant \cdots \leqslant \breve{T}_{(m_0)}$ of $\breve{T}_1, \ldots, \breve{T}_{m_0}$, we moreover let $B_i \equiv I(S_{(i)} > 0.5)$ for $i = 1, \ldots, m_0$, where the order of $S_{(i)}$'s here is inherited from the order of the $\breve{T}_{(i)}$'s, rather than the magnitudes of the S_i's themselves. Let $1 \leqslant J \leqslant m_0$ be the index such that

$$\breve{T}_{(1)} \leqslant \cdots \leqslant \breve{T}_{(J)} \leqslant \hat{t}_\alpha < \breve{T}_{(J+1)} \leqslant \cdots \leqslant \breve{T}_{(m_0)},$$

we then have

$$\frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1+\#\{i \text{ null: } S_i \geqslant 1-c_i(\hat{t}_\alpha)\}} = \frac{\#\{i \text{ null: } T_i \leqslant \hat{t}_\alpha\}}{1+\#\{i \text{ null: } c_i^{-1}(1-S_i) \leqslant \hat{t}_\alpha\}}$$
$$= \frac{(1-B_1)+\cdots+(1-B_J)}{1+B_1+\cdots+B_J} = \frac{1+J}{1+B_1+\cdots+B_J} - 1,$$

considering that $\hat{t}_\alpha$ must be less than $t_{\max} := \max\{t : c_i(t) \leqslant 0.5 \text{ for all } i\}$. Hence it amounts to showing $E[\frac{1+J}{1+B_1+\cdots+B_J}] \leqslant 2$. This final step can be shown by applying Barber and Candès (2019, Lemma 1), since *conditional on* (i) $\breve{T}_{(1)} \ldots \breve{T}_{(m_0)}$ and (ii) $\{T_i : i \text{ is non null}\}$, $B_1, \ldots, B_{m_0}$ are independent Bernoulli(0.5) random variables, and J can be seen as a stopping time in reverse time with respect to the filtrations $\{\mathcal{F}_j\}_{j=1}^{m_0}$, where $\mathcal{F}_j \equiv \{B_1+\cdots+B_j, B_{j+1}, \ldots, B_{m_0}\}$.

Appendix C. Properties of the working model

In this section we will develop some properties of the beta-mixture model in Section 3.2 and the assessor functions it induces. To simplify notation, we will use $\pi_{li}, \pi_{ri}, k_{li}, k_{ri}, h_{li}, h_{ri}$ to respectively denote the quantities and functions π_{l,X_i}, π_{r,X_i}, k_{l,X_i}, k_{r,X_i}, h_{l,X_i}, h_{r,X_i} from Model (3.6) when the observed covariate X_i is used, where the underlying parameters $\{\theta_l, \theta_r, \beta_l, \beta_r\}$ are unspecified but common for all $i = 1, \ldots, m$. Likewise, we also use

$$a_i(u) \equiv a_{X_i}(u) = \frac{1-\pi_{li}-\pi_{ri}}{(1-\pi_{li}-\pi_{ri})+\pi_{li}h_{li}(u)+\pi_{ri}h_{ri}(u)} \tag{C.1}$$

to denote the assessor function constructed with them, and T_i, S_i and $c_i(\cdot)$ will denote the test statistics and null distribution function based on $a_i(\cdot)$ as in Section 3.1.

First, the following lemma states properties concerning the left and right alternative functions h_{li} and h_{ri}.

Lemma C.1 (Properties of the non-null component densities)**.** *For $\gamma_l > 2$, $h_{li}(\cdot)$ is a strictly convex and strictly decreasing function with the properties*

$$\lim_{u\to 0} h_{li}(u) = \infty \text{ and } \lim_{u\to 1} h_{li}(u) = 0.$$

Similarly, for $\gamma_r > 2$, $h_{r_i}(\cdot)$ is a strictly convex and strictly increasing function with the properties

$$\lim_{u\to 0} h_{ri}(u) = 0 \text{ and } \lim_{u\to 1} h_{ri}(u) = \infty.$$

Proof of Lemma C.1. It suffices to show the facts for h_{li} since those for h_{ri} can be proven exactly analogously. Recall that

$$h_{li}(u) := B(k_{li}, \gamma_l)^{-1} u^{k_{li}-1}(1-u)^{\gamma_l-1},$$

where for brevity we have suppressed the dependence on X_i in notations. Differentiating with respect to u we get

$$h'_{li}(u) = B(k_{li}, \gamma_l)^{-1}[(k_{li}-1)u^{k_{li}-2}(1-u)^{\gamma_l-1} - (\gamma_l-1)(1-u)^{\gamma_l-2}u^{k_{li}-1}], \tag{C.2}$$

which, given $\gamma_l > 2$ (actually $\gamma_l > 1$ is sufficient), can be seen to be always negative for any $u \in (0,1)$ and hence proves that h_{li} is strictly decreasing. For convexity, we differentiate one more time to get

$$\begin{aligned} h''_{li}(u) &= B(k_l,\gamma_l)(k_{li}-1)\left[(k_{li}-2)u^{k_{li}-3}(1-u)^{\gamma_l-1} - u^{k_{li}-2}(\gamma_l-1)(1-u)^{\gamma_l-2}\right] \\ &\quad - B(k_l,\gamma_l)(\gamma_l-1)\left[(2-\gamma_l)(1-u)^{\gamma_l-3}u^{k_{li}-1} + (k_{li}-1)u^{k_{li}-2}(1-u)^{\gamma_l-2}\right] \\ &= B(k_l,\gamma_l)\underbrace{u^{k_{li}-3}(1-u)^{\gamma_l-3}}_{>0} \times \\ &\qquad \left[\underbrace{(k_{li}-1)(k_{li}-2)(1-u)^2}_{>0} - \underbrace{2(k_{li}-1)(\gamma_l-1)u(1-u)}_{>0} + (\gamma_l-1)(\gamma_l-2)u^2\right], \end{aligned}$$

where in the last equality, the positive terms are positive since $0 < k_{li} < 1$. As such, $h''_{li}(u)$ is strictly positive for all $u \in (0,1)$ as long as $\gamma_l > 2$, which proves the strict convexity of h_{li}. □

A closer inspection of the proof above will reveal that for $\gamma_l \in (1,2)$, h_{li} may not even be convex, and the same is true for h_{ri}. Hence we have required that $\gamma_l, \gamma_r > 2$ in our model. To facilitate the proof in later sections we will also define the *reciprocal* assessor function

$$b_i(u) \equiv 1/a_i(u). \tag{C.3}$$

By the properties of h_{li} and h_{ri} in Lemma C.1, one can readily conclude the following lemma, which will help us develop some useful facts later:

Lemma C.2 (Properties of the reciprocal assessor)**.** *The reciprocal assessor function defined in* (C.3) *(for $\gamma_l, \gamma_r > 2$) is strictly convex and smooth, with the property that*

$$\lim_{u\to 0} b_i(u) = \lim_{u\to 1} b_i(u) = \infty. \tag{C.4}$$

Hence, there exists a unique minimal $\underline{u}_i$ such that

$$b_i(\underline{u}_i) < b_i(u) \text{ for all } u \in (0,1).$$

$T_i \equiv a_i(U_i)$ as a random variable has the range $(0, a_i(\underline{u}_i)]$ in light of Lemma C.2. By construction, $a_i(\cdot)$'s level sets can only be of Lebesgue measure 0, so $c_i(\cdot)$ is continuous under the uniform null distribution of U_i. With the strict convexity of $b_i(\cdot)$, one can also conclude that $c_i(\cdot)$, its null distribution function, is invertible (or equivalently, strictly increasing), since no interval in the range $(0, a_i(\underline{u}_i)]$ will have zero measure under the law of $a_i(U_i)$ induced by the uniform null distribution of U_i by the intermediate value theorem. The smooth "bowl" shape of $b_i(\cdot)$ also implies that, for any $t \in (0, a_i(\underline{u}_i)]$, the event $\{T_i > t\}$ is equivalent to U_i taking values in a certain sub-interval of $(0,1)$. One can define two smooth functions to describe this fact:

Definition C.1 (Expression for the event $\{T_i > t\}$)**.** For each i, $\omega_{iL} : (0, a_i(\underline{u}_i)] \to (0, \underline{u}_i]$ and $\omega_{iR} : (0, a_i(\underline{u}_i)] \to [\underline{u}_i, 1)$ are respectively two smooth functions such

that for any $t \in (0, a_i(\underline{u}_i)]$,

$$\{T_i > t\} = \{\omega_{iL}(t) < U_i < \omega_{iR}(t)\} = \{S_i > c_i(t)\},$$

with $\omega_{iL}(\cdot)$ and $\omega_{iR}(\cdot)$ being strictly increasing and strictly decreasing, respectively. Since S_i and U_i are uniformly distributed when $H_i = 0$,

$$\omega_{iR}(t) - \omega_{iL}(t) = 1 - c_i(t).$$

Moreover, $b_i(\omega_{iL}(t)) = b_i(\omega_{iR}(t)) = 1/t$.

Of course the variable $S_i = c_i(T_i)$ has the range $(0, 1]$, and we can define functions to describe events of the form $\{S_i > s\}$ similar to Definition C.1:

Definition C.2 (Expression for the event $\{S_i > s\}$)**.** For each i, $\psi_{iL} : (0, 1] \to (0, \underline{u}_i]$ and $\psi_{iR} : (0, 1] \to [\underline{u}_i, 1)$ are respectively two smooth functions such that for any $s \in (0, 1]$,

$$\{S_i > s\} = \{\psi_{iL}(s) < U_i < \psi_{iR}(s)\} = \{T_i > c_i^{-1}(s)\},$$

with $\psi_{iL}(\cdot)$ and $\psi_{iR}(\cdot)$ being strictly increasing and strictly decreasing, respectively. Since S_i and U_i are uniformly distributed when $H_i = 0$,

$$\psi_{iR}(s) - \psi_{iL}(s) = 1 - s. \tag{C.5}$$

Moreover, $b_i(\psi_{iL}(s)) = b_i(\psi_{iR}(s)) = 1/c_i^{-1}(s)$.

Appendix D. Proof for the asymptotic method

Before proving Theorem 3.1, we remark that the theorem is established by assuming that the mirror statistic $\hat{T}_i^{\mathfrak{m}}$ is the *exact* reflection of $\hat{T}_i$ under the null distribution $\hat{c}_i$. In practice, $\hat{T}_i^{\mathfrak{m}}$ can be determined up to arbitrary precision in Algorithm 1 as long as the number of uniform realizations N is set to be very large, as recommended in Section 3.5.

We will make heavy use of the notation and results in Appendix C. We will also use $a_i^*(\cdot)$, $b_i^*(\cdot)$, $c_i^*(\cdot)$, T_i^* and S_i^* to denote the respective functions and statistics when $\{\theta, \beta\}$ is taken to be the pair $\{\theta^*, \beta^*\}$ to construct $a_i(\cdot)$, $b_i(\cdot)$, $c_i(\cdot)$, T_i and S_i. Similarly, the quantities and functions appearing in Lemma C.2 and Definitions C.1 and C.2 all have their "star" versions: $\underline{u}_i^*$, $\omega_{iL}^*(\cdot)$, $\omega_{iR}^*(\cdot)$, $\psi_{iL}^*(\cdot)$ and $\psi_{iR}^*(\cdot)$. Generally speaking, $C, c > 0$ will denote unspecified universal constants required for the asymptotic arguments in this section.

D.1. Additional assumptions for Theorem 3.1.

Assumption 2 (Regularity conditions)**.**

(i) $\max_i \|X_i\|_\infty \leqslant C$ *almost surely for some universal constant* $C > 0$*, where* $\|\cdot\|_\infty$ *indicates the sup norm.*

(ii) Let $[\delta_1, \delta_2]$ *be any fixed compact interval in* $(0, 1)$*. For each* i*, let* V_{1i} *and* V_{2i} *be two measurable subsets in* (δ_1, δ_2)*. Then for large enough* m*,*

$$\left| \frac{1}{m} \sum_{i=1}^{m} P(U_i \in V_{1i} | X_i) - P(U_i \in V_{2i} | X_i) \right| \leqslant C(\delta_1, \delta_2) \max_{1 \leqslant i \leqslant m} \lambda\left(V_{i1} \Delta V_{i2}\right),$$

where $\lambda(\cdot)$ is the Lebesgue measure, $C(\delta_1, \delta_2)$ is a constant that may depends on δ_1 and δ_2, and $\mathcal{S}_1 \Delta \mathcal{S}_2$ is their symmetric difference for any two sets $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}$.

(iii) $\mathbb{E}[\log(U_i)]$ *and* $\mathbb{E}[\log(1 - U_i)]$ *are finite.*

Assumption 3 (Strong laws of large numbers)**.** *Let $\pi_0 \equiv P(H_i = 0) > 0$. For any $t \in (0, 1]$, it holds that*

$$\text{(D.1)} \quad m^{-1} \sum_{i=1}^{m} I(T_i^* \leqslant t) \longrightarrow G(t),$$

$$m^{-1} \sum_{i=1}^{m} (1 - H_i) I(T_i^* \leqslant t) \longrightarrow \pi_0 G_0(t) \text{ and}$$

$$m^{-1} \sum_{i=1}^{m} I(S_i^* \geqslant 1 - c_i^*(t)) \longrightarrow \bar{G}_0(t)$$

almost surely, where $G(t)$, $G_0(t)$ and $\bar{G}_0(t)$ are positive continuous functions in t. Moreover, $\bar{G}_0(t_0)/G(t_0) < \alpha$ for some $t_0 > 0$, and the limiting threshold

$$t_\alpha^\infty \equiv \sup \left\{ t \in (0, 1] : \frac{\bar{G}_0(t)}{G(t)} \leqslant \alpha \right\}$$

is such that $\max_{i \in \mathbb{N}} c_i^*(t_\alpha^\infty) < \bar{s}$ for $\bar{s} < 1$, or equivalently, $t_\alpha^\infty < \min_i {c_i^*}^{-1}(\bar{s})$. Note that the strong laws above, as well as the marginal probability π_0, are with respect to the joint law of $\{H_i, Z_i, X_i\}$.

Assumption 2 regulates the tail behaviors of the random variables X_i and U_i; in particular, (ii) implies that conditional on X_i, the density of U_i can be unbounded at the two tails, which is natural for multiple testing as it provides room for non-null tail behaviors. Assumption 3 states properties of the strong law limits involved. For technical reasons, that $\sup_{i \in \mathbb{N}} c_i^*(t_\alpha^\infty)$ is bounded away from 1 ensures that our result won't rely on the strong law limits for very large values of t, which is hardly restrictive in practice: any sensible multiple testing procedure should only consider rejecting S_i^*, which is uniformly distributed under the null, if it is much less than the typically small target FDR level α. Similar assumptions have also appeared in the works of Storey et al. (2004), Zhang and Chen (2020).

We remark that our current assumptions for Theorem 3.1 are no stronger than those in Zhang and Chen (2020) in any essential way, and can conceivably be further relaxed; for example, if Assumption 2(i) is phrased as a probabilistic bound, one can still likely establish a version of Theorem 3.1 which says that the FDR is less than α with probability approaching 1. Moreover, we have assume, as stated in Section 2.1, that $\{H_i, Z_i, X_i\}$ are independent across i, which can be further relaxed to a generic weak dependence condition under which the strong laws in Assumption 3 hold. In fact it is possible to prove an FDR bound in terms of the conditional expectation $\mathbb{E}[\cdot | H_1, \ldots, H_m]$, treating the hypotheses as fixed. These embellishments have not been pursued here for a more streamlined presentation.

D.2. **Technical lemmas.** Under Assumption 1 and Assumption 2(i), all $k_{ri} = k_r(X_i)$, $k_{li} = k_l(X_i)$, $\pi_{ri} = \pi_r(X_i)$, $\pi_{li} = \pi_l(X_i)$ are bounded away from one and zero, i.e. $k_{ri}, k_{li} \in [\underline{k}, \bar{k}]$, $\pi_{ri}, \pi_{li} \in [\underline{\pi}, \bar{\pi}]$ for some compact intervals $[\underline{k}, \bar{k}], [\underline{\pi}, \bar{\pi}] \subset (0,1)$, by the compactness of $\Theta \times \mathbf{B}$. Moreover, by continuity of the beta functions we can define

$$B_{\max} := \max_{k \in [\underline{k}, \bar{k}]} (B(k, \gamma_l) \vee (B(\gamma_r, k)) \text{ and } B_{\min} := \max_{k \in [\underline{k}, \bar{k}]} (B(k, \gamma_l) \vee (B(\gamma_r, k)),$$

which are both positive numbers.

The following "uniformity" properties will be heavily relied on later:

Lemma D.1 (Uniformity properties)**.** *Under Assumption 1 and Assumption 2(i), the following are true for any $(\theta, \beta) \in \Theta \times \boldsymbol{B}$:*

(i) For any $t_0 > 0$, there exists a $u_0 = u_0(t_0) > 0$ not depending on (θ, β) such that for all $t \geqslant t_0$, $(\omega_{iL}(t), \omega_{iR}(t)) \subset [u_0, 1 - u_0]$ for all i.

(ii) For any $s_0 > 0$, there exists a $u_0 = u_0(s_0) > 0$ not depending on (θ, β) such that for all $s \geqslant s_0$, $(\psi_{iL}(s), \psi_{iR}(s)) \subset [u_0, 1 - u_0]$ for all i.

(iii) There exists a small positive constant $u_0 \in (0, 0.5)$ not depending on (θ, β) such that,

$$\underline{u}_i \in [u_0, 1 - u_0]$$

for all i, where $\underline{u}_i$ is as in Lemma C.2.

(iv)

$$\lim_{\epsilon \to 0} \max_{i \in \mathbb{N}, \theta \in \Theta, \beta \in \boldsymbol{B}} \lambda\Big({b_i}^{-1}\big([b_i(\underline{u}_i), b_i(\underline{u}_i) + \epsilon)\big)\Big) = 0,$$

where λ is the Lebesgue measure, $\underline{u}_i$ is as in Lemma C.2 and

$${b_i}^{-1}(\mathcal{T}) \equiv \{u : b_i(u) \in \mathcal{T}\}.$$

for any interval $\mathcal{T}$ in $\mathbb{R}$.

Proof. (i): Note that

$$b_i(u) \geqslant \max\left(\frac{\underline{\pi} u^{\bar{k}-1}(1-u)^{\gamma_l - 1}}{(1 - 2\underline{\pi}) B_{\max}}, \frac{\underline{\pi} u^{\gamma_r - 1}(1-u)^{\bar{k}-1}}{(1 - 2\underline{\pi}) B_{\max}} \right) \text{ for all } i, \tag{D.2}$$

which implies $\lim_{u \to 0} \min_i b_i(u) = \lim_{u \to 1} \min_i b_i(u) = \infty$, since the right hand side of (D.2) tends to ∞ as u tends to 0 or 1. Hence one must be able to find a small enough $u_0 > 0$ such that $\min(b_i(u_0), b_i(1 - u_0)) > 1/t_0$ for all i, which implies that $(\omega_{iL}(t), \omega_{iR}(t)) \subset [u_0, 1 - u_0]$ for all i by Definition C.1. This proves (i).

(ii): Suppose towards a contradiction, such a u_0 doesn't exist. Without loss of generality, we assume there is a subsequence $\{i_1, i_2, \dots\}$ such that $\lim_{j \to \infty} \psi_{i_j L}(s_0) = 0$. As such, $\lim_{j \to \infty} \psi_{i_j R}(s_0) = 1 - s_0$ by the property stated in Definition C.2, which implies

$$\limsup_j b_{i_j}(\psi_{i_j R}(s_0)) \leqslant 1 + \frac{\bar{\pi}(1-s_0)^{\underline{k}-1}(s_0)^{\gamma_l - 1}}{(1 - 2\bar{\pi}) B_{\min}} + \frac{\bar{\pi}(1-s_0)^{\gamma_r - 1}(s_0)^{\underline{k}-1}}{(1 - 2\bar{\pi}) B_{\min}}. \tag{D.3}$$

On the other hand,

$$\lim_j b^*_{i_j}(\psi^*_{i_j L}(s_0)) = \infty \tag{D.4}$$

in consideration of (D.2) and $\lim_{j\to\infty}\psi^*_{i_jL}(s_0) = 0$. (D.3) and (D.4) together reach a contradiction since it must be that $\lim_j b^*_{i_j}(\psi^*_{i_jL}(s_0)) = \lim_j b^*_{i_j}(\psi^*_{i_jR}(s_0))$ as $b^*_{i_j}(\psi^*_{i_jL}(s_0)) = b^*_{i_j}(\psi^*_{i_jR}(s_0))$by Definition C.2.

(iii): By the fact that that $b'_i(b_i(\underline{u}_i)) = 0$ for all i, it suffices to show that

$$\lim_{u\to 0}\max_{i,\theta,\beta} b'_i(u) = -\infty \text{ and } \lim_{u\to 1}\min_{i,\theta,\beta} b'_i(u) = \infty. \tag{D.5}$$

Note that

$$b'_i(u) = \frac{\pi_{li}}{1-\pi_{ri}-\pi_{li}}h'_{li}(u) + \frac{\pi_{ri}}{1-\pi_{ri}-\pi_{li}}h'_{ri}(u),$$

where h'_{li} has the form

$$h'_{li}(u) = B(k_{li},\gamma_l)^{-1}[(k_{li}-1)u^{k_{li}-2}(1-u)^{\gamma_l-1} - (\gamma_l-1)(1-u)^{\gamma_l-2}u^{k_{li}-1}]$$

as shown in the proof of Lemma C.1. Define, for $u\in(0,1)$, the functions

$$\bar{h}'_l(u) = B^{-1}_{\max}[(\bar{k}-1)u^{\bar{k}-2}(1-u)^{\gamma_l-1} - (\gamma_l-1)(1-u)^{\gamma_l-2}u^{\bar{k}-1}],$$
$$\underline{h}'_l(u) = B^{-1}_{\min}[(\underline{k}-1)u^{\underline{k}-2}(1-u)^{\gamma_l-1} - (\gamma_l-1)(1-u)^{\gamma_l-2}u^{\underline{k}-1}]$$

so that $\underline{h}'_l(u) \leqslant h'_{li}(u) \leqslant \bar{h}'_l(u) < 0$ for all i. Note that

$$\lim_{u\to 0}\bar{h}'_l(u) = -\infty \text{ and } \lim_{u\to 1}\underline{h}'_l(u) = 0.$$

One can similarly define functions $\bar{h}'_r$ and $\underline{h}'_r$ on $(0,1)$ such that $0 < \underline{h}'_r(u) \leqslant h'_{ri}(u) \leqslant \bar{h}'_r(u)$ for all i and

$$\lim_{u\to 0}\bar{h}'_r(u) = 0 \text{ and } \lim_{u\to 1}\underline{h}'_r(u) = \infty.$$

The fact that $b'_i(u) \leqslant \frac{\bar{\pi}}{1-2\bar{\pi}}\bar{h}'_l(u) + \frac{\bar{\pi}}{1-2\bar{\pi}}\bar{h}'_r(u)$, together with two of the limit results above, has shown the first limit in (D.5). Similarly, that $b'_i(u) \geqslant \frac{\underline{\pi}}{1-2\underline{\pi}}\underline{h}'_l(u) + \frac{\underline{\pi}}{1-2\underline{\pi}}\underline{h}'_r(u)$, together with the other two limit results above, has shown the second limit in (D.5).

(iv): By Lemma D.1(iii), pick $u_0\in(0,0.5)$ such that $\underline{u}_i \in [u_0, 1-u_0]$ for all i. Note that

$$b''_i(u) = \frac{\pi_{li}}{1-\pi_{li}-\pi_{ri}}h''_{li}(u) + \frac{\pi_{ri}}{1-\pi_{li}-\pi_{ri}}h''_{ri}(u),$$

where the dependence on X_i has been suppressed in notations for brevity. By Lemma C.1 , b''_i is always positive, hence there exists a universal positive number $c>0$ such that

$$b''_i(u) > c$$

for all i and all $u\in[u_0/2, 1-u_0/2]$, considering that Θ, B and $[u_0/2, 1-u_0/2]$ are all compact. Now for each i consider the quadratic function

$$f_i(u) = \frac{c(u-\underline{u}_i)^2}{2} + b_i(\underline{u}_i)$$

defined on $[u_0/2, 1-u_0/2]$. Then on the interval $[u_0/2, 1-u_0/2]$, $b_i \geqslant f_i$ since $g_i = b_i - f_i$ is strictly convex with $g'_i(\underline{u}_i) = g_i(\underline{u}_i) = 0$. Then

$$\lambda\left(b_i^{-1}([m_i, m_i+\epsilon))\right) \leqslant \lambda\left(f_i^{-1}([m_i, m_i+\epsilon))\right) = \sqrt{\frac{8\epsilon}{c}}$$

where the right hand side obviously converges to zero as $\epsilon \to 0$.

□

We will now state two crucial "event inclusion" lemmas that involve the most delicate proofs in this paper, and may be skipped at first reading. To state them, we conveniently define the long vectors $\Pi = \Pi(\theta) = (\pi_{li}, \pi_{ri})_{i=1}^m$ and $K = K(\beta) = (k_{li}, k_{ri})_{i=1}^m$ with $2m$ components. Note that they implicitly depend on the unspecified parameters $\{\theta, \beta\}$. As such, we can also define $\Pi^* = \Pi(\theta^*)$ and $K^* = K(\beta^*)$ to be the versions evaluated at θ^* and β^*.

Lemma D.2 (First event inclusion lemma)**.** *Suppose Assumptions 1 and 2(i) are true and let $(\theta, \beta) \in \Theta \times B$. For given $\underline{t} > 0$ and $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$,*

$$\{S_i^* \leqslant c_i^*(t) - \epsilon\} \subset \{S_i \leqslant c_i(t)\} \subset \{S_i^* \leqslant c_i^*(t) + \epsilon\} \text{ for all } i \text{ and all } t > \underline{t}.$$

Proof of Lemma D.2. By Definition C.1, we will show, equivalently, that there exists $\delta > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$,

(D.6)
$$\{U_i \in \underbrace{(\omega_{iL}^*(t_{i,+\epsilon}), \omega_{iR}^*(t_{i,+\epsilon}))}_{\text{length}=1-c_i^*(t)-\epsilon}\} \subset \{U_i \in \underbrace{(\omega_{iL}(t), \omega_{iR}(t))}_{\text{length}=1-c_i(t)}\} \subset \{U_i \in \underbrace{(\omega_{iL}^*(t_{i,-\epsilon}), \omega_{iR}^*(t_{i,-\epsilon}))}_{\text{length}=1-c_i^*(t)+\epsilon}\},$$

where we define $t_{i,-\epsilon} := c_i^{*-1}(c_i^*(t) - \epsilon)$ and $t_{i,+\epsilon} := c_i^{*-1}(c_i^*(t) + \epsilon)$ that are respectively less and greater than t. In particular, we will first focus on showing the second inclusion in in (D.6) , which amounts to showing

$$b_i(\omega_{iL}^*(t_{i,-\epsilon})) \wedge b_i(\omega_{iR}^*(t_{i,-\epsilon})) \geqslant 1/t \tag{D.7}$$

whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$, in light of the fact that

$$b_i^*(\omega_{iL}^*(t_{i,-\epsilon})) = b_i^*(\omega_{iR}^*(t_{i,-\epsilon})) = 1/t_{i,-\epsilon}, \tag{D.8}$$

by the definition of $\omega_{iL}^*(\cdot)$ and $\omega_{iR}^*(\cdot)$ in Definition C.1 and properties of b_i from Lemma C.2.

Let $u_0 > 0$ be a small positive number such that $(\omega_{iL}^*(\underline{t}), \omega_{iR}^*(\underline{t})) \subset [u_0, 1 - u_0]$ by Lemma D.1(i), and consider the even larger compact interval $[u_0/2, 1 - u_0/2]$. Consider each $b_i(u) = b(u; \pi_i, k_i)$ as a function in (u, π_i, k_i), and let

$$\nabla_{\pi,k} b_i(u) \equiv \left(\frac{\partial}{\partial \pi_i} b(u; \pi_i, k_i)^T, \frac{\partial}{\partial k_i} b(u; \pi_i, k_i)^T \right)^T$$

be the gradient of b_i with respect to (π_i, k_i) evaluated at u. Using the compactness of $[u_0/2, 1 - u_0/2] \times \Theta \times \mathcal{B}$ and Assumption 2(i) again, one can find a universal constant $C(u_0) > 0$ such that the gradient bounds

$$\|\nabla_{\pi,k} b_i(u)\|_1 < C(u_0) \text{ for all } i, \text{ all } u \in \left[\frac{u_0}{2}, 1 - \frac{u_0}{2}\right], \text{ all } (\theta, \beta) \in \Theta \times \boldsymbol{B}. \tag{D.9}$$

On the other hand, without loss of generality, we will let

$$\epsilon < u_0/2 \tag{D.10}$$

and, with Lemma D.1(iv), take $\tilde{\epsilon} > 0$ be a small enough constant such that

$$\lambda(b_i^{*-1}([b_i^*(\underline{u}_i^*), b_i^*(\underline{u}_i^*) + \tilde{\epsilon})) < \epsilon \text{ for all } i. \tag{D.11}$$

By the mean-value theorem and the gradient bound (D.9), one can then find $\delta > 0$ such that when $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$

$$|b_i(u) - b_i^*(u)| < \tilde{\epsilon} \text{ for all } u \in \left[\frac{u_0}{2}, 1 - \frac{u_0}{2}\right]. \tag{D.12}$$

By the construction of $\tilde{\epsilon}$ in (D.11) and convexity properties from Lemma C.2, one must have for all i

$$b_i^*(\omega_{iL}^*(t_{i,-\epsilon})) - b_i^*(\omega_{iL}^*(t)) = b_i^*(\omega_{iR}^*(t_{i,-\epsilon})) - b_i^*(\omega_{iR}^*(t)) \geqslant \tilde{\epsilon},$$

which implies

$$1/t_{i,-\epsilon} \geqslant \tilde{\epsilon} + 1/t. \tag{D.13}$$

by the last property in Definition C.1. Since $\{S_i^* > c_i^*(t) - \epsilon\} \supset \{S_i^* > c_i^*(t)\}$, from the property (C.5) in Definition C.2 both

$$\omega_{iL}^*(t_{i,-\epsilon}) \in (\omega_{iL}^*(t) - \epsilon, \omega_{iL}^*(t)) \text{ and } \omega_{iR}^*(t_{i,-\epsilon}) \in (\omega_{iR}^*(t), \omega_{iR}^*(t) + \epsilon)$$

are true, which implies

$$[\omega_{iL}^*(t_{i,-\epsilon}), \omega_{iR}^*(t_{i,-\epsilon})] \subset [u_0/2, 1 - u_0/2],$$

considering (D.10) and $\omega_{iL}^*(t), \omega_{iR}^*(t) \in [u_0, 1 - u_0]$ (as $t \geqslant \underline{t}$). Therefore by (D.12), we must have

$$b_i(\omega_{iL}^*(t_{i,-\epsilon})) \wedge b_i(\omega_{iR}^*(t_{i,-\epsilon})) \geqslant 1/t_{i,-\epsilon} - \tilde{\epsilon} \tag{D.14}$$

given (D.8). Combining (D.13) and (D.14) gives (D.7).

The proof for the first inclusion in (D.6) follows an analogous argument but is with less resistance, since $[\omega_{iL}^*(t_{i,+\epsilon}), \omega_{iR}^*(t_{i,+\epsilon})] \subset [u_0, 1 - u_0]$ for all i. We leave it to the reader.

□

Lemma D.3 (Second event inclusion lemma)**.** *Suppose Assumptions 1 and 2(i) are true and let $(\theta, \beta) \in \Theta \times B$. For any fixed $\bar{t} < \min_i c_i^{*-1}(\bar{s})$ with $\bar{s} < 1$ and any $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$,*

$$\{S_i^* > 1 - c_i^*(t) - \epsilon\} \supset \{S_i > 1 - c_i(t)\} \supset \{S_i^* > 1 - c_i^*(t) + \epsilon\},$$

for all i and all $t \leqslant \bar{t}$.

Proof of Lemma D.3. Note that from (C.5) in Definition C.2 and Lemma D.2 we can conclude there exists a $\delta_1 > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta_1$,

$$1 - c_i^*(t) - \epsilon \leqslant 1 - c_i(t) \leqslant 1 - c_i^*(t) + \epsilon. \tag{D.15}$$

Based on (D.15), it suffices to show that there exists a $\delta_2 > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta_2$,

$$\{S_i > 1 - c_i^*(t) - \epsilon\} \subset \{S_i^* > 1 - c_i^*(t) - 2\epsilon\}, \tag{D.16}$$

and

$$\{S_i > 1 - c_i^*(t) + \epsilon\} \supset \{S_i^* > 1 - c_i^*(t) + 2\epsilon\} \tag{D.17}$$

which conclude the lemma by taking $\delta = \delta_1 \vee \delta_2$ and replacing ϵ with $\epsilon/2$. In fact, since $t \leqslant \bar{t}$ and $c_i^*(\bar{t})$ are bounded away from 1, we will show the more general statement: For a given $\underline{s} > 0$, there exists $\delta = \delta(\underline{s}) > 0$ such that whenever $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$,

$$\{S_i > s\} \subset \{S_i^* > s - \epsilon\} \tag{D.18}$$

and

$$\{S_i > s\} \supset \{S_i^* > s + \epsilon\} \tag{D.19}$$

for all $s \geqslant \underline{s}$ and all i. This will necessitate (D.16) and (D.17) for $t \leqslant \bar{t}$.

We will first show (D.18) which amounts to

$$(\psi_{iL}(s), \psi_{iR}(s)) \subset (\psi_{iL}^*(s - \epsilon), \psi_{iR}^*(s - \epsilon)) \tag{D.20}$$

in light of Definition C.2. In particular, it suffices to only consider the case where $\epsilon < \underline{s}$, since if $s - \epsilon \leqslant 0$, $\{S_i^* > s - \epsilon\} = \{S_i^* \geqslant 0\}$ becomes the whole underlying probability space which makes (D.20) trivially true. Now for each i, let

$$s(i) := \sup\{s' \in (0,1) : (\psi_{iL}(s), \psi_{iR}(s)) \subset (\psi_{iL}^*(s'), \psi_{iR}^*(s'))\}.$$

By Definition C.2 it must be the case that

$$s(i) = 1 - \psi_{iR}^*(s(i)) + \psi_{iL}^*(s(i)) \leqslant 1 - \psi_{iR}(s) + \psi_{iL}(s) = s, \tag{D.21}$$

and only one of the following possibilities can be true:

(i) $\psi_{iL}^*(s(i)) = \psi_{iL}(s)$ and $\psi_{iR}(s) < \psi_{iR}^*(s(i))$,
(ii) $\psi_{iL}^*(s(i)) < \psi_{iL}(s)$ and $\psi_{iR}(s) = \psi_{iR}^*(s(i))$,
(iii) $\psi_{iL}^*(s(i)) = \psi_{iL}(s)$ and $\psi_{iR}(s) = \psi_{iR}^*(s(i))$.

In light of the monotone properties in Definition C.2, it suffices to show that

$$s - s(i) < \epsilon, \tag{D.22}$$

which will then imply (D.20). Obviously, if (iii) is true then (D.22) must be true in light of (D.21). We will focus on showing (D.22) in the case of (i) since the proof for the case of (ii) follows a parallel argument.

By Lemma D.1(ii), there exists a $u_0 = u_0(\underline{s}) > 0$ such that

$$(\psi_{iL}(s), \psi_{iR}(s)) \subset [u_0, 1 - u_0] \text{ for all } i \in \mathbb{N} \text{ and } s \geqslant \underline{s}. \tag{D.23}$$

Consider each $b_i(u) = b(u; \pi_i, k_i)$ as a function in (u, π_i, k_i), and let

$$\nabla_{\pi,k} b_i(u) := \left(\frac{\partial}{\partial \pi_i} b(u; \pi_i, k_i)^T, \frac{\partial}{\partial k_i} b(u; \pi_i, k_i)^T \right)^T$$

be the gradient of b_i with respect to (π_i, k_i) evaluated at u. Using the compactness of $[u_0, 1-u_0] \times \Theta \times \mathcal{B}$ and Assumption 2(i) again, one can find a constant $C(u_0) > 0$ such that the gradient bounds

$$\|\nabla_{\pi,k} b_i(u)\|_1 < C(u_0) \text{ for all } i \text{ and for all } u \in \left[\frac{u_0}{2}, 1 - \frac{u_0}{2} \right]. \tag{D.24}$$

On the other hand, without loss of generality, with Lemma D.1(iv), let $\tilde{\epsilon} > 0$ be a small enough constant such that

$$\lambda(b_i^{*-1}([m_i^*, m_i^* + \tilde{\epsilon})) < \frac{\epsilon}{2} \text{ for all } i. \tag{D.25}$$

By the mean-value theorem and the gradient bound (D.24), one can then find $\delta > 0$ such that when $\|\Pi - \Pi^*\|_\infty \vee \|K - K^*\|_\infty < \delta$

$$|b_i(u) - b_i^*(u)| < \tilde{\epsilon} \text{ for all } u \in [u_0, 1 - u_0]. \tag{D.26}$$

Since $\psi_{iL}^*(s(i)) = \psi_{iL}(s) \in [u_0, 1-u_0]$, (D.26) and the last property in Definition C.2 suggest that

$$b_i^*\Big(\psi_{iR}^*(s(i))\Big) = b_i^*\Big(\psi_{iL}^*(s(i))\Big) < b_i\Big(\psi_{iL}(s)\Big) + \tilde{\epsilon} = b_i\Big(\psi_{iR}(s)\Big) + \tilde{\epsilon}. \tag{D.27}$$

But since $\psi_{iR}(s)$ is also in the interval $[u_0, 1 - u_0]$, we must have

$$b_i^*\Big(\psi_{iR}(s)\Big) > b_i\Big(\psi_{iR}(s)\Big) - \tilde{\epsilon}. \tag{D.28}$$

Combining (D.27) and (D.28), we get that $2\tilde{\epsilon} > b_i^*\Big(\psi_{iR}^*(s(i))\Big) - b_i^*\Big(\psi_{iR}(s)\Big)$ which in light of the construction of $\tilde{\epsilon}$ in (D.25) and convexity properties from Lemma C.2 gives that

$$\psi_{iR}^*(s(i)) - \psi_{iR}(s) < \epsilon,$$

which in turn implies (D.22) by the property (C.5) in Definition C.2. The proof of (D.19) is similar. It amounts to showing

$$(\psi_{iL}^*(s + \epsilon), \psi_{iR}^*(s + \epsilon)) \subset (\psi_{iL}(s), \psi_{iR}(s)).$$

We will alternatively define

$$\tilde{s}(i) := \inf\{s' \in (0, 1) : (\psi_{iL}^*(s'), \psi_{iR}^*(s')) \subset (\psi_{iL}(s), \psi_{iR}(s))\}.$$

then show $\tilde{s}(i) - s < \epsilon$. We leave the details to the reader. □

D.3. **A Glivenko-Cantelli theorem.**

Lemma D.4 (Pre-Glivenko-Cantelli theorem)**.** *Under Assumptions 1-3 , for any $\epsilon > 0$ and positive numbers $0 < \underline{t} < \bar{t} < \min_i c_i^{*-1}(\bar{s})$, there exists $\delta = \delta(\epsilon) > 0$ such that, for sufficiently large m,*

$$\sup_{\substack{\max(\|K-K^*\|, \|\Pi-\Pi^*\|)<\delta \\ \underline{t} \leqslant t \leqslant \bar{t}}} \left| \frac{1}{m} \sum_{i=1}^m I(T_i \leqslant t) - G(t) \right| \leqslant \epsilon \tag{D.29}$$

$$\sup_{\substack{\max(\|K-K^*\|, \|\Pi-\Pi^*\|)<\delta \\ \underline{t} \leqslant t \leqslant \bar{t}}} \left| \frac{1}{m} \sum_{i=1}^m (1 - H_i) I(T_i \leqslant t) - \pi_0 G_0(t) \right| \leqslant \epsilon, \tag{D.30}$$

$$\sup_{\substack{\max(\|K-K^*\|, \|\Pi-\Pi^*\|)<\delta \\ \underline{t} \leqslant t \leqslant \bar{t}}} \left| \frac{1}{m} \sum_{i=1}^m I(S_i \geqslant 1 - c_i(t)) - \bar{G}_0(t) \right| \leqslant \epsilon \tag{D.31}$$

with probability 1.

Proof of Lemma D.4. In this proof, for any function $F(\cdot)$, $F(t-)$ denotes the left limit at the point t.

Proof of (D.29): Let

$$G_{\theta,\beta}(t) = \frac{1}{m}\sum_{i=1}^{m} I(T_i \leqslant t),$$

where the subscript emphasizes that the T_i's are defined with an unspecified (θ, β), to distinguish from G in Assumption 3. Let n be large enough such that $1/n < \epsilon/2$ and consider $G^{\leftarrow}(1/n) \leqslant \cdots \leqslant G^{\leftarrow}(n/n)$. If we define

$$n' := \min\{i : \underline{t} < G^{\leftarrow}(i/n) < \bar{t}, i = 1, \ldots, n\},$$

$$d' := |\{i : \underline{t} < G^{\leftarrow}(i/n) < \bar{t}, i = 1, \ldots, n\}|.$$

Define $t_1 := G^{\leftarrow}(n'/n), t_2 := G^{\leftarrow}((n'+1)/n), \ldots, t_{d'} := G^{\leftarrow}((n'+d'-1)/n)$, as well as $t_0 = \underline{t}$ and $t_d = \bar{t}$ with $d = d' + 1$. Following the proof of the Glivenko-Cantelli theorem in Resnick (2019, p.224), we have

(D.32)

$$\sup_{\underline{t} \leqslant t \leqslant \bar{t}} |G_{\theta,\beta}(t) - G(t)| \leqslant \left(\bigvee_{v=0}^{d} |G_{\theta,\beta}(t_v) - G(t_v)| \vee |G_{\theta,\beta}(t_v-) - G(t_v-)| \right) + 1/n$$

We will first bound the terms of the form $|G_{\theta,\beta}(t_v) - G(t_v)|$ in (D.32). The strong law of large numbers for $G(\cdot)$ in Assumption 3 suggests that

$$|G_{\theta,\beta}(t_v) - G(t_v)| \leqslant \left| m^{-1} \sum_{i=1}^{m} (I(T_i \leqslant t_v) - I(T_i^* \leqslant t_v)) \right| + R_v, \tag{D.33}$$

where the remainder term $R_v \longrightarrow 0$ almost surely. Now, realizing $\{T_i^* \leqslant t_v\} = \{S_i^* \leqslant c_i^*(t_v)\}$, by Lemma D.2 and $\underline{t} > 0$, pick $\delta > 0$ such that

$$\left| \frac{1}{m} \sum_{i=1}^{m} (I(T_i \leqslant t_v) - I(T_i^* \leqslant t_v)) \right| \leqslant \tag{D.34}$$

$$\underbrace{\left| m^{-1} \sum_{i=1}^{m} [I(S_i^* \leqslant c_i^*(t_v) + \epsilon) - I(S_i^* \leqslant c_i^*(t_v))] \right|}_{(A)} \vee \underbrace{\left| m^{-1} \sum_{i=1}^{m} [I(S_i^* \leqslant c_i^*(t_v) - \epsilon) - I(S_i^* \leqslant c_i^*(t_v))] \right|}_{(B)}.$$

for $\max(\|\Pi - \Pi^*\|, \|K - K^*\|) < \delta$. This is because $\sum_{i=1}^{m} [I(S_i^* \leqslant c_i^*(t_v) + \epsilon) - I(S_i^* \leqslant c_i^*(t_v))] \geqslant \sum_{i=1}^{m} (I(T_i \leqslant t_v) - I(T_i^* \leqslant t_v))$ if the latter term is greater than 0; likewise, $\sum_{i=1}^{m} [I(S_i^* \leqslant c_i^*(t_v) - \epsilon) - I(S_i^* \leqslant c_i^*(t_v))] \leqslant \sum_{i=1}^{m} (I(T_i \leqslant t_v) - I(T_i^* \leqslant t_v))$ if the latter is less than 0.

We will first develop a bound for term (A). One have

$$\begin{aligned}(A) &\leqslant \mathbb{E}\Big|\frac{1}{m}\sum_{i=1}^{m}\big[P(S_i^* > c_i^*(t_v)|X_i) - P(S_i^* > c_i^*(t_v)+\epsilon|X_i)\big]\Big| + Q_v^+ \\ &= \mathbb{E}\Big|\frac{1}{m}\sum_{i=1}^{m}\big[P\big(U_i \in (\psi_{iL}^*(c_i^*(t_v)), \psi_{iR}^*(c_i^*(t_v))|X_i\big) - \\ &\qquad\qquad P(U_i \in (\psi_{iL}^*(c_i^*(t_v)+\epsilon), \psi_{iR}^*(c_i^*(t_v)+\epsilon)|X_i)\big]\Big| + Q_v^+ \\ &\leqslant C(\underline{t})\epsilon + Q_v^+ \qquad \text{(D.35)}\end{aligned}$$

where $Q_v^+ = o_{a.s.}(1)$ is a remainder term coming from the strong law of $G(\cdot)$ in Assumption 3. The second equality comes from Definition C.2. Note that the intervals

$$(\psi_{iL}^*(c_i^*(t_v)), \psi_{iR}^*(c_i^*(t_v)), i = 1, \ldots, m$$

can be equivalently represented as

$$(\omega_{iL}^*(t_v), \omega_{iR}^*(t_v)), i = 1, \ldots, m$$

by Definition C.1, which all belong to a compact sub-interval in $(0, 1)$ by Lemma D.1(i) and the fact that $\underline{t} > 0$. As such, Assumption 2(ii) can be applied to give the last inequality (D.35). By realizing, from Definitions C.1 and C.2, that the event

$$\{S_i^* > c_i^*(t_v) - \epsilon\}$$

is equivalent to U_i belonging to an interval that is $(\omega_{iL}^*(t_v), \omega_{iR}^*(t_v))$ expanded by a further ϵ width, it is obvious that one can analogously develop the bound

$$(B) \leqslant C(\underline{t})\epsilon + Q_v^- \qquad \text{(D.36)}$$

for a constant $C(\underline{t})$ and $Q_v^- = o_{a.s.}(1)$. Combining (D.33), (D.35) and (D.36) gives (by appropriately adjusting ϵ)

$$|G_{\theta,\beta}(t_v) - G(t_v)| \leqslant \epsilon/2 \text{ a.s.}, \qquad \text{(D.37)}$$

for sufficiently large m. A similar bound

$$|G_{\theta,\beta}(t_v-) - G(t_v-)| \leqslant \epsilon/2 \text{ a.s.}, \qquad \text{(D.38)}$$

can be derived in much the same way with no difficulty by first writing

$$G_{\theta,\beta}(t_v-) - G(t_v-) = m^{-1}\sum_{i=1}^{m} I(T_i < t_v) - G(t_v)$$

using the continuity of $G(\cdot)$, and the proof is omitted for brevity. Combining (D.32), (D.37) and (D.38) give (D.29). (D.30) can be proved in the same way by first writing

$$\sum_{i=1}^{m}(1 - H_i)I(T_i \leqslant t) = \sum_{H_i=0} I(T_i \leqslant t),$$

and noting that U_i is uniformly distributed given $H_i = 0$ and X_i, and thus omitted.

Proving (D.31) is similar. One proceed by developing the bound

$$\left|\frac{1}{m}\sum_i [I(S_i > 1 - c_i(t_v)) - P(S_i^* > 1 - c_i^*(t_v))]\right|$$
$$\leqslant \left|\frac{1}{m}\sum_i [I(S_i^* > 1 - c_i^*(t_v) - \epsilon) - P(S_i^* > 1 - c_i^*(t_v))]\right| \vee$$
$$\left|\frac{1}{m}\sum_i [I(S_i^* > 1 - c_i^*(t_v) + \epsilon) - P(S_i^* > 1 - c_i^*(t_v))]\right|$$

with Lemma D.3 for t_v that are now quantiles of $\bar{G}_0$, for sufficiently small $\delta > 0$ and $\max(\|\Pi - \Pi^*\|, \|K - K^*\|) < \delta$. From Definition C.2, the events $\{S_i^* > 1 - c_i^*(t_v)\}$ have the form

$$\{U_i \in (\psi_{iL}^*(1 - c_i^*(t_v)), \psi_{iR}^*(1 - c_i^*(t_v)))\},$$

and to show that the intervals $(\psi_{iL}^*(1 - c_i^*(t_v)), \psi_{iR}^*(1 - c_i^*(t_v))$ can be placed in a compact sub-interval of $(0, 1)$ by Lemma D.1(ii) to apply Assumption 2(ii), one need to show that $c_1^*(t_v), \ldots, c_m^*(t_v)$ are bounded away from 1. This is true because $c_i^*(t_v) \leqslant c_i^*(\bar{t}) \leqslant \bar{s} < 1$ for all i by definition. The same proof rundown goes through, again, by realizing that $(\psi_{iL}^*(1 - c_i^*(t_v) - \epsilon), \psi_{iR}^*(1 - c_i^*(t_v) - \epsilon)$ are just ϵ-expansion of $(\psi_{iL}^*(1 - c_i^*(t_v)), \psi_{iR}^*(1 - c_i^*(t_v))$ from Definition C.2. The rest of the proof goes thru with no resistance.

□

Lemma D.5 (Glivenko-Cantelli Theorems)**.** *Under Assumptions 1-3, for any* $0 < \underline{t} < \bar{t} \leqslant \min_i {c_i^*}^{-1}(\bar{s})$*, we have*

$$\sup_{\underline{t} \leqslant t \leqslant \bar{t}} \left|\frac{1}{m}\sum_{i=1}^m I(\hat{T}_i \leqslant t) - G(t)\right| \longrightarrow 0, \tag{D.39}$$

$$\sup_{\underline{t} \leqslant t \leqslant \bar{t}} \left|\frac{1}{m}\sum_{i=1}^m (1 - H_i) I(\hat{T}_i \leqslant t) - \pi_0 G_0(t)\right| \longrightarrow 0, \tag{D.40}$$

$$\sup_{\underline{t} \leqslant t \leqslant \bar{t}} \left|\frac{1}{m}\sum_{i=1}^m I(\hat{S}_i \geqslant 1 - \hat{c}_i(t)) - \bar{G}_0(t)\right| \longrightarrow 0, \tag{D.41}$$

almost surely.

Proof of Lemma D.5. Let $\hat{\theta} = (\hat{\theta}_l^T, \hat{\theta}_r^T)^T, \hat{\beta} = (\hat{\beta}_l^T, \hat{\beta}_r^T)^T$, and define $\hat{\Pi} = \Pi(\hat{\theta})$ and $\hat{K} = K(\hat{\beta})$. We will first show that

$$\|\hat{\Pi} - \Pi^*\|_\infty \vee \|\hat{K} - K^*\|_\infty \longrightarrow 0 \text{ a.s.}, \tag{D.42}$$

which is a consequence of

$$\hat{\theta} \longrightarrow \theta^* \text{ and } \hat{\beta} \longrightarrow \beta^* \quad a.s. \tag{D.43}$$

by the mean value theorem, the compactness of $\Theta \times \boldsymbol{B}$ and Assumption 2(i). To show (D.43), it suffices to bound

$$|\log[(1-\pi_{li}-\pi_{ri})+\pi_{li}h_{li}(U_i)+\pi_{ri}h_{ri}(U_i)]| = |\log h_{X_i}(U_i)| \tag{D.44}$$

by an integrable function in U_i that doesn't depend on (θ,β) (White, 1981, Theorem 2.1). We first let $u_0>0$ be as in Lemma D.1(iii). By the compactness of $\Theta\times\boldsymbol{B}\times[u_0,1-u_0]$ and Assumption 2(i), there exists universal constants $C,c>0$ such that $\log(c)<0$, $\log(C)>0$, and

$$C \geqslant h_{X_i}(U_i) \geqslant c$$

for all (θ,β,X_i) whenever $U_i\in[u_0,1-u_0]$. Note that Lemma C.2 also implies that

$$h_{X_i}(U_i)\geqslant c \text{ for } U_i\in(0,u_0)\cup(1-u_0,1).$$

Moreover, for $U_i\in(0,u_0)$, $\pi_{ri}h_{ri}(U_i)\leqslant C$ by Lemma C.1, hence borrowing notations from Section D.2,

$$\begin{aligned}\log h_{X_i}(U_i) &\leqslant \log\left(1+C+\bar{\pi}\frac{(1-U_i)^{\gamma_l-1}U_i^{\underline{k}-1}}{B_{\min}}\right)\\ &\leqslant \log(2+C)+\log\left(\tilde{C}\frac{(1-U_i)^{\gamma_l-1}U_i^{\underline{k}-1}}{(1-u_0)^{\gamma_l-1}u_0^{\underline{k}-1}}\right)\equiv \underbrace{m_l(U_i)}_{>0},\end{aligned}$$

for a constant $\tilde{C}>1\vee\frac{\bar{\pi}}{B_{\min}}$. Similarly, there exists a positive function $m_r(\cdot)$ such that

$$\log h_{X_i}(U_i)\leqslant m_r(U_i)$$

for $U_i\in(1-u_0,1)$. Combining these facts we have for all $U_i\in(0,1)$,

$$|\log h_{X_i}(U_i)|\leqslant |\log(c)|\vee\log(C)\vee m_r(U_i)\vee m_l(U_i),$$

where the right hand side is integrable by Assumption 2 (iii) and (D.43) is proved.

We will only prove (D.39), and (D.40) and (D.41) can be shown the same way. Let $D_m = D_m(\omega) = \sup_{\underline{t}\leqslant t\leqslant\bar{t}}\left|\frac{1}{m}\sum_{i=1}^m I(\hat{T}_i\leqslant t)-G(t)\right|$, where ω denotes a point in the underlying probability space Ω. It suffices to show that for any $\epsilon>0$, there exists a subspace $\Lambda(\epsilon)\subset\Omega$ such that $P(\Lambda(\epsilon))=1$ and $D_m(\omega)<\epsilon$ for sufficiently large m and every $\omega\in\Lambda(\epsilon)$. By Lemma D.4, there exists Λ_1 with $P(\Lambda_1)=1$ such that (D.29) holds on Λ_1 for $\delta(\epsilon)>0$. By (D.42), there exists Λ_2 with $P(\Lambda_2)=1$ such that $\|\hat{\Pi}-\Pi\|_\infty\vee\|\hat{K}-K\|_\infty<\delta(\epsilon)$ on Λ_2 for sufficiently large m. Take $\Lambda(\epsilon)=\Lambda_1\cap\Lambda_2$. □

D.4. **Proof of Theorem 3.1.** The proof is similar to that of Storey et al. (2004, Theorem 4) but is a bit more sutble. Recall the ratio in (3.8). We shall first show that under all the assumptions of Theorem 3.1, for any $\underline{t}>0$ and any $\bar{t}\in(t_\alpha^\infty,\min_i {c_i^*}^{-1}(\bar{s})]$,

$$\liminf_{m\to\infty}\inf_{\underline{t}\leqslant t\leqslant\bar{t}}\left\{\widehat{\mathrm{FDP}}_{asymp}(t)-FDP(t)\right\}\geqslant 0 \text{ a.s.} \tag{D.45}$$

where $FDP(t)\equiv\frac{\sum_i(1-H_i)I(\hat{T}_i\leqslant t)}{\left(\sum_i I(\hat{T}_i\leqslant t)\right)\vee 1}$ for any $t>0$. From the first two Glivenko-Cantelli statements in Lemma D.5, we see that

$$\text{(D.46)}\quad \lim_{m\to\infty}\sup_{\underline{t}\leqslant t\leqslant\bar{t}}\left|FDP(t)-\frac{\pi_0 mG_0(t)}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant t)\right)\vee 1}\right|$$

$$=\lim_{m\to\infty}\sup_{\underline{t}\leqslant t\leqslant\bar{t}}\left|\frac{\sum_i(1-H_i)I(\hat{T}_i\leqslant t)}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant t)\right)\vee 1}-\frac{\pi_0 mG_0(t)}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant t)\right)\vee 1}\right|$$

$$\leqslant\lim_{m\to\infty}\left|\frac{m}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant \underline{t})\right)\vee 1}\right|\sup_{\underline{t}\leqslant t\leqslant\bar{t}}\left|\frac{\sum_i(1-H_i)I(\hat{T}_i\leqslant t)}{m}-\pi_0 G_0(t)\right|=0 \text{ a.s.}$$

since $\lim_{m\to\infty}\left|\frac{m}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant\underline{t})\right)\vee 1}\right| = 1/G(\underline{t}) < \infty$ almost surely, given that $G(\underline{t}) > 0$. On the other hand, it must be that $\bar{G}_0(t) \geqslant \pi_0 G_0(t)$ considering that $P(S_i^* \geqslant 1-c_i^*(t)|H_i=0) = P(T_i^*\leqslant t|H_i=0)$, which, together with the last Glivenko-Cantelli statement in Lemma D.5, gives

$$\liminf_m \inf_{\underline{t}\leqslant t\leqslant\bar{t}}\left\{\frac{\sum_{i=1}^m I(\hat{S}_i\geqslant 1-\hat{c}_i(t))}{m}-\pi_0 G_0(t)\right\}\geqslant 0.$$

The preceding display and (D.46) will lead to

$$\liminf_{m\to\infty}\inf_{\underline{t}\leqslant t\leqslant\bar{t}}\left\{\widehat{\text{FDP}}_{asymp}(t)-FDP(t)\right\}$$

$$\geqslant\liminf_{m\to\infty}\inf_{\underline{t}\leqslant t\leqslant\bar{t}}\left\{\frac{\sum_{i=1}^m I(\hat{S}_i\geqslant 1-\hat{c}_i(t))}{\left(\sum_{i=1}^m I(\hat{T}_i\leqslant t)\right)\vee 1}-FDP(t)\right\}\geqslant 0,$$

which is (D.45).

Towards finishing, we will establish that, almost surely,

$$\text{(D.47)}\qquad \liminf_m \hat{t}_{asymp}(\alpha) > 0 \text{ and } \limsup_m \hat{t}_{asymp}(\alpha)\leqslant \min_i c_i^{*-1}(\bar{s}).$$

Fix $t_1\in(t_\alpha^\infty, \min_i c_i^{*-1}(\bar{s}))$. By the definition of t_α^∞ in Assumption 3 it must be the case that

$$\frac{\bar{G}_0(t_1)}{G(t_1)} > \alpha = \frac{\bar{G}_0(t_\alpha^\infty)}{G(t_\alpha^\infty)}$$

and we can let $\frac{\bar{G}_0(t_1)}{G(t_1)}-\alpha=\epsilon_1>0$. For sufficiently large m, because of Lemma D.5 we can get that $\left|\frac{\sum_i I(\hat{S}_i\geqslant 1-\hat{c}_i(t_1))}{\left(\sum_i I(\hat{T}_i\leqslant t_1)\right)\vee 1}-\frac{\bar{G}_0(t_1)}{G(t_1)}\right| < \epsilon_1/2$ a.s., which implies that $\widehat{\text{FDP}}_{asymp}(t_1) > \alpha$ almost surely to give the "limsup" statement in (D.47). On the other hand, let $\alpha-\frac{\bar{G}_0(t_0)}{G(t_0)}=\epsilon_0>0$ for t_0 in Assumption 3. Since $t_0 < t_\alpha^\infty$ (by continuity of the "G" functions), Lemma D.5 also suggests that for large enough m, $|\frac{\bar{G}_0(t_0)}{G(t_0)}-\widehat{\text{FDP}}_{asymp}(t_0)| < \epsilon_0/2$ almost surely, which implies $\widehat{\text{FDP}}_{asymp}(t_0) < \alpha$ almost surely and hence the "liminf" statement in (D.47).

Now given (D.47) is true, since $\widehat{\text{FDP}}_{asymp}(\hat{t}_{asymp}(\alpha)) \leqslant \alpha$, by (D.45) it must be true that

$$\limsup_m FDP(\hat{t}_{asymp}(\alpha)) \leqslant \alpha \text{ a.s.}$$

By the reverse Fatou's lemma, this implies

$$\limsup_m \mathbb{E}[FDP(\hat{t}_{asymp}(\alpha))] \leqslant \mathbb{E}\left[\limsup_m FDP(\hat{t}_{asymp}(\alpha))\right] \leqslant \alpha,$$

and Theorem 3.1 is proved.

Appendix E. Proof for the finite-sample method

The proof is almost exactly the same as that of Lei and Fithian (2018, Theorem 1) which relies on the key lemma in that paper (Lei and Fithian, 2018, Lemma 2), and we will only define the notation required to apply their argument. First, for each $t = 0, 1, \ldots,$ let $\mathbb{V}_t = \#\{i : U_i \in \mathcal{R}_t \text{ and } H_i = 0\}$ and $\mathbb{U}_t = \#\{i : U_i \in \mathcal{A}_t \text{ and } H_i = 0\}$ which are respectively the numbers of true nulls in the rejection set and acceptance set at step t. Define

$$m_i = I(U_i \geqslant 0.5)(\breve{U}_i \vee U_i) + I(U_i < 0.5)(\breve{U}_i \wedge U_i)$$

and

$$b_i = I(0.25 \leqslant U_i \leqslant 0.75)$$

so that

$$U_i = b_i\{I(m_i \geqslant 0.5)(1.5 - m_i) + I(m_i < 0.5)(0.5 - m_i)\} + (1 - b_i)m_i.$$

Also define $\mathcal{C}_t = \{i : i \in \mathcal{A}_t \cup \mathcal{R}_t \text{ and } H_i = 0\}$ to give

$$\mathbb{U}_t = \sum_{i \in \mathcal{C}_t} b_i \text{ and } \mathbb{V}_t = \sum_{i \in \mathcal{C}_t} (1 - b_i) = |\mathcal{C}_t| - \mathbb{U}_t.$$

If we set the initial sigma-algebra $\mathcal{G}_{-1} = \sigma\{(X_i, m_i)_{i \in \{1,\ldots,m\}}, (b_i)_{i:H_i \neq 0}\}$, then $P(b_i = 1|\mathcal{G}_{-1}) = 0.5$ almost surely for a null i under the uniform null distribution of U_i. With these ingredients, the arguments in the proof of Lei and Fithian (2018, Theorem 1) will follow line by line, where the U_i's will take the role of the p-values in that paper.

Appendix F. Supplementary algorithms

We will inherit the simplified notation in Appendix C. The *complete data log-likelihood* for Model (3.6), treated as a function of $\{\theta, \beta\}$, has the form

$$\begin{aligned} l(\theta, \beta) = &\sum_{i=1}^{m} H_{li}\left[(k_{li} - 1)\log(U_i) + (\gamma_l - 1)\log(1 - U_i) - \log B(k_{li}, \gamma_l)\right] + \\ &\sum_{i=1}^{m} H_{ri}\left[(k_{ri} - 1)\log(1 - U_i) + (\gamma_r - 1)\log(U_i) - \log B(\gamma_r, k_{ri})\right] + \\ &\left[\sum_{i=1}^{m}(1 - H_{li} - H_{ri})\log(1 - \pi_{li} - \pi_{ri}) + H_{li}\log(\pi_{li}) + H_{ri}\log(\pi_{ri})\right], \end{aligned} \tag{F.1}$$

where for each i, H_{li} and H_{ri} are Bernoulli random variables with respective success probabilities π_{li} and π_{ri}, and H_{li} and H_{ri} cannot be both equal to 1 at the same time. Note that the last line in (F.1) amounts to a multinomial logistic regression with three classes.

F.1. **EM algorithm for asymptotic ZAP.**

Algorithm 3: EM algorithm for asymptotic ZAP

Data: $U_1, \ldots, U_m, X_1, \ldots, X_m$

Input: initial guess $\beta^{(0)}$, $\theta^{(0)}$

while $(\beta^{(j)}, \theta^{(j)})$ *not converged* **do**

E step: Let $\pi_{li}^{(j)}$, $\pi_{ri}^{(j)}$, $h_{li}^{(j)}$, $h_{ri}^{(j)}$ and $h_{X_i}^{(j)}$ be as defined in Section 3.2 evaluated at $(\beta^{(j)}, \theta^{(j)})$. Compute

$$Q^{(j)}(\beta, \theta) = \sum_{i=1}^{m} \left\{ w_{li}^{(j)} \log[\pi_{li} h_{li}(U_i)] + w_{ri}^{(j)} \log[\pi_{ri} h_{ri}(U_i)] + (1 - w_{li}^{(j)} - w_{ri}^{(j)}) \log(1 - \pi_{li} - \pi_{ri}) \right\},$$

where

$$w_{li}^{(j)} = \mathbb{E}_{\beta^{(j)}, \theta^{(j)}}[H_{li} \mid X_i, U_i] = \frac{\pi_{li}^{(j)} \cdot h_{li}^{(j)}(U_i)}{h_{X_i}^{(j)}(U_i)},$$

$$w_{ri}^{(j)} = \mathbb{E}_{\beta^{(j)}, \theta^{(j)}}[H_{ri} \mid X_i, U_i] = \frac{\pi_{ri}^{(j)} \cdot h_{ri}^{(j)}(U_i)}{h_{X_i}^{(j)}(U_i)}.$$

M step: Compute $(\beta^{(j+1)}, \theta^{(j+1)}) = \arg\max_{\beta, \theta} Q^{(j)}(\beta, \theta))$.

end

Output: Estimated coefficients β and θ

F.2. **Updating the thresholding functions in finite-sample ZAP.** We recommend using Algorithm 4 below to update the thresholding functions, which performs estimations of our beta-mixture model, although finite-sample FDR control is guaranteed as long as the conditions in the Theorem 3.2 are met. As seen in Algorithm 4, assessor functions for the hypotheses are first constructed based on expression (3.7), using an EM algorithm that acts on the masked data $\{\tilde{U}_{t,i}, X_i\}_{i=1}^m$ (Appendix F.3) to estimate the parameters. Next, for each masked $i \in \mathcal{A}_t \cup \mathcal{R}_t$, evaluated assessor value T_i' at whichever U_i or $\breve{U}_i$ is closer to the extreme ends of the interval $(0, 1)$ is computed, and among them the hypothesis j with the largest such value is selected. This step aims to locate the hypothesis in the current masked set that is the most likely to be a true null if all masked hypotheses are presumed to be from the candidate rejection set $\mathcal{R}_t$. Finally, one of the two thresholding functions $s_{l,t}$ and $s_{r,t}$ will be updated, in a manner that satisfies condition (ii) in Theorem 3.2, to give a different $s_{l,t+1}$ or $s_{r,t+1}$: If $U_j > 0.5$, $s_{r,t+1}$ will be updated from $s_{r,t}$ at the point X_j as $s_{r,t+1}(X_j) = U_j \vee \breve{U}_j$, and remains the same at all other covariate values; otherwise, $s_{l,t+1}$ will update from $s_{l,t}$ in a similar fashion using the value $U_j \wedge \breve{U}_j$. As such, at the next step $t + 1$, one of $\mathcal{A}_{t+1}$ or $\mathcal{R}_{t+1}$ will be shrunk

by exactly one element which is j. This is intuitive since if $\widehat{\mathrm{FDP}}_{finite}(t) > \alpha$ at step t, one would hope to reduce the size of $\mathcal{R}_t$.

Algorithm 4: Update thresholding functions at step t with Model (3.6)

Input: The masked data $\{\tilde{U}_{t,i}, X_i\}_{i=1}^m$

Compute $\{\hat{\theta}_l, \hat{\theta}_r, \hat{\beta}_l, \hat{\beta}_r\}$ using the EM algorithm in Appendix F.3.

Construct $\{\hat{a}_{X_i}(\cdot)\}_{i\in\mathcal{A}_t\cup\mathcal{R}_t}$ with (3.7) by setting the underlying parameters as the estimates in the prior step.

Find $j \equiv \arg\max_{i\in\mathcal{A}_t\cup\mathcal{R}_t} T_i'$ for $T_i' = \hat{a}_{X_i}(U_i')$, where

$$U_i' \equiv I(U_i < 0.5)U_i \wedge \breve{U}_i + I(U_i \geqslant 0.5)U_i \vee \breve{U}_i$$

if $U_j > 0.5$ **then**

  $s_{rt+1}(X_i) = I(i=j)(U_i \vee \breve{U}_i) + I(i\neq j)s_{rt}(X_i)$, $s_{l,t+1} \equiv s_{l,t}$;

else

  $s_{lt+1}(X_i) = I(i=j)(U_i \wedge \breve{U}_i) + I(i\neq j)s_{lt}(X_i)$, $s_{r,t+1} \equiv s_{r,t}$;

end

Output: s_{lt+1}, $s_{r,t+1}$

F.3. **EM algorithm for finite-sample ZAP.** We will lay out aspects of the EM algorithm required for Algorithm 4.

E-step computations. Let $\mathcal{D}_{ti} = (X_i, \tilde{U}_{t,i})$ be the available data for i at step t of the finite-sample ZAP algorithm. To update from the parameters $(\theta^{(j)}, \beta^{(j)})$ at the j-th EM iteration, we need to compute the following quantities:

$$\begin{aligned}(\text{F.2})\quad &\mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}|\mathcal{D}_{ti}],\ \ \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}|\mathcal{D}_{ti}],\\ &\mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}\log(U_i)|\mathcal{D}_{ti}],\ \ \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}\log(1-U_i)|\mathcal{D}_{ti}],\\ &\mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}\log(U_i)|\mathcal{D}_{ti}],\ \ \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}\log(1-U_i)|\mathcal{D}_{ti}].\end{aligned}$$

These quantities are straightforward to compute when $\tilde{U}_{t,i}$ is a singleton, so we will only focus on computing them when $\tilde{U}_{t,i}$ is a two-element set, i.e. corresponding to a masked U_i. Like Algorithm 3, we shall let $\pi_{li}^{(j)}$, $\pi_{ri}^{(j)}$, $h_{li}^{(j)}$, $h_{ri}^{(j)}$ and $h_{X_i}^{(j)}$ be as defined in Section 3.2 evaluated at $(\beta^{(j)}, \theta^{(j)})$. We will have

$$\hat{H}_{li}^{(j)} \equiv \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}|\mathcal{D}_{ti}] = P_{(\theta^{(j)},\beta^{(j)})}[H_{li}=1|\mathcal{D}_{ti}] = \frac{\pi_{li}^{(j)}[h_{li}^{(j)}(U_i) + h_{li}^{(j)}(\breve{U}_i)]}{h_{X_i}^{(j)}(\breve{U}_i) + h_{X_i}^{(j)}(U_i)}$$

$$\hat{H}_{ri}^{(j)} \equiv \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}|\mathcal{D}_{ti}] = P_{(\theta^{(j)},\beta^{(j)})}[H_{ri}=1|\mathcal{D}_{ti}] = \frac{\pi_{ri}^{(j)}[h_{ri}^{(j)}(U_i) + h_{ri}^{(j)}(\breve{U}_i)]}{h_{X_i}^{(j)}(\breve{U}_i) + h_{X_i}^{(j)}(U_i)}.$$

Moreover, to express the last four quantities in (F.2), we define

$$y_{li,A}^{(j)} \equiv \mathbb{E}_{\theta^{(j)},\beta^{(j)}}[H_{li}\log(U_i)|\mathcal{D}_{ti}, H_{li}=1] = \frac{h_{li}^{(j)}(U_i)\log(U_i) + h_{li}^{(j)}(\breve{U}_i)\log(\breve{U}_i)}{h_{li}^{(j)}(U_i) + h_{li}^{(j)}(\breve{U}_i)},$$

$$y_{ri,A}^{(j)} \equiv \mathbb{E}_{\theta^{(j)},\beta^{(j)}}[H_{ri}\log(U_i)|\mathcal{D}_{ti}, H_{ri}=1] = \frac{h_{ri}^{(j)}(U_i)\log(U_i) + h_{ri}^{(j)}(\breve{U}_i)\log(\breve{U}_i)}{h_{li}^{(j)}(U_i) + h_{li}^{(j)}(\breve{U}_i)},$$

$$y_{li,B}^{(j)} \equiv \mathbb{E}_{\theta^{(j)},\beta^{(j)}}[H_{li}\log(1-U_i)|\mathcal{D}_{ti}, H_{li}=1] = \frac{h_{li}^{(j)}(U_i)\log(1-U_i) + h_{li}^{(j)}(\breve{U}_i)\log(1-\breve{U}_i)}{h_{li}^{(j)}(U_i) + h_{li}^{(j)}(\breve{U}_i)},$$

$$y_{ri,B}^{(j)} \equiv \mathbb{E}_{\theta^{(j)},\beta^{(j)}}[H_{ri}\log(1-U_i)|\mathcal{D}_{ti}, H_{ri}=1] = \frac{h_{ri}^{(j)}(U_i)\log(1-U_i) + h_{ri}^{(j)}(\breve{U}_i)\log(1-\breve{U}_i)}{h_{ri}^{(j)}(U_i) + h_{ri}^{(j)}(\breve{U}_i)},$$

then one can express

$$\mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}\log(U_i)|\mathcal{D}_{ti}] = y_{li,A}^{(j)}\hat{H}_{li}^{(j)}, \quad \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}\log(U_i)|\mathcal{D}_{ti}] = y_{ri,A}^{(j)}\hat{H}_{ri}^{(j)},$$
$$\mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{li}\log(1-U_i)|\mathcal{D}_{ti}] = y_{li,B}^{(j)}\hat{H}_{li}^{(j)}, \quad \mathbb{E}_{(\theta^{(j)},\beta^{(j)})}[H_{ri}\log(1-U_i)|\mathcal{D}_{ti}] = y_{ri,B}^{(j)}\hat{H}_{ri}^{(j)}.$$

Initialization. We now discuss how to specify values for $\pi_{li}^{(0)}, \pi_{ri}^{(0)}$, $\beta_l^{(0)}$ and $\beta_r^{(0)}$ to initialize the algorithm. Specifying $\beta_l^{(0)}$ and $\beta_r^{(0)}$ is easy: For β_l, we only consider the left group $\mathfrak{L} = \{i : \breve{U}_i \vee U_i \leqslant 0.5\}$, and fit the left-leaning beta density $h_{li}(\cdot)$ to the data points $\{U_i \wedge \breve{U}_i : i \in \mathfrak{L}\}$ to obtain an estimate for β_l as the initial value $\beta_l^{(0)}$, with a given value for γ_l (such as 4). Note that we fit the model to the smaller point $U_i \wedge \breve{U}_i$ instead of $U_i \vee \breve{U}_i$ for each $i \in \mathfrak{L}$ with the goal of having a more "aggressive" left alternative distribution. The initial value $\beta_r^{(0)}$ can be obtained similarly by considering the right group $\mathfrak{R} \equiv \{i : \breve{U}_i \wedge U_i > 0.5\}$, the data $\{U_i \vee \breve{U}_i : i \in \mathfrak{R}\}$ and $h_{ri}(\cdot)$.

To specifiy $\pi_{li}^{(0)}, \pi_{ri}^{(0)}$, for each i, we first define

(F.3)

$$\pi_{ri}^{+} \equiv P(H_{ri}=1|X_i, U_i > 0.5), \quad \pi_{li}^{+} \equiv P(H_{li}=1|X_i, U_i > 0.5), \quad \pi_i^{+} \equiv P(U_i > 0.5|X_i),$$
$$\pi_{li}^{-} \equiv P(H_{li}=1|X_i, U_i \leqslant 0.5), \quad \pi_{ri}^{-} \equiv P(H_{ri}=1|X_i, U_i \leqslant 0.5), \quad \pi_i^{-} \equiv P(U_i \leqslant 0.5|X_i)$$

and note that, by definition, $\pi_{ri} \geqslant \pi_{ri}^{+}\pi_i^{+}$ and $\pi_{li} \geqslant \pi_{li}^{-}\pi_i^{-}$. We will form estimates for $\hat{\pi}_{ri}^{+}, \hat{\pi}_i^{+}, \hat{\pi}_{li}^{-}, \hat{\pi}_i^{-}$ and let $\pi_{ri}^{(0)} = \hat{\pi}_{ri}^{+}\hat{\pi}_i^{+}$ and $\pi_{li}^{(0)} = \hat{\pi}_{li}^{+}\hat{\pi}_i^{+}$ be conservative estimates for π_{ri} and π_{li}. The estimates $\hat{\pi}_i^{+}$ and $\hat{\pi}_i^{-}$ can be obtained as predicted probabilities by fitting a logistic regression on the indicator responses $D_i \equiv I(U_i > 0.5)$ with covariates X_i. For the rest of this section we will focus on the estimates $\hat{\pi}_{ri}^{+}$, since the estimates $\hat{\pi}_{li}^{-}$ can be obtain analogously.

Let $J_i \equiv I(\tilde{U}_{0,i}$ has one element). Then

$$\mathbb{E}[J_i|X_i, D_i=1] = P(J_i=1|X_i, D_i=1) \geqslant (1-\pi_{ri}^{+}-\pi_{li}^{+})\left(\frac{0.5-2(1-s_{r,0}(X_i))}{0.5}\right)$$

which is equivalent to

$$\pi_{ri}^{+} + \pi_{li}^{+} \geqslant \mathbb{E}\left[\tilde{J}_i \middle| X_i, D_i=1\right],$$

where $\tilde{J}_i \equiv 1 - \frac{0.5 J_i}{0.5 - 2(1 - s_{r,0}(X_i))}$ and

$$\mathbb{E}\left[\tilde{J}_i \middle| X_i, D_i = 1\right] = \underbrace{P(\tilde{U}_{0,i} \text{ has two elements } | X_i, D_i = 1)}_{\equiv 1 - \pi^+_{J_i=1}} + \underbrace{P(\tilde{U}_{0,i} \text{ has one element } | X_i, D_i = 1)}_{\pi^+_{J_i=1}} \left(1 - \frac{0.5}{0.5 - 2(1 - s_{r,0}(X_i))}\right)$$

As the probability of $H_{li} = 1$ should be small under $U_i > 0.5$, π^+_{li} is likely to be negligible, hence the right hand side of the previous display should still be a conservative estimate for π^+_{ri}, i.e.

$$\pi^+_{ri} \approx (1 - \pi^+_{J_i=1}) + \pi^+_{J_i=1} \left(1 - \frac{0.5}{0.5 - 2(1 - s_{r,0}(X_i))}\right).$$

Now estimates for $\pi^+_{J_i=1}$ for any $i \in \{1, \ldots, m\}$ can be obtained as the fitted probability of the logistic regression on J_i with covariates X_i restricted to samples with $U_i > 0.5$.

Appendix G. Further numerical results

G.1. **Further simulations for Section 3.5.** We will perform extra simulations to test how our model in Section 3.2 can robustly estimate the non-null probabilities. We generate 8000 i.i.d. z-values from a normal mixture model with density

$$(1 - w) \times f_0(z) + \underbrace{(w \cdot (1 - \rho))}_{\equiv w_l} \times \phi(z; \mu_l, 1) + \underbrace{(w \cdot \rho)}_{\equiv w_r} \times \phi(z; \mu_r, 1), \tag{G.1}$$

where the simulation parameters μ_l, μ_r, w, ρ range as

$$\mu_l \in \{-2.5, -2, -1.5, -1, -0.5\}, \quad \mu_r \in \{0.5, 1, 1.5, 2, 2.5\}, \quad w \in \{0.1, 0.15, 0.2\}, \quad \rho \in \{0.5, 0.7, 0.9\}.$$

Apparently, w is the non-null probability, μ_l and μ_r are respectively the mean parameters for the alternative normals on the left and right, and ρ parametrizes the degree of asymmetry reflected in the mixing probabilities w_l and w_r. For each set of 8000 z-values, the beta mixture model (3.6) for $(\gamma_l, \gamma_r) = (4, 4)$ is fitted with regression intercepts only by an EM algorithm, and the resulting left and right model-based non-null probabilities $\hat{\pi}_l$ and $\hat{\pi}_r$ estimates serve as estimates for w_l and w_r.

From the results in Tables 2 - 4, one can see that our beta mixture produces fairly accurate non-null probability estimates $(\hat{\pi}_l, \hat{\pi}_r)$ for (w_l, w_r). Generally speaking, the estimates are the most inaccurate when one of μ_l or μ_r has a small magnitude, which is reasonable since one of the two non-null components has a weak signal and many z-values which are non-nulls could be regarded as null by the EM fitting

TABLE 2. Estimated probabilities $\hat{\pi}_l$ and $\hat{\pi}_r$ based on the beta mixture (3.6) with $(\gamma_l, \gamma_r) = (4, 4)$ and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with $\rho = 0.5$. $\hat{\pi}_l$ and $\hat{\pi}_r$ are respectively the left and right entries in each cell.

$\mu_l \backslash \mu_r$	0.5		1		1.5		2		2.5	
	$w = 0.1$, $(w_l, w_r) = (0.05, 0.05)$									
-2.5	0.073	0.056	0.066	0.055	0.064	0.064	0.071	0.065	0.080	0.069
-2	0.061	0.037	0.062	0.038	0.057	0.058	0.059	0.061	0.071	0.079
-1.5	0.040	0.022	0.048	0.037	0.042	0.047	0.049	0.065	0.058	0.070
-1	0.044	0.029	0.037	0.039	0.034	0.046	0.047	0.064	0.066	0.072
-0.5	0.023	0.024	0.024	0.047	0.041	0.073	0.043	0.072	0.041	0.076
	$w = 0.15$, $(w_l, w_r) = (0.075, 0.075)$									
-2.5	0.110	0.050	0.103	0.072	0.100	0.087	0.106	0.095	0.102	0.095
-2	0.085	0.045	0.087	0.058	0.089	0.070	0.086	0.091	0.100	0.096
-1.5	0.080	0.052	0.076	0.065	0.082	0.064	0.074	0.082	0.082	0.094
-1	0.065	0.040	0.044	0.041	0.041	0.064	0.068	0.087	0.068	0.102
-0.5	0.022	0.024	0.049	0.065	0.050	0.075	0.041	0.071	0.050	0.091
	$w = 0.2$, $(w_l, w_r) = (0.1, 0.1)$									
-2.5	0.116	0.073	0.118	0.101	0.123	0.119	0.127	0.132	0.122	0.136
-2	0.112	0.053	0.113	0.086	0.115	0.127	0.113	0.132	0.120	0.133
-1.5	0.089	0.047	0.090	0.074	0.086	0.104	0.096	0.124	0.086	0.135
-1	0.067	0.042	0.055	0.068	0.060	0.095	0.071	0.117	0.090	0.134
-0.5	0.025	0.046	0.042	0.083	0.032	0.106	0.068	0.124	0.069	0.142

algorithm. This should not be too concerning, as it simply means some of the hypotheses pose hard testing problems to begin with.

TABLE 3. Estimated probabilities $\hat{\pi}_l$ and $\hat{\pi}_r$ based on the beta mixture (3.6) with $(\gamma_l, \gamma_r) = (4, 4)$ and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with $\rho = 0.7$. $\hat{\pi}_l$ and $\hat{\pi}_r$ are respectively the left and right entries in each cell.

$\mu_l \backslash \mu_r$	0.5		1		1.5		2		2.5	
	$w = 0.1$, $(w_l, w_r) = (0.03, 0.07)$									
-2.5	0.049	0.043	0.037	0.046	0.047	0.060	0.050	0.079	0.058	0.091
-2	0.046	0.042	0.036	0.059	0.037	0.072	0.046	0.077	0.042	0.089
-1.5	0.065	0.042	0.023	0.044	0.028	0.072	0.051	0.076	0.037	0.089
-1	0.036	0.041	0.057	0.061	0.034	0.072	0.049	0.086	0.047	0.087
- 0.5	0.031	0.040	0.037	0.050	0.042	0.077	0.036	0.074	0.030	0.091
	$w = 0.15$, $(w_l, w_r) = (0.045, 0.105)$									
-2.5	0.066	0.054	0.057	0.058	0.062	0.106	0.058	0.124	0.070	0.137
-2	0.047	0.051	0.057	0.079	0.058	0.112	0.067	0.124	0.057	0.134
-1.5	0.043	0.055	0.044	0.075	0.034	0.091	0.048	0.118	0.050	0.128
-1	0.053	0.058	0.024	0.069	0.049	0.100	0.053	0.129	0.069	0.136
-0.5	0.038	0.057	0.034	0.076	0.034	0.111	0.035	0.125	0.045	0.135
	$w = 0.2$, $(w_l, w_r) = (0.06, 0.14)$									
-2.5	0.078	0.070	0.079	0.116	0.088	0.152	0.079	0.159	0.088	0.185
-2	0.062	0.050	0.071	0.104	0.072	0.143	0.075	0.164	0.073	0.181
-1.5	0.049	0.052	0.058	0.114	0.059	0.146	0.059	0.164	0.074	0.184
-1	0.029	0.041	0.028	0.088	0.046	0.128	0.075	0.174	0.087	0.192
-0.5	0.031	0.059	0.021	0.106	0.026	0.137	0.050	0.164	0.059	0.186

TABLE 4. Estimated probabilities $\hat{\pi}_l$ and $\hat{\pi}_r$ based on the beta mixture (3.6) with $(\gamma_l, \gamma_r) = (4, 4)$ and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with $\rho = 0.9$. $\hat{\pi}_l$ and $\hat{\pi}_r$ are respectively the left and right entries in each cell.

$\mu_l \backslash \mu_r$	0.5		1		1.5		2		2.5	
	$w = 0.1$, $(w_l, w_r) = (0.01, 0.09)$									
-2.5	0.019	0.044	0.017	0.072	0.015	0.093	0.030	0.100	0.026	0.114
-2	0.037	0.065	0.008	0.064	0.017	0.102	0.009	0.108	0.035	0.123
-1.5	0.022	0.050	0.011	0.071	0.018	0.086	0.034	0.117	0.053	0.120
-1	0.019	0.044	0.019	0.066	0.021	0.079	0.026	0.112	0.041	0.125
-0.5	0.010	0.050	0.011	0.062	0.019	0.092	0.035	0.120	0.049	0.116
	$w = 0.15$, $(w_l, w_r) = (0.015, 0.135)$									
-2.5	0.024	0.060	0.022	0.096	0.024	0.115	0.023	0.149	0.024	0.164
-2	0.020	0.071	0.021	0.098	0.020	0.128	0.017	0.150	0.035	0.167
-1.5	0.029	0.080	0.015	0.092	0.025	0.141	0.035	0.148	0.035	0.156
-1	0.021	0.060	0.026	0.112	0.017	0.124	0.044	0.150	0.034	0.163
-0.5	0.011	0.065	0.017	0.105	0.012	0.127	0.018	0.139	0.032	0.158
	$w = 0.2$, $(w_l, w_r) = (0.02, 0.18)$									
-2.5	0.018	0.059	0.024	0.134	0.021	0.162	0.026	0.194	0.039	0.223
-2	0.015	0.046	0.017	0.125	0.026	0.162	0.020	0.196	0.043	0.226
-1.5	0.012	0.072	0.010	0.134	0.027	0.154	0.039	0.214	0.032	0.216
-1	0.014	0.077	0.010	0.120	0.021	0.175	0.037	0.213	0.035	0.225
-0.5	0.016	0.104	0.013	0.115	0.023	0.167	0.033	0.209	0.049	0.230

G.2. **Extended numerical results for Section 4.** For good measure we have also experimented with the following methods:

(a) BH: the vanilla BH procedure (Benjamini and Hochberg, 1995)
(b) oracle: The oracle procedure $\boldsymbol{\delta}^{\mathcal{Z}}$
(c) SABHA: Structure adaptive BH procedure with $\tau = 0.5$, $\epsilon = 0.1$ and stepwise constant weights (Li and Barber, 2019)
(d) BL: Boca and Leek procedure (Boca and Leek, 2018).

Similar to IHW, SABHA is a method only applicable to univariate covariates. Hence for the simulations in Section 4.1, it is applied with the covariate sum $X_{\bullet i}$; for the applications in Section 4.2, it is applied with the original log mean normalized read count for the RNA-seq data and not applied for the neural data. Of course, the oracle procedure is only applicable to simulated data.

The FDR and TPR plots for the whole set of methods are shown in Figure G.1 for the simulated data. Note that except for the oracle procedure, all the additional methods are p-value based, and they cannot dominate the z-value based methods in power as expected. In Setup 3, none among the extended set of experimented methods has power comparable to the oracle procedure, suggesting that the data generating mechanism poses a hard multiple testing problem. One additional observation is that FDRreg in fact has slightly more power than the oracle procedure in Setup 1 when the covariates are the most informative. Of course, this has come at the expense of violating the FDR bound.

The numbers of rejections for the extended list of methods applied to the real data are shown in Figure G.2, and the conclusions we can arrive at are essentially the same as those from Figure 4.2 in the main text. Lastly, one can refer to Lei and Fithian (2018) for access to the `bottomly` and `airway` datasets. The other two real datasets are available at:

(i) `hippo`: `https://www.raynamharris.com/DissociationTest/`
(ii) `scott`: `https://github.com/jgscott/FDRreg`

G.3. **Differential expression analysis of RNA-Seq data.** We shall first briefly discuss the importance of pre-filtering genes with excessively low read counts before applying FDR methodologies for differential expression (DE) analysis of RNA-Seq data. For the unfamiliar reader, a good open resource on the relevant analysis pipeline can be found on `https://github.com/hbctraining/DGE_workshop`. It typically begins with a raw "count matrix" with the expression read counts as entries, where each row corresponds to a mapped gene and each column corresponds to a sample/library that is either in the treatment or the control group. This count matrix is taken as an input to a suite of statistical analysis tools available from one of the `R` packages for DE analysis that differ by their underlying modelling assumptions, to produce test statistics that are re-scaled measures of differential expression between the two groups for all the genes involved. The two most popular such R pipelines which can produce the z-values considered by the current paper are `limma` (with the "voom" function therein) (Ritchie et al., 2015) and `DESeq2` Love et al. (2014). `limma` operates with a linear model to produce t-statistics, and

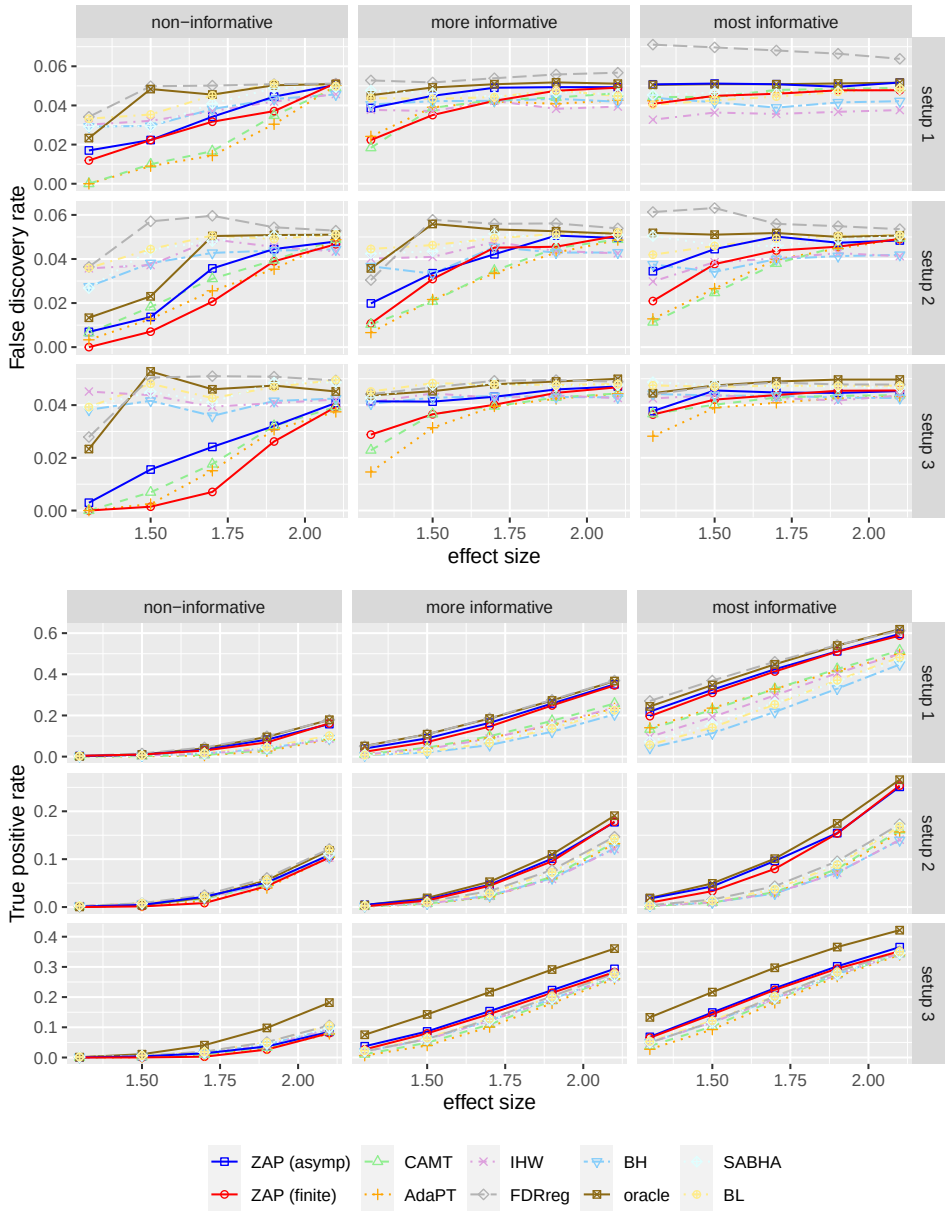

FIGURE G.1. FDR and TPR performances of an extended list of methods under **Setup 1 - 3**. All methods are applied at a targeted FDR level of 0.05. The x-axes show the values of ϵ. `non-informative`, `more informative` and `most informative` correspond to different values of ζ from the smallest to the largest.

`DESeq2` operates with a negative binomial model to produce Wald statistics. These

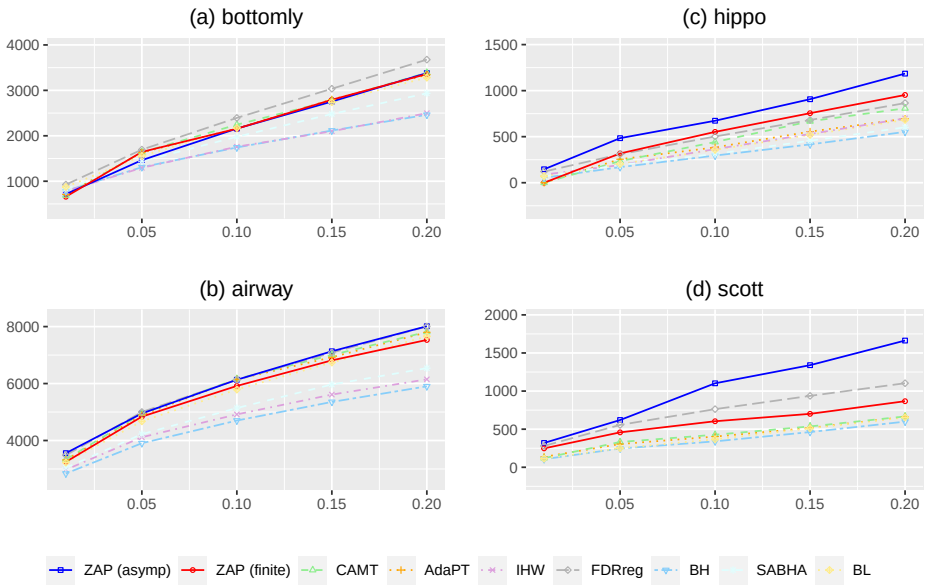

FIGURE G.2. (a)-(d) plot the numbers of rejections for different methods across datasets, against targeted FDR level at $0.01, 0.05, 0.1, 0.15, 0.2$.

primary statistics can then undergo the further transformation in Section 3.1 to give the u-values, on which our ZAP methods can be applied.

However, without suitable pre-processing, a raw count matrix will typically produce unusual distributions for the u-values (or p-values). Figure G.3 plots histograms of the u-values produced by the original raw count matrices of the three RNA-Seq datasets in the main text processed with `DESeq2`. Normally, if the test statistics are well-calibrated, the null u-values should be approximately uniformly distributed, and one should only expect spikes near the two ends of the interval $(0, 1)$ (or only close to 0 if the histogram is for two-sided p-values) which represent genes that are non-null. This is clearly not the case in Figure G.3, and the spurious spikes in the middle of the unit interval for all three histograms are typically results of genes that have excessive low read counts for which reliable DE analysis is impossible and can at best be considered as nulls. In particular, the presence of such spikes will make a procedure like the BH overly conservative. A standard practice is to filter out these genes according to some rules of thumb which have been discussed by Chen et al. (2016) in some length. In the analysis of the main text, we have adopted a simple convention of filtering out genes with a total raw counts less than 15 using the function `filterByExpr` in the R package `edgeR`, which implements the method in Chen et al. (2016). Apparently, spurious structures in the u-value histograms have been more or less removed as a result, as is evident by comparing Figure G.3 (c) and Figure 4.2(e), the latter of which has its u-values produced by the filtered version of `hippo` dataset.

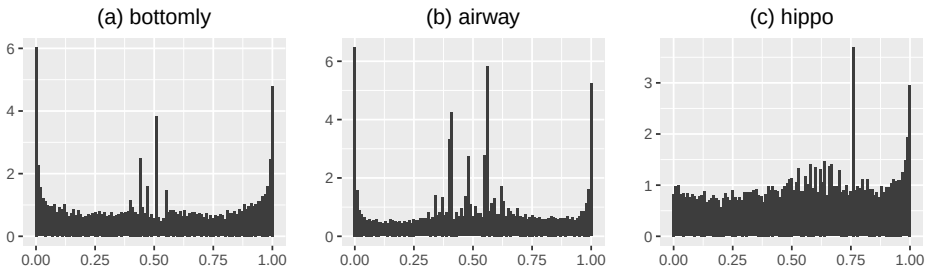

FIGURE G.3. Histograms of u-values for the original unfiltered versions of the three RNA-Seq datasets in the main text. In principle, one should only see at most two "spikes" on the two ends of the interval $(0, 1)$. Spikes not located close to 0 or 1 in any histogram result from genes with extremely low read counts.

SCHOOL OF MATHEMATICS AND STATISTICS, UNIVERSITY OF MELBOURNE
Email address: dennis.leung@unimelb.edu.au

DEPARTMENT OF DATA SCIENCES AND OPERATIONS, USC MARSHALL SCHOOL OF BUSINESS
Email address: wenguans@marshall.usc.edu