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 misspecied. We demonstrate the state-of-the-art performance of ZAP using both simulated and real data, which shows that the eciency gain can be substantial in comparison with p-value based methods. Our methodology is implemented in the R package zap.

1. Introduction

In modern scientic studies, a ubiquitous task is to test a multitude of two-sided hypotheses regarding the presence of nonzero eects. The problem of multiple test-ing with covariates has received much recent attention, as leveraging contextual information beyond what is oered 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 dierential 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 eciency of existing methods. The importance of this direction has been reected 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 Classication. 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 test-ing, 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 eectively incorporates both the primary statistics and covariates into inference. ZAP is deployed with a working model, whose potential misspecication will not inval-idate 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, lling an important gap in the literature. Particularly, we propose the rst z-value based procedure with nite-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 θ_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

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

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:

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

where $w_x = \mathbb{P}(H_i = 1 \mid X_i = x)$ is the conditional probability of having a non-zero effect given $X_i = x$ and $f_{1,x}(z) = f(z \mid H_i = 1, X_i = x)$ 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 $f_0(z) = \varphi(z)$, the density of a $N(0,1)$ 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 $\mathcal{R} \subseteq \{1, \ldots, m\}$ be the set of hypotheses rejected by a multiple testing procedure. In large-scale testing problems, the widely used FDR is defined as

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

where $V = \sum_{i=1}^m (1 - H_i)\mathbb{1}(i \in \mathcal{R})$ and $R = \sum_{i=1}^m \mathbb{1}(i \in \mathcal{R})$ are respectively the number of false positives and the number of rejections. Throughout, $\mathbb{E}_{\mathrm{rs}}$ 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 $\mathrm{ETD} = \mathbb{E}_{\mathrm{rs}}[R - V]$ or the true positive rate

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

Our goal is to devise a powerful procedure that can control the FDR under a prespecified 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 $N(0,1)$ 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 $\Phi^{-1} \circ G_0$ if the primary test statistic has a known null distribution function G_0, e.g. a t-distribution, where Φ 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 $X_i \sim \mathrm{Unif}(-1, 1)$. Our study examines three situations:

Example 2.1 Asymmetric alternatives: $f(z|x) = \frac{8}{10} f_0(z) + \frac{2}{10} f_1(z - 1.5)$.

Example 2.2 Unbalanced covariate effects on the non-null proportions:

$$f(z|x) = 0.8 f_0(z) + \frac{1}{2}\left(\frac{1-x}{10}\right) f_1(z - 1.5) + \frac{1}{2}\left(\frac{1+x}{10}\right) f_1(z + 1.5).$$

Example 2.3 Unbalanced covariate effects on the alternative means:

$$f(z|x) = 0.9 f_0(z) + 0.1 f_1(z - 1.5\,\mathrm{sgn}(x)), \quad \text{where } \mathrm{sgn}(x) = 1(x > 0) - 1(x < 0).$$

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

$$\delta^P = \{1(P(H_i = 0|P_i, X_i) \le t_P)\}_{i=1}^m \quad \text{and} \quad \delta^Z = \{1(P(H_i = 0|Z_i, X_i) \le t_Z)\}_{i=1}^m,$$

where the rejection decisions are expressed by indicators, and the thresholds t_P and t_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: $\mathrm{TPR}_P = 4.4\%$; $\mathrm{TPR}_Z = 11.7\%$.
Example 2.2: $\mathrm{TPR}_P = 3.4\%$; $\mathrm{TPR}_Z = 5.5\%$.
Example 2.3: $\mathrm{TPR}_P = 0.6\%$; $\mathrm{TPR}_Z = 2.6\%$.

Apparently, δ^Z is more powerful than δ^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

$$Z_i - P(S(X_i)) \in R \tag{2.3}$$

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

[2] In practice these posterior probabilities are unknown.

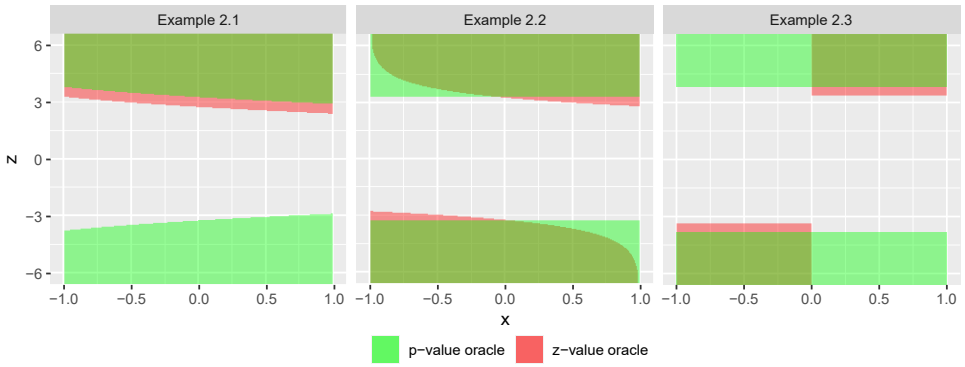

Figure 2.1. Comparisons of rejection regions. For each covariate value $x \in [-1, 1]$, the rejection regions $S^P(x)$ and $S^Z(x)$ are respectively marked in green and red on the z-value scale. The overlapped region is depicted in yellow.

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 $S^Z(x)$ changes with x, while $S^P(x)$ is completely insensitive to the changes in x; see Appendix A.2 for the relevant calculations. Hence, the covariates are only informative for ^{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 ^{P} in fully capturing the original interactive information between Z_i and X_i. When the latter interactive structure represents the bulk of the informa-tion 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 ^{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 mod-elling 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 ^{Z} while circumventing the direct estimation of $P(H_i = 0|Z_i, X_i)$. Next we first out-line 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: $U_i = p(Z_i)$. The transformed statistic U_i is referred to as a u-value[3], which, according to (2.2), obeys the induced mixture model

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

with $h_0(u)$ and $h_{1;x}(u) = h(u|H_i = 1; X_i = x)$ 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)

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

[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 $\mathrm{CLfdr}_x(u) : (0,1] \to (0,1]$ 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 $f_x(z)$ (or equivalently $h_x(z)$), 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 $\mathrm{Clfdr}_x(u)$ 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 $c_i(t) \equiv P(T_i \le t \mid H_i = 0)$ 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 Candes, 2015):

$$\widehat{\mathrm{FDP}}(t) \equiv \frac{1 + \#\{i : S_i \ge 1 - c_i(t)\}}{\#\{i : T_i \le t\} \vee 1} \equiv \frac{1 + \#\{i : T_i^m \le t\}}{\#\{i : T_i \le t\} \vee 1}; \tag{3.4}$$

[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, $S_i = c_i(T_i)$ is the probability of realizing a smaller significance index and $T_i = c_i^{-1}(1 - S_i)$ is the mirror statistic that "reflects" T_i's position in the distribution c_i. Define

$$\hat{t}(q) = \min\{t \in \{0, |T_1|, \ldots, |T_m|\} : \widehat{\mathrm{FDP}}(t) \le q\}, \tag{3.5}$$

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

However, the assessor $a_x(u)$, 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 Candes (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 $\hat{a}_x(u)$. From there we can test the hypotheses in a data-driven manner by either implementing the prototype procedure directly using $\hat{a}_x(u)$ 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 $\hat{a}_x(u)$, 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:

$$h_x(u) = (1 - \pi_{0;x} - \pi_{r;x})h_0(u) + \pi_{0;x}h_{0;x}(u) + \pi_{r;x}h_{r;x}(u), \tag{3.6}$$

where, given $X_i = x$, $\pi_{0;x}$ and $\pi_{r;x}$ respectively denote the mixing probabilities that $\theta_i = 0$ and $\theta_i > 0$[6], and $h_{0;x}$ 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 $\pi_{0;x}$ and $\pi_{r;x}$ are multinomial probabilities with regression parameter vectors β_0 and β_r:

$$\pi_{0;x} = \frac{\exp(\tilde{x}^T\beta_0)}{1 + \exp(\tilde{x}^T\beta_r) + \exp(\tilde{x}^T\beta_0)}, \quad \pi_{r;x} = \frac{\exp(\tilde{x}^T\beta_r)}{1 + \exp(\tilde{x}^T\beta_r) + \exp(\tilde{x}^T\beta_0)},$$

where $\tilde{x} = (1, x^T)^T$ is the intercept-augmented covariate vector. Further, $h_{0;x}$ and $h_{r;x}$ are chosen to be beta densities with regression parameters γ_0 and γ_r:

$$h_{0;x}(u) = \frac{1}{B(k_{0;x}, \gamma_0)} u^{k_{0;x}-1}(1-u)^{\gamma_0-1}, \quad h_{r;x}(u) = B(\gamma_r, k_{r;x})\, u^{\gamma_r-1}(1-u)^{k_{r;x}-1},$$

[6]The different symbols $\pi_{0;x}$ 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 $k_{l;x} \ t1 \ \exp(x^T_l u^1$ and $k_{r;x} \ t1 \ \exp(x^T_r u^1$, for two xed shape parameters $_l$ and $_r$. $h_{l;x}$ and $h_{r;x}$ are respectively left-leaning (right-skewed) and right-leaning (left-skewed) functions. We require that $_l \ i \ 2$ and $_r \ i \ 2$ 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 $t_{l;r}u$ 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 identied as a exible 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 dened 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 CLfdr$_x$pq function with respect to our working model (3.6). Since $h_0 \ 1$, it follows that

$$a_x(u) \ \frac{1 \ _{l;x} \ _{r;x}}{1 \ _{l;x} \ _{r;x} \ _{l;x} h_{l;x}(u) \ _{r;x} h_{r;x}(u)}; \quad 0 \ u \ 1: \tag{3.7}$$

The corresponding data-driven assessor is denoted by $\hat{a}_x$puq if the parameters $t_{l;r;l;r}u$ 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 $\hat{a}_x$puq, we rst obtain the maximum likelihood estimates (MLE) of the unknown regression parameters $t_{l;r;l;r}u$ with the data $tU_i; X_i u_{i1}$; the EM algorithm for their computations are provided in Appendix F.1. Denote $T_i \ \hat{a}_{X_i}(U_i)$, and let $\hat{c}$ pq be its null distribution by treating $\hat{a}_{X_i}$ pq as if it is pre-determined. With $\hat{S}_i \ \hat{c}_i(T_i)$, the estimated mirror statistics are correspondingly dened as $\hat{T}^m_i \ \hat{c}^1_i(1 \ S_i)$, which can be computed numerically by performing quantile estimation. The FDP for a candidate threshold t can be estimated as

$$\widehat{\mathrm{FDP}}_{\mathrm{asymp}}(t) \ \frac{1 \ \#\{i : T_i \ t\}}{\#\{i : \hat{T}^m_i \ t\} \ 1}. \tag{3.8}$$

Dene $\hat{t}_{\mathrm{asymp}}(\alpha) \ \sup\{0 \ t \ 1 : \widehat{\mathrm{FDP}}_{\mathrm{asymp}}(t) \ \alpha\}$, and reject i whenever $T_i \ \hat{t}_{\mathrm{asymp}}(\alpha)$. In practice, it suces to consider only the values of $T_1; \ldots; T_m$ as candidate thresholds. This procedure is summarized in Algorithm 1.

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

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

$$E\{\log(1 \ _{l;X_i} \ _{r;X_i}) \ _{l;X_i} h_{l;X_i}(U_i) \ _{r;X_i} h_{r;X_i}(U_i)\}$$

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 $T_i = \hat{a}_{X_i}(U_i)$ for each i.
2 Compute the mirror statistics $\{T_i, u_i\}_{i=1}^m$;
(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 T_i^m via e.g. `quantile()` in R.
3 Order $\{T_i, u_i\}_{i=1}^m$ as $T_{(1)} \le \cdots \le T_{(m)}$. Reject i if $T_i \le T_{(k)}$, where $k = \max\left\{0 \le t \le m : \frac{1 + \#\{i : T_i \le T_{(t)}, u_i > \hat{\tau}\}}{1 - \hat{\tau}} \le t\right\}$.

of the beta-mixture model (3.6) has a unique maximum at (τ, u) over $\mathcal{P}$ and $\mathcal{P}_B$ for compact spaces $\mathcal{P}$ and $\mathcal{B}$, where $p_\ell = (\pi^\tau, q^\tau)^\tau$ and $p^\tau = (\pi^\tau, q^\tau)^\tau$. 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 $\hat{a}(\cdot) = \hat{a}(\cdot; \hat{\tau}, \hat{u})$ be constructed with the MLE $(\hat{\tau}, \hat{u}) = \operatorname{argmax}_{\tau, u} \sum_{i=1}^m \log \hat{h}_{X_i}(U_i)$ of the beta-mixture model (3.6). Under Assumptions 1-3, the procedure that rejects i whenever $T_i \le \hat{t}_{\text{asymp}}$ controls the FDR asymptotically in the sense that $\limsup_{m \to \infty} \text{FDR} \le \alpha$.

We highlight two aspects of this result. First, it doesn't require the estimated assessor function to be a good proxy for $\text{CLfdr}_x(\cdot)$. 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

$$\left\{m^{-1} \sum_{i=1}^m \mathbb{1}\{T_i \le t\}\right\}_{t \in [0,1]} \quad \left\{m^{-1} \sum_{i=1}^m \mathbb{1}\{H_i = 1, T_i \le t\}\right\}_{t \in [0,1]} \quad \left\{m^{-1} \sum_{i=1}^m \mathbb{1}\left\{S_{\hat{a}}(\cdot) \le \hat{c}_i(t)\right\}\right\}_{t \in [0,1]},$$

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 (τ, u); 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 dened on the same space, $g_1 \preceq g_2$ means $g_1(x) \leq g_2(x)$ for all x in that space. If C is a constant, $g_1 \preceq C$ means $g_1(x) \leq C$ for all x. Similarly we can dene $g_1 \succeq g_2$ 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 rst divide the covariate values into a left and right group based on the observed u-values:

$$X_l = \{X_i : U_i \leq 0.5\} \text{ and } X_r = \{X_i : U_i > 0.5\}.$$

At each step $t = 0, 1, \ldots$, let $s_{l,t} : X_l \to [0, 0.25]$ and $s_{r,t} : X_r \to [0.75, 1]$ denote two corresponding thresholding functions, and dene the candidate rejection set $R_t = R_{l,t} \cup R_{r,t}$, where

$$R_{l,t} = \{i : U_i \leq s_{l,t}(X_i)\} \wedge 0.5 \text{ and } R_{r,t} = \{i : U_i \geq s_{r,t}(X_i)\} \vee 0.5. \tag{3.9}$$

Let $A_t = A_{l,t} \cup A_{r,t}$ be the corresponding set of "accepted" hypotheses, where

$$A_{l,t} = \{i : 0.5 - s_{l,t}(X_i) \leq U_i \leq 0.5\} \text{ and } A_{r,t} = \{i : 0.5 \leq U_i \leq 1.5 - s_{r,t}(X_i)\}.$$

Intuitively, $|A_{l,t}|$ estimates the number of false rejections in the left candidate rejection set $R_{l,t}$. Given $H_i = 0$ and $U_i \leq 0.5$, the events $\{U_i \leq s_{l,t}(X_i)\}$ and $\{U_i \geq 0.5 - s_{l,t}(X_i)\}$ are equally likely. The FDP of possibly rejecting R_t at step t can then be estimated as

$$\widehat{\mathrm{FDP}}_{\mathrm{finite}}(t) = \frac{1 + |A_t|}{|R_t| \vee 1}. \tag{3.10}$$

If $\widehat{\mathrm{FDP}}_{\mathrm{finite}}(t) \leq \alpha$, the algorithm terminates and the hypotheses in 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 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 $|R_t|$, $|A_t|$ and the partially masked data $\{(U_{t,i}, X_i)\}_{i=1}^m$ only, where

$$U_{t,i} = \begin{cases} U_i & \text{if } U_i \in A_t \cup R_t \\ \{U_i, 1 - U_i\} & \text{if } U_i \notin A_t \cup R_t \end{cases} \tag{3.11}$$

is a singleton or a two-element set depending on whether i is in the "masked" set $A_t \cup R_t$, and $1 - U_i$ is the "reection" 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:

$$1 - U_i = (1.5 - U_i)\mathbb{1}(U_i > 0.5) + (0.5 - U_i)\mathbb{1}(U_i \leq 0.5).$$

For example, if the underlying $U_i = 0.1$ and i is masked at step t, the algorithm can only update for $s_{l,t+1}$ and $s_{r,t+1}$ with the partial knowledge that U_i is either 0.1 or its reection 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 A_t, R_t and the unmasked set $\{1, \ldots, m\} \setminus (A_t \cup R_t)$ 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 $\{(U_i, X_i) : i \in A_t \cup R_t\}$ are also shown.

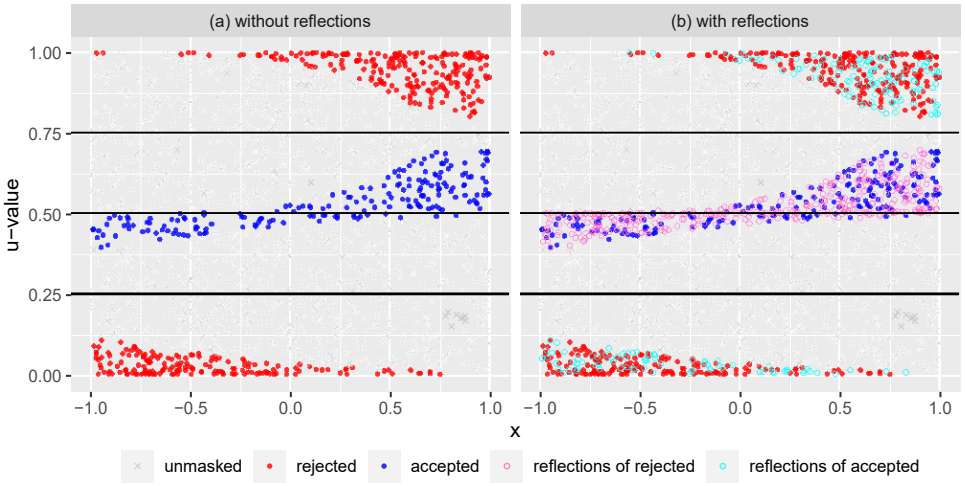

Figure 3.1. Illustration of Algorithm 2 at a step t based on Example 2.2, $m = 2000$. (a): The red, blue and grey are points in the respective sets R_t, A_t and $\{1, \ldots, m\} \setminus (A_t \cup R_t)$, where $\widehat{\mathrm{FDP}}_{\mathrm{finite}}(t) = \frac{1}{193} / \frac{389}{0.5}$. (b): The reflections of the points in R_t and A_t are respectively shown in pink and cyan. A pink 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 $s_{l,t-1} \geq s_{l,t}$ and (ii) $s_{r,t-1}$ and $s_{l,t-1}$ are updated based on $|R_t|$, $|A_t|$ and $\{(U_{t,i}, X_i)\}_{i=1}^m$ only, Algorithm 2 controls the FDR under for finite samples. Specifically, we have

$$\mathrm{E}\left[\mathrm{FDP}\{H_i, (X_i, U_i)_{i=1}^m\}\right] \leq .$$

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 $\{(X_i, U_i)\}_{i=1}^m$ such that $s_t \geq 0.5$, and the ratio $\frac{1 + |\{i : P_i \geq 1 - s_t(X_i)(U_i)\}|}{1 \vee |\{i : P_i \leq s_t(X_i)(U_i)\}|}$ is used as an FDP estimator for the candidate rejection set $\{i : P_i \leq s_t(X_i)(U_i)\}$. It is easy to see that

$$P_i \leq s_t(X_i)(U_i) \quad \text{if} \quad U_i \leq s_t(X_i)(U_i) \text{ or } U_i \geq 1 - s_t(X_i)(U_i). \tag{3.12}$$

Algorithm 2: Finite-sample ZAP

Data: $\{U_i; X_i\}_{i=1}^m$

Input: FDR level α, initial thresholding functions $s_{l;0} \equiv 0.5$ and $s_{r;0} \equiv 0.5$;

for $t = 0, 1, \ldots,$ do

  Compute $\widehat{\mathrm{FDP}}_{\mathrm{finite}}(t)$ in (3.10);

  if $\widehat{\mathrm{FDP}}_{\mathrm{finite}}(t) \le \alpha$ 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 R_t; break;

  end

end

Output: Reject all hypotheses in 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 (γ_l, γ_r) of the working model need to be pre-specified before running the EM algorithms. While requiring $\gamma_l, \gamma_r \ge 2$ 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

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

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),

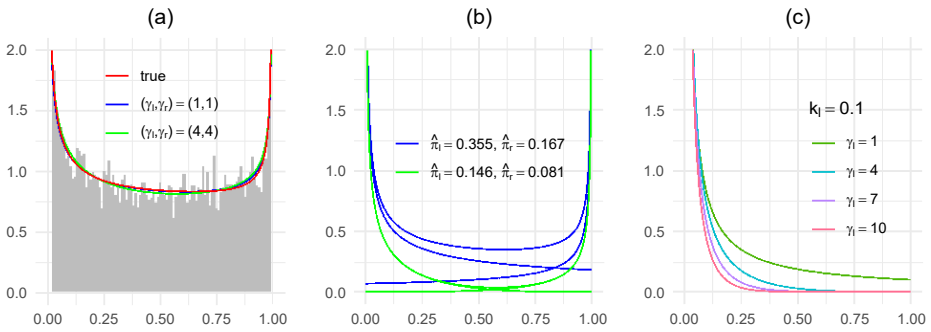

Figure 3.2. (a): Histogram of U_i pZ_iq generated by (3.13). The red curve is the true density; while the blue and green curves respectively correspond to the estimated densities of our beta-mixture model with $p_l; _rq$ set at p1; 1q and p4; 4q. (b): The long dashed blue curve is the entire non-null component of the solid blue estimated density in paq. 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 paq. The legend shows the estimated probabilities for the left and right-leaning components, where the subscript \x" is omitted from $\hat{}_l$ and $\hat{}_r$ as the t uses intercepts only. (c): Plot of the left-leaning beta density $Bpk_l; _lq^1u^{k_l1}p1\ uq^{l1}$ for k_l 0:1 and dierent values of $_l$.

we present the estimated quantities pertaining to the non-null components. We can see that setting $p_l; _rq$ p1; 1q has drastically overestimated the left and right non-null probabilities, whereas setting $p_l; _rq$ p4; 4q provides good approxima-tions 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

$$Bpk_l; _lq^1u^{k_l1}p1\ uq^{l1}$$

for dierent values of $_l$ and a xed k_l 0:1, which supposedly captures the negative eects in two-sided tests. We can see that small values of $_l$ 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 eects 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 p0; 1q akin to the U-shaped blue curve in Figure 3.2pbq. This inates the non-null probability estimates. In contrast, larger values of $_l$ and $_r$ eectively 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 $p_l; _rq$ p4; 4q 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 $s_{l0} = 0.2$ and $s_{r0} = 0.8$, but other values close to 0.25 and 0.75 tend to be equally effective. We also update the thresholding functions every $\lfloor 100s \rfloor$ 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 (nite): 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 $\lfloor 100s \rfloor$ 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 $N(0,1)$ 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 $N\left((0,0)^T, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$. Conditional on $X_i = x = (x_1, x_2)^T$, Z_i is generated with a normal mixture density

$$p(z) = (1 - w_{l;x} - w_{r;x}) f_0(z) + w_{l;x}\, p(z \mid \mu_{l;x}) + w_{r;x}\, p(z \mid \mu_{r;x}), \tag{4.1}$$

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

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

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

$$f_{1;x}(z) = \frac{w_{l;x}\varphi(z - \mu_{l;x}) + w_{r;x}\varphi(z - \mu_{r;x})}{w} \cdot \frac{1 - \mu_{l;x} - \mu_{r;x}}{w} \tag{4.2}$$

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 $X_i = X_{1i} + X_{2i}$ of the covariate components is $N(0, 1)$-distributed, and $x = x_1 + x_2$ will denote a realized value of it in what follows.

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

$$w_{r;x} = \frac{1}{1 + \exp(\beta x)}, \quad \mu_{r;x} = \frac{2\mu}{1 + \exp(\beta x)}, \quad w_{l;x} = 0, \quad \mu_{l;x} = 0,$$

with the simulation parameters ranging as

$$\beta \in \{0, 0.5, 1\}, \quad \mu \in \{1.3, 1.5, 1.7, 1.9, 2.1\} \text{ and } \mu = 2.$$

Since $w_{l;x} = 0$, all the non-null statistics come from the right centered alternative density $\varphi(z - \mu_{r;x})$. We briefly explain the simulation parameters. μ is an effect size parameter. Generally, β controls the informativeness of the covariates in relation to both the non-null probabilities and alternative means: when $\beta > 0$, a greater value of x makes the signals denser and stronger (i.e. $w_{r;x}$ and $\mu_{r;x}$ become larger). The value of β controls the sparsity levels. For example, when the covariates are non-informative at $\beta = 0$, setting $\mu = 2$ 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

$$w_{r;x} = \frac{\exp(\beta x)}{\exp(\beta) + \exp(\beta x) + \exp(\beta x)}, \quad w_{l;x} = \frac{\exp(\beta x)}{\exp(\beta) + \exp(\beta x) + \exp(\beta x)},$$

$\mu_{r;x} = \mu$ and $\mu_{l;x} = -\mu$. We fix $\mu = 2.5$ and vary other parameters in the range $\beta \in \{0, 0.7, 1\}$, $\mu \in \{1.3, 1.5, 1.7, 1.9, 2.1\}$.

Only $w_{l;x}$ and $w_{r;x}$ depend on the covariate value: for $\beta > 0$, $w_{r;x}$ increases and $w_{l;x}$ decreases as x increases, and vice versa as x 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 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, since both very negative and positive x 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 $\mu = 2$ corresponds to a baseline signal proportion of roughly 14% when $\beta = 0$.

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

$$w_{r;x} \quad 1 \quad \frac{1\{2}{\text{exppq}} ; w_{l;x} \quad 1 \quad \frac{1\{2}{\text{exppq}} ; {}_{r;x} \quad 1 \quad \frac{2"}{\text{exppxq}} ; {}_{l;x} \quad 1 \quad \frac{2"}{\text{exppxq}} \div$$

The simulation parameters range as

$$P\ t0; 1{:}5; 3u; \quad " \ P\ t1{:}3; 1{:}5; 1{:}7; 1{:}9; 2{:}1u \text{ and } \quad 2{:}$$

Our choice of corresponds to the a baseline signal proportion of roughly 12% when 0. When the covariates are informative (i 0), $_{r;x}$ and $_{l;x}$ respec-tively become more positive and less negative as x 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_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_i, which is an eective summary covariate in all three setups. The simulation results are reported in Figure 4.1, where the empirical FDR and TPR levels of dierent 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 nite-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 nite-sample ZAP is in its theoretical properties.
(b) Both asymptotic and nite-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 conicting 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 $f_{1;x}pzq$ f_1pzq 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,

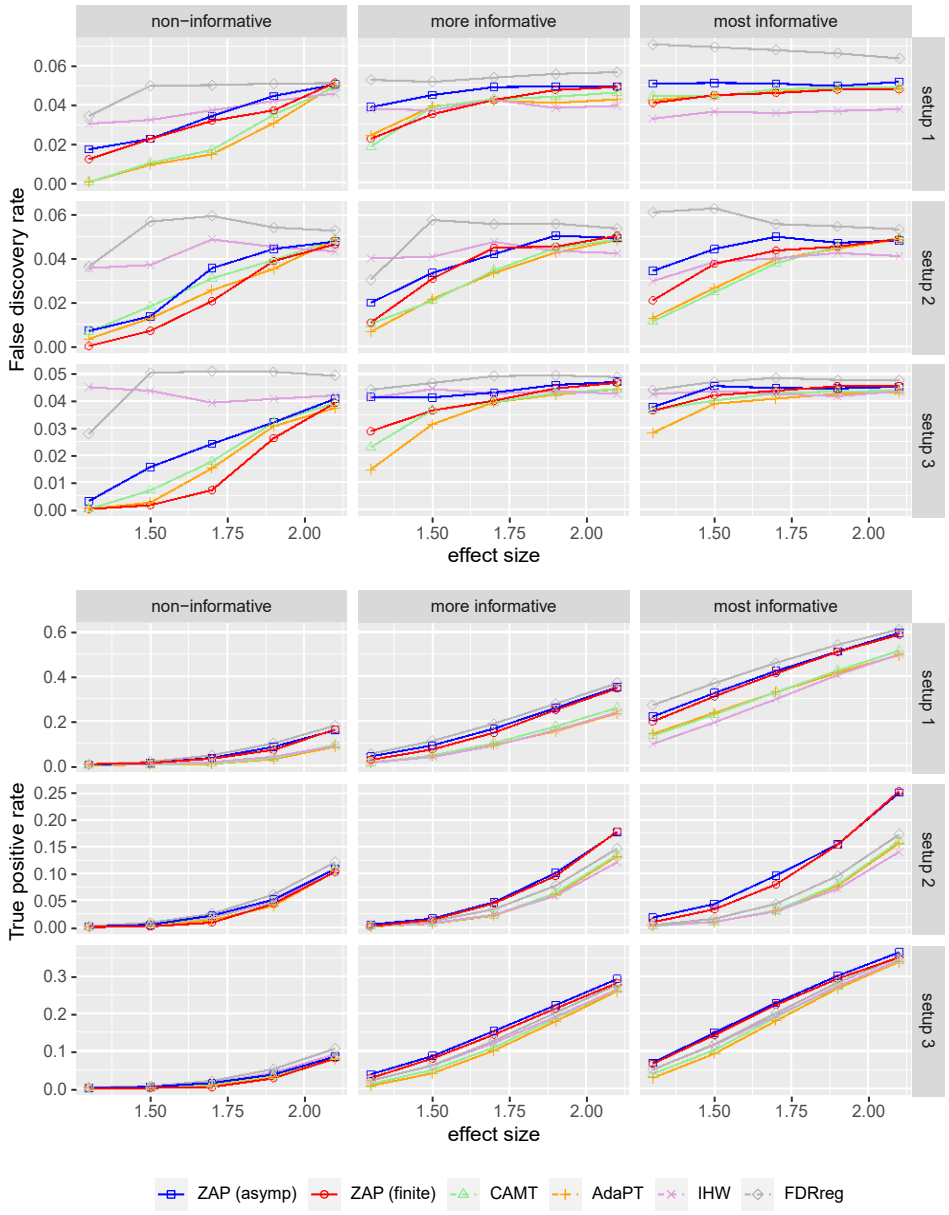

Figure 4.1. FDR and TPR performances of dierent methods under Setup 1 - 3 . All methods are applied at a targeted FDR level of 0:05. The x-axes show the values of ". non-informative, more informative and most informative correspond to dierent values of from the smallest to the largest.

airway, hippo) are generated by RNA sequencing (RNA-Seq) experiments for detecting dierential 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 ring 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 = Dierential Expression.

measures the observed dierence 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 dierential expressions in mouse hippocampus. For all datasets above, we have adopted the standard data pre-processing step, which lters out genes with excessively low read counts across samples before further downstream analyses (Chen et al., 2016) such as model tting 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.

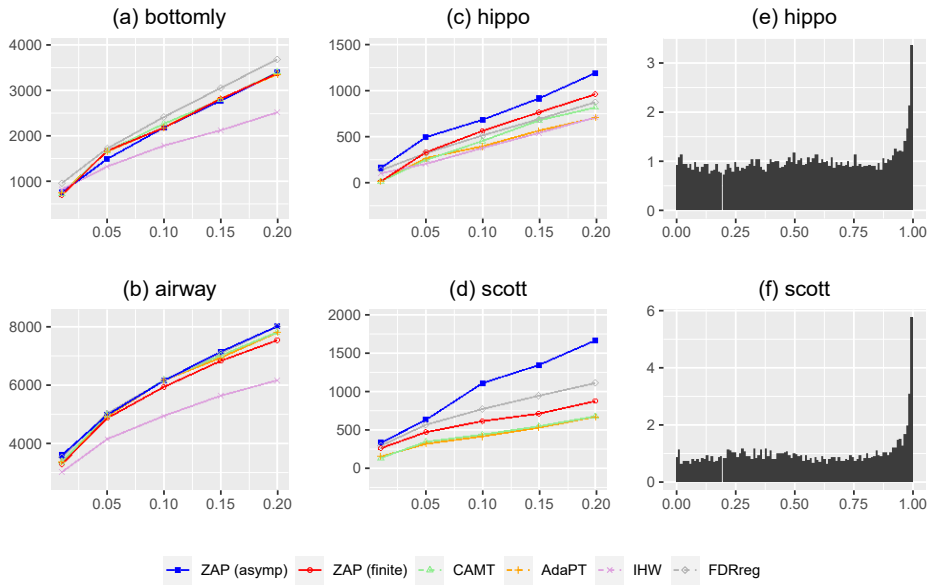

Figure 4.2. paq pdq plot the numbers of rejections for different methods across datasets, against targeted FDR level at 0:01; 0:05; 0:1; 0:15; 0:2. peq and pfq are respectively the histograms of the \u-values" for the hippo and scott datasets.

The number of rejections for the various methods are shown in Figure 4.2paq-pdq, and ZAP has attained top power performance. For the datasets bottomly and airway, we do not see substantial dierences 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 overow 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 4.2peq and pfq, show that the u-value distribution is asymmetric. This explains why ZAP exhibits considerable power improvement over the p-value methods. Specically, the patterns in Figure 4.2 pcq and pdq 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 nite-sample ZAP, whose validity is based on fewer assumptions (Theorem 3.2). In summary, the real data analyses arm that z-value based approaches to covariate-adaptive testing can better exploit the full data $tZ_i; X_i u_{i1}$ to^{m}boost testing power.

5. Discussion

We have introduced ZAP, which is a z-value based covariate-adaptive testing framework that oers FDR controls under minimal assumptions. In particular, our method for nite-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-ts-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 misspecications. For example, normal mixtures are another popular class of models used by researchers for FDR testing in dierent 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 dened 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 exible 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 briey 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 $\mathrm{CLfdr}_x(p, q)$ 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 eciency 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 justication 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 nite-sample FDR control, we rst 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"

$$\underbrace{(0, s_t \cdot \mathrm{pX}(q, s))}_{\text{\left" rejection region}} \quad / \quad \underbrace{(0.5 - s_t \cdot \mathrm{pX}(q, 0.5), 0.5)}_{\text{\left" acceptance region}};$$

and similarly for a \right" u-value. By requiring that $s_{l;t\ 1} \¤ s_{l;t}$ and $s_{rt\ 1}{}' s_{rt}$, 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 nite-sample FDR control. Hence, other working models, which may lead to dierent 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 Candes, E. J. (2015). \Controlling the false discovery rate via knockos." The Annals of Statistics, 43(5): 2055{2085.

| (2019). \A knocko lter 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: dierential 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 proling identies 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 eld 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-dierentially 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 dierential 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 dierential 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 signicance 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 unied 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 misspecied nonlinear regression models." Journal of the American Statistical Association, 76(374): 419{433.

| (1982). \Maximum likelihood estimation of misspecied 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 ^{P} and ^{Z}. In this section we briey review the optimality properties of the oracle procedures ^{P} and ^{Z} in Section 2.2 and how their thresh-olds t_P and t_Z are determined. To streamline the discussion, we will let $tM_iu_{i1}^m$ 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 $tM_i; X_iu_{i1}^m$, 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 $P pH_i = 0|M_i; X_iq$ falls below a data-dependent threshold t_M. We will use ^{M} (in a similar way as ^{P} or ^{Z}) to denote the procedure that thresholds the quantities $tP pH_i = 0|M_i; X_iqu_{i1}^m$ with t_M.

There are subtly dierent ways to dene \optimality", depending on the particu-lar false discovery (e.g. FDR, mFDR, pFDR) and power (e.g. TPR, ETD, mFNR) measures used, which may lead to dierent ways of setting t_M. A most recent result of Heller and Rosset (2021, Theorem 3.1) suggests that among all the testing procedures that are functions of $tM_i; X_iu_{i1}^m$, the t_M that renders an ETD-maximing ^{M} with $FDR \le$ can be found by solving an integer optimization problem (Heller and Rosset, 2021, Theorem 3.1). For our purpose, by letting $L_{p1q} \ge \cdots \ge L_{pmq}$ be the order statistics of the posterior probabilities $tP pH_i = 0|M_i; X_iqu_{i1}^m$, we have considered the computationally simpler optimal procedure rst proposed in Sun and Cai (2007) which takes $t_M = L_{pjq}$, where

$$j = \max\left\{ i^1 \in \{1,\ldots,m\} : \frac{\sum_{i=1}^{i^1} L_{piq}}{i^1} \le \right\}. \tag{A.1}$$

This procedure has $FDR \le$ because for any procedure that produces a rejection set R based on $tM_i; X_iu_{i1}^m$, its FDR can be written as

$$FDR = E\Big[\underbrace{\frac{\sum_{i\in R}(1 - H_i)}{|R| \vee 1}}_{ErFDP\,|\,tM_i;X_iu_{i1}^m s}\Big] = E\Big[\frac{\sum_{i\in R} P pH_i = 0|M_i; X_iq}{|R| \vee 1}\Big].$$

Conditional on any instance of the data $tM_i; X_iu_{i1}^m$, ^{M} prioritizes rejections of the hypotheses that are least likely to be true nulls, all the while controlling the conditional FDR $ErFDP\,|\,tM_i; X_iu_{i1}^m s$ below by setting $t_M = L_{pjq}$, as the ratio $\sum_{i1}^{i^1} L_{piq}\{i^1$ in (A.1) is precisely the conditional FDR of rejecting the most promis-ing ^1i hypotheses. As a result, its controls the FDR under since the conditional version $ErFDP\,|\,tM_i; X_iu_{i1}^m s$ is always not larger than .

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 R is the ratio

$$mFDR = \frac{ErV\,s}{ErRs},$$

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

$$\Lambda_M = \sup\left\{P \in (0,1) : \frac{\sum_i E[(1-H_i)\mathbb{1}(P_i \le P) \mid H_i = 0, M_i, X_i]}{\sum_i E[\mathbb{1}(P_i \le P) \mid H_i = 0, M_i, X_i]} \le \alpha\right\}; \tag{A.2}$$

see Sun and Cai (2007) and Heller and Rosset (2021). In practice, since Λ_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 $t_{\Lambda_M} \approx L_{p|q}$ 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

$$f_{P_i|X_i}(p|x) = g_x(p)q + g_{0|x}(p|x)(1-w_x)q = (1-w_x)q + w_x g_{1;x}(p)q, \tag{A.3}$$

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

$$S^Z(x) = \{z : P(H_i = 0 \mid Z_i = z, X_i = x) \le \alpha\} = \left\{z : \frac{(1-w_x)f_{0|x}(z)q}{(1-w_x)f_{0|x}(z)q + w_x f_{1|x}(z)q} \le \alpha\right\}$$

and

$$S^P(x) = \{p : P(H_i = 0 \mid P_i = p, X_i = x) \le \alpha\} = \left\{p : \frac{(1-w_x)g_{0|x}(p|z)q}{(1-w_x)g_{0|x}(p|z)q + w_x g_{1|x}(p|z)q} \le \alpha\right\}.$$

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

$$\left\{z : \frac{0.2\,\frac{1}{2}\exp\left(-\frac{(z-\mu)^2}{2}\right)}{0.8\exp\left(-\frac{z^2}{2}\right) + 0.2\,\frac{1}{2}\exp\left(-\frac{(z-\mu)^2}{2}\right)} \le \frac{1}{1+\tau}\right\},$$

where $\mu = 1.5$. By setting $\tau = \frac{w_x f_{1;x}(z)}{(1-w_x)f_{0|x}(z)}$, one arrives at the equation

$$\frac{1}{2}\exp\left(-\frac{(z-\mu)^2}{2}\right) - \exp\left(-\frac{z^2}{2}\right) \le 0,$$

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

$$\exp(\mu z) \le \frac{4 + \sqrt{16 + 8\exp(\mu^2)(1-x)x}}{1-x};$$

which in turn implies the two boundary points

$$\frac{1}{1-x}\left(\Phi(z_{\alpha}) - \sqrt{\Phi(z_{\alpha})^2 - \left(1-\Phi(z_{\alpha}+x)+\Phi(z_{\alpha})\right)^2 + 2x\left(1-\Phi(z_{\alpha}+x)+\Phi(z_{\alpha})\right)}\right),$$

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

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

$$P(H_i = 0 \mid P_i = p, X = x) = \frac{(1-w_x)g_0(p)}{g_{w_x}(p)} \Big/ \frac{(1-w_x)g_0(p)}{(1-w_x)g_0(p) + w_x g_{1,x}(p)}.$$

Hence, if w_x and $g_{1,x}$ do not depend on x, it is apparent that S_x^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

$$g_{w_x}(p) = \frac{\phi(z_{p/2} - 1.5) + \phi(z_{p/2} + 1.5)}{2\phi(z_{p/2})} \Big/ \frac{\phi(|z_p| - 1.5) + \phi(|z_p| + 1.5)}{2\phi(|z_p|)}$$

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 $\hat{t} = \hat{t}(c)$, can be written as

$$\begin{aligned} FDP &= \frac{\#\{i : \text{null}, T_i \ge \hat{t}\}}{\#\{i : T_i \ge \hat{t}\}} \\ &= \frac{\#\{i : \text{null}, T_i \ge \hat{t}\}}{\#\{i : S_i \le c_0(\hat{t})\}} \Big/ \underbrace{\frac{1 + \#\{i : \text{null}, S_i \le c_0(\hat{t})\}}{1 + \#\{i : T_i \ge \hat{t}\}}}_{\ge \text{ by the definition of our procedure}} \\ &\le \frac{\#\{i : \text{null}, T_i \ge \hat{t}\}}{1 + \#\{i : \text{null}, S_i \le c_0(\hat{t})\}}. \end{aligned}$$

We only have to show that $E\left[\frac{\#\{i : \text{null}, T_i \ge \hat{t}\}}{1 + \#\{i : \text{null}, S_i \le c_0(\hat{t})\}}\right]$ is bounded by 1 using the stopping time argument from (Barber and Candes, 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

$$S_{0i} = \begin{cases} S_i & \text{when } S_i \ge 0.5 \\ 1 - S_i & \text{when } S_i < 0.5 \end{cases}$$

and $\pi_i = c_0(S_{0i})$. Using the order statistics $T_{\pi(1)} \ge \cdots \ge T_{\pi(m_0)}$ of $T_{\pi(1)}, \ldots, T_{\pi(m_0)}$, we moreover let $B_i = \mathbb{1}(S_{\pi(i)} < 0.5)$ for $i = 1, \ldots, m_0$, where the order of $S_{\pi(i)}$'s here is inherited from the order of the $T_{\pi(i)}$'s, rather than the magnitudes of the S_i's themselves. Let $1 \le J \le m_0$ be the index such that

$$T_{\pi(1)} \ge \cdots \ge T_{\pi(J)} \ge \hat{t} > T_{\pi(J+1)} \ge \cdots \ge T_{\pi(m_0)},$$

we then have

$$\frac{\#\{i \text{ null}: T_i \geq \hat{t}\}}{1 \vee \#\{i \text{ null}: S_i \geq 1 - c_i(\hat{t})\}} \leq \frac{\#\{i \text{ null}: T_i \geq \hat{t}\}}{1 \vee \#\{i \text{ null}: c_i^{-1}(1 - S_i) \geq u\}\big|_{u=1}} \cdot \frac{B_1 + \cdots + B_J}{1 \vee (B_1 + \cdots + B_J)} \leq \frac{1 \vee J}{1 \vee (B_1 + \cdots + B_J)} \leq 1,$$

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

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 $\lambda_i, \tau_i, k_{\lambda i}, k_{\tau i}, h_{\lambda i}, h_{\tau i}$ to respectively denote the quantities and functions $\lambda_{i,X_i}, \tau_{i,X_i}$, $k_{\lambda,X_i}, k_{\tau,X_i}, h_{\lambda,X_i}, h_{\tau,X_i}$ from Model (3.6) when the observed covariate X_i is used, where the underlying parameters $\{\lambda, \tau, \gamma, \kappa\}$ are unspecified but common for all $i = 1, \ldots, m$. Likewise, we also use

$$a_i(u) = a_{X_i}(u) = \frac{1 - \lambda_i + \tau_i}{(1 - \lambda_i - \tau_i) + \lambda_i h_{\lambda i}(u) + \tau_i h_{\tau i}(u)} \tag{C.1}$$

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 $h_{\lambda i}$ and $h_{\tau i}$.

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

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

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

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

Proof of Lemma C.1. It suffices to show the facts for $h_{\lambda i}$ since those for $h_{\tau i}$ can be proven exactly analogously. Recall that

$$h_{\lambda i}(u) = B(k_{\lambda i}, \gamma_\lambda)^{-1} u^{k_{\lambda i} - 1} (1 - u)^{\gamma_\lambda - 1},$$

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

$$h'_{\lambda i}(u) = B(k_{\lambda i}, \gamma_\lambda)^{-1} \left[(k_{\lambda i} - 1) u^{k_{\lambda i} - 2} (1 - u)^{\gamma_\lambda - 1} - (\gamma_\lambda - 1) u^{k_{\lambda i} - 1} (1 - u)^{\gamma_\lambda - 2}\right], \tag{C.2}$$

which, given $\lambda \geq 2$ (actually $\lambda \geq 1$ is sucient), can be seen to be always negative for any $u \in [0,1]$ and hence proves that $h_{k\lambda}$ is strictly decreasing. For convexity, we dierentiate one more time to get

$$h''_{k\lambda}(u) = B(k+\lambda, \lambda)\left[k(k-1)u^{k-2}(1-u^k)^{\lambda-1} - 2k\lambda u^{k-1}(1-u^k)^{\lambda-2}ku^{k-1} + (\lambda-1)\lambda u^{2k-2}(1-u^k)^{\lambda-3}ku^{k-1}\right]$$

$$= B(k+\lambda, \lambda)\,u^{k-2}(1-u^k)^{\lambda-3}\Big[k(k-1)(1-u^k)^2 - 2k\lambda u^k(1-u^k) + (\lambda-1)\lambda u^{2k}\Big]$$

$$= B(k+\lambda, \lambda)\,u^{k-2}(1-u^k)^{\lambda-3}\Big[k(k-1) - 2k(k+\lambda-1)u^k + (k+\lambda-1)(k+\lambda)u^{2k}\Big] \geq 0,$$

where in the last equality, the positive terms are positive since $0 \leq k_{k\lambda} \leq 1$. As such, $h''_{k\lambda}(u)$ is strictly positive for all $u \in [0,1]$ as long as $\lambda \geq 2$, which proves the strict convexity of $h_{k\lambda}$.

A closer inspection of the proof above will reveal that for $\lambda \in [1,2)$, $h_{k\lambda}$ may not even be convex, and the same is true for $h_{r\lambda}$. Hence we have required that $\lambda_l, \lambda_r \geq 2$ in our model. To facilitate the proof in later sections we will also dene the reciprocal assessor function

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

By the properties of $h_{k\lambda}$ and $h_{r\lambda}$ 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 dened in (C.3) (for $\lambda_l, \lambda_r \geq 2$) 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

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

$T_i = a_i(U_i)$ 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 $\{T_i \leq t\}$ is equivalent to U_i taking values in a certain sub-interval of $[0,1]$. One can dene two smooth functions to describe this fact:

Denition C.1 (Expression for the event $\{T_i \leq t\}$). 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 $t \in [0, a_i \underline{u}_i]$,

$$\{T_i \le t\} = \{U_i \le \omega_{iL}(t)\} \cup \{U_i \ge \omega_{iR}(t)\} = \{S_i \le c_i(t)\},$$

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 $\{S_i \le s\}$ similar to Definition C.1:

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

$$\{S_i \le s\} = \{U_i \le \gamma_{iL}(s)\} \cup \{U_i \ge \gamma_{iR}(s)\} = \{T_i \ge c_i^{-1}(s)\},$$

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

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

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

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 $\hat{T}_i^m$ is the exact reflection of $\hat{T}_i$ under the null distribution c_i. In practice, $\hat{T}_i^m$ 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 (t, u) is taken to be the pair (t, u) 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: $\underline{u}_i$, $\omega_{iL}(\cdot)$, $\omega_{iR}(\cdot)$, $\gamma_{iL}(\cdot)$ and $\gamma_{iR}(\cdot)$. 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) $\max_i \|X_i\|_\infty \le C$ almost surely for some universal constant $C > 0$, where $\|\cdot\|_\infty$ indicates the sup norm.

(ii) Let $[\varepsilon_1, \varepsilon_2]$ be any fixed compact interval in $(0, 1)$. For each i, let V_{1i} and V_{2i} be two measurable subsets in $[\varepsilon_1, \varepsilon_2]$. Then for large enough m,

$$\frac{1}{m-1} \sum_{i \ne 1} |P(U_i \in V_{1i} \mid X_i) - P(U_i \in V_{2i} \mid X_i)| \le C(\varepsilon_2 - \varepsilon_1) \max(|V_{i1} \triangle V_{i2}|, 1/m),$$

where $p\cdot q$ is the Lebesgue measure, $C(\epsilon_1, \epsilon_2)$ is a constant that may depends on ϵ_1 and ϵ_2, and $S_1 \triangle S_2$ is their symmetric difference for any two sets $S_1, S_2 \subseteq \mathbb{R}$.

(iii) $\mathrm{Var}[\log(P_i)]$ and $\mathrm{Var}[\log(1 - P_i)]$ are finite.

Assumption 3 (Strong laws of large numbers). Let $\mu_0 = \mathbb{P}(H_i = 0) > 0$. For any $t \in (0, 1]$, it holds that

$$\text{(D.1)} \quad \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}(P_i \le t) / \mu_i(t) \to G(t),$$

$$\frac{1}{m} \sum_{i=1}^{m} (1 - H_i) \mathbb{I}(P_i \le t) / \mu_i(t) \to G_0(t) \text{ and}$$

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{I}(S_i \ge 1 - c(t)) / c(t) \to G_0(t)$$

almost surely, where $G(t)$, $G_0(t)$ and $G_0(t)$ are positive continuous functions in t. Moreover, $G_0(t_0)/G(t_0) \ne \mu_0$ for some $t_0 > 0$, and the limiting threshold

$$t^* = \sup\left\{ t \in (0, 1] : \frac{G_0(t)}{G(t)} \le t \right\}$$

is such that $\max_{i \in \mathcal{N}} c_i(t^*) \le s$ for $s < 1$, or equivalently, $t^* < \min_i c_i^{-1}(s)$. Note that the strong laws above, as well as the marginal probability μ_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 $\sup_{i \in \mathcal{N}} c_i(t)$ 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(ii) 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 $\mathrm{E}[\cdot | H_1, \ldots, H_m]$, 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 2piq, all k_{ri} $k_r pX_i q$, k_{li} $k_l pX_i q$, $_{ri}$ $_r pX_i q$, $_{li}$ $_l pX_i q$ are bounded away from one and zero, i.e. $k_{ri}; k_{li} \in \underline{r}k; ks$, $_{ri}; _{li} \in r; s$ for some compact intervals $\underline{r}k; ks; \underline{r}; s \subset p0; 1q$, by the compactness of B. Moreover, by continuity of the beta functions we can dene

$$B_{max} : \max_{k \in \underline{r}k;ks} pBpk; q \dashv pBp ; kqq \text{ and } B_{min} : \max_{k \in \underline{r}k;ks} pBpk; q \dashv pBp ; kqq;$$

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 2piq, the following are true for any $p; q \in B$:

(i) For any $t_0 > 0$, there exists a $u_0 \in u_0 pt_0 q > 0$ not depending on $p; q$ such that for all $t \geq t_0$, $p!_{iL} ptq; !_{iR} ptqq \subset ru_0; 1 - u_0 s$ for all i.

(ii) For any $s_0 > 0$, there exists a $u_0 \in u_0 ps_0 q > 0$ not depending on $p; q$ such that for all $s \geq s_0$, $p_{iL} psq; {}_{iR} psqq \subset ru_0; 1 - u_0 s$ for all i.

(iii) There exists a small positive constant $u_0 \in p0; 0.5q$ not depending on $p; q$ such that,

$$\underline{u}_i \in ru_0; 1 - u_0 s$$

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

(iv)

$$\lim_{N \to 0} \max_{i \in N; P;} \quad \sup_{P \in B} \quad b_i^{-1} \, rb_i pu_i q; b_i pu_i q \quad = 0, \text{ where}$$

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

$$b_i^{-1} pT q \equiv tu : b_i puq \in T u:$$

for any interval T in R.

Proof. piq: Note that

$$b_i \geq \max \left\{ \frac{u^{k-1} \, p1 - uq^{l-1}}{p1 - \underline{2}qB}, \frac{u q^{k-1} \, p1 - uq^{l}}{\max p1 - \underline{2}qB} \right\} \quad \text{for all } i; \tag{D.2}$$

which implies $\lim_{u \to 0} \min_i b_i puq = \lim_{u \to 1} \min_i b_i puq = 8$, since the right hand side of (D.2) tends to 8 as u tends to 0 or 1. Hence one must be able to nd a small enough $u_0 > 0$ such that $\min pb_i pu_0 q; b_i p1 - u_0 qq > 1\{t_0$ for all i, which implies that $p!_{iL} ptq; !_{iR} ptqq \subset ru_0; 1 - u_0 s$ for all i by Denition C.1. This proves piq.

piiq: Suppose towards a contradiction, such a u_0 doesn't exist. Without loss of generality, we assume there is a subsequence $ti_1; i_2; : : : u$ such that $\lim_{j \to 8} {}_{i_j L} ps_0 q = 0$. As such, $\lim_{j \to 8} {}_{i_j R} ps_0 q = 1 - s_0$ by the property stated in Denition C.2, which implies

$$\lim_j \left(\frac{p1 - s_0 q^{k-1} ps_0 q^{l-1}}{p1 - 2qB_{min}} - \frac{p1 - s_0 q^{l-1} ps_0 q^{k-1}}{p1 - 2qB_{min}} \right)^{i_j} = 0 \tag{D.3}$$

On the other hand,

$$\limsup_j b_{i_j} p \quad ps \quad qq \Big/ \left(1 - \frac{}{} - \frac{}{} \right) :$$

$$\lim_j b_{i_j} p \quad {}_{i_j L} ps_0 qq = 8 \tag{D.4}$$

in consideration of (D.2) and lim$_{j Ñ 8}$ $_i$ $_L$ps$_0$q 0. (D.3) and (D.4) together reach a contradiction since it must be that lim$_j$ b^{j}p $_L$ps$_0$qq$_{i_j}$ lim$_j$ b p $_R$ps$_0$qq as b p$_{i_j}$ $_L$ps$_0$qq b p $_i$ $_R$ps$_0$qqby Denition C.2.

piiiq: By the fact that that b$_i$pb$_i$pu$_i$qq 0 for all i, it suces to show that

(D.5) $$\lim_{u Ñ 0} \max_{i;;} b_i^1 puq \; 8 \quad \text{and} \quad \lim_{u Ñ 1} \min_{i;;} b_i puq^1 \; 8 :$$

Note that

$$b_i^1 puq \quad 1 \frac{i}{ri \; li} h_{li}^1 puq \quad 1 \; ri \; li \quad h_{ri} puq; \frac{ri}{} \quad 1$$

where h^1_{li} has the form

h^1_{li}puq Bpk$_{li}$; $_l$q^{1}rpk$_{li}$ 1qu$^{k_{li}2}$p1 uq^{l1} p$_l$ 1qp1 uq^{l2}u$^{k_{li}1}$s as shown in the proof of Lemma C.1. Dene, for u P p0; 1q, the functions

h^1_lpuq B$_{max}^1$rpk 1qu^{k2}p1 uq^{l1} p$_l$ 1qp1 uq^{l2}u^{k1}s;

$\underline{h}^1_l$puq B$_{min}^1$rpk 1qu^{k2}p1 uq^{l1} p$_l$ 1qp1 uq^{l2}u^{k1}s so that $\underline{h}^1$puq ⁄ h$_{li}$puq ⁄ h$_l$puq 1 1 0 for all i. Note that

$$\lim_{u Ñ 0} h_l^1 puq \; 8 \quad \text{and} \quad \lim_{u Ñ 1} \underline{h}_l puq^1 \; 0:$$

One can similarly dene functions h$_r$ ^{1}and $\underline{h}_r$ ^{1}on p0; 1q such that 0 $\underline{h}^1_r$puq ⁄ h^1_{ri}puq ⁄ h^1_rpuq for all i and

$$\lim_{u Ñ 0} h_r^1 puq \; 0 \text{ and } \lim_{u Ñ 1} \underline{h}_r puq^1 \; 8 :$$

The fact that b^1_ipuq ⁄ $_{12}$ h^{1}puq $_{12}$ h^1 puq, together with two of the limit results above, has shown the rst limit in (D.5). Similarly, that b$_i$puq^1 ¥ $_{12}$ $\underline{h}_l$puq^1 $_{12}$ $\underline{h}_r$puq^1, together with the other two limit results above, has shown the second limit in (D.5).

pivq: By Lemma D.1piiiq, pick u$_0$ P p0; 0:5q such that $\underline{u}_i$ P ru$_0$; 1 u$_0$s for all i. Note that

$$b_i^2 \; p \; u \; q \quad 1 \frac{h_{li}^{li} \; p \; uq}{li \; ri} \; 2 \quad 1 \; li \; ri \quad h_{ri} pu \; q; \frac{ri}{} \quad 2$$

where the dependence on X$_i$ has been suppressed in notations for brevity. By Lemma C.1 , b^2_i is always positive, hence there exists a universal positive number c ¡ 0 such that

$$b_i^2 puq \; ¡ \; c$$

for all i and all u P ru$_0${2; 1 u$_0${2s, considering that , B and ru$_0${2; 1 u$_0${2s are all compact. Now for each i consider the quadratic function

$$f_i puq \quad \frac{c \;\; u \; u_i}{p \quad 2} \; q^2 \; b_i p\underline{u}_i q$$

dened on ru$_0${2; 1 u$_0${2s. Then on the interval ru$_0${2; 1 u$_0${2s , b$_i$ ¥ f$_i$ since g$_i$ b$_i$ f$_i$ is strictly convex with g$_i$p$\underline{u}_i$q g$_i$p$\underline{u}_i$q^1 0. Then

$$b^1_i prm_i; m_i \quad qq ⁄ \quad f^1_i prm_i; m_i \quad qq \qquad \frac{c \; 8}{c}$$

where the right hand side obviously converges to zero as Ñ 0.

We will now state two crucial \event inclusion" lemmas that involve the most delicate proofs in this paper, and may be skipped at rst reading. To state them, we conveniently dene the long vectors $\mathbf{p}(q) \equiv (p_{li}; r_i q_{i1}^m$ and $\mathbf{K} \equiv \mathbf{K}(q) \equiv (k_{li}; k_{ri} q_{i1}^m$ with 2m components. Note that they implicitly depend on the unspec-ied parameters $t; u$. As such, we can also dene $\mathbf{p}(q)$ and $\mathbf{K} \equiv \mathbf{K}(q)$ to be the versions evaluated at and .

Lemma D.2 (First event inclusion lemma). Suppose Assumptions 1 and 2(i)(ii) are true and let $p; q \in B$. For given $\underline{t} > 0$ and > 0, there exists a > 0 such that whenever $\| \|_8 < \| \mathbf{K} - \mathbf{K} \|_8 <$,

$$S_i(c_i(p,q),u) \le S_i(c_i(p,q),u) \le S_i(c_i(p,q),u) \text{ for all } i \text{ and all } t \ge \underline{t}.$$

Proof of Lemma D.2. By Denition C.1, we will show, equivalently, that there exists > 0 such that whenever $\| \|_8 < \| \mathbf{K} - \mathbf{K} \|_8 <$,

(D.6)
$$U_i(\underbrace{p_{li}(t, q_i; \mathbf{l}_{-i}(t, q_{-i}))}_{\text{length-}1 c_i(p,q)}, q_i; \mathbf{l}_{-i}(t, q_{-i})), u) \le U_i(\underbrace{p_{li}(t, q_i; \mathbf{l}_{-i}(t, q_{-i}))}_{\text{length-}1 c_i(p,q)}, q_i; \mathbf{l}_{-i}(t, q_{-i})), u) \le U_i(\underbrace{p_{li}(t, q_i; \mathbf{l}_{-i}(t, q_{-i}))}_{\text{length-}1 c_i(p,q)}, q_i; \mathbf{l}_{-i}(t, q_{-i})), u)$$

where we dene $t_{i;} \equiv c_i^{-1}(c_i(p,q) - q)$ and $t_{i;} \equiv c_i^{-1}(c_i(p,q) - q)$ that are respectively less and greater than t. In particular, we will rst focus on showing the second inclusion in in (D.6) , which amounts to showing

(D.7) $b_i(p, !_{iL}(t_{i;}, q)) \wedge b_i(p, !_{iR}(t_{i;}, q)) \ge 1\{t \text{ whenever}$

$\| \|_8 < \| \mathbf{K} - \mathbf{K} \|_8 <$, in light of the fact that (D.8)

$$b_i(p, !_{iL}(t_{i;}, q)) \le b(p, !_{iR}(t_{i;}, q)) \le 1\{t \ge t_{i;}\},$$

by the denition of $!_{iL}(p,q)$ and $!_{iR}(p,q)$ in Denition C.1 and properties of b_i from Lemma C.2.

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

$$r_{;k} b_i(u) \equiv \left(\frac{\partial}{\partial _i} b(u; _i; k_i) \; ; \; \frac{\partial}{\partial k_i} b(u; _i; k_i) \right)^T$$

be the gradient of b_i with respect to $(_i; k_i)$ evaluated at u. Using the compactness of $[u_0/2; 1 - u_0/2]^s \times B$ and Assumption 2(i)(ii) again, one can nd a universal constant $C(u_0) > 0$ such that the gradient bounds

(D.9) $\|r_{;k} b_i(u)\|_1 \le C(p,q)$ for all i; all $u \in [u_0/2; 1 - u_0/2]^s$; all $p, q \in B$: On the other hand, without loss of generality, we will let

(D.10) $u_0/2$

and, with Lemma D.1pivq, take ~ ¡ 0 be a small enough constant such that

(D.11) $\quad pb_i^{\ 1}prb_i\ pu q; bpu q\ _i\ \ ~qq$ $\quad$ for all i:

By the mean-value theorem and the gradient bound (D.9), one can then nd ¡ 0 such that when } }8 _ }K K }8

(D.12) $\quad |b_i pu q\ \ b_i\ pu q|\quad$ ~ for all u P $\ u_0/2;1\ \ u_0/2\ $:

By the construction of ~ in (D.11) and convexity properties from Lemma C.2, one must have for all i

$$bp!_{i_L}pt_{i;}qq\ \ bp!_{iL}ptqq\ \ bp!_{iR}pt_{i;}qq\ \ bp!_{i_R}ptqq\ ¥\ ~;$$ which implies

(D.13) $\quad 1\{t_{i;}\ ¥\ \ ~\quad 1\{t:$

by the last property in Denition C.1. Since tS$_i$ ¡ c$_i$ ptqu • tS$_i$ ¡ c$_i$ ptqu, from the property (C.5) in Denition C.2 both

$$!_{i_L}pt_{i;}q\ P\ p!_{i_L}ptq\ \ ;!_{iL}ptqq \text{ and } !_{iR}pt_{i;}q\ P\ p!_{iR}ptq;!_{iR}ptq\quad q \text{ are true,}$$

which implies

$$r!_{i_L}pt_{i;}q;\ !_{iR}pt_{i;}qs\ €\ ru_0\{2;1\ \ u_0\{2s;$$

considering (D.10) and $!_{iL}ptq;\ !_{iR}ptq$ P ru$_0$; 1 u$_0$s (as t ¥ t). Therefore by (D.12), we must have

(D.14) $\quad b_i p!_{iL}pt_{i;}qq\ \wedge\ b_i p!_{iR}pt_{i;}qq\ ¥\ \ 1\{t_{i;}\ \ ~$

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

The proof for the rst inclusion in (D.6) follows an analogous argument but is with less resistance, since r!$_{iL}$pt$_{i;}$ q; !$_{iR}$pt$_{i;}$ qs € ru$_0$; 1 u$_0$s for all i. We leave it to the reader.

Lemma D.3 (Second event inclusion lemma). Suppose Assumptions 1 and 2piq are true and let p;q P B. For any xed t $\qquad$ min$_i$ c$_i^{\ 1}$pq with s $\quad$ 1 and any ¡ 0, there exists a ¡ 0 such that whenever } }8 _ }K K }8 $\quad$,

$$tS_i\ \ ¡\ \ 1\ \ c_i\ ptq\ \ u\ •\ tS_i\ \ ¡\ \ 1\ \ c_i ptqu\ •\ tS_i\ \ ¡\ \ 1\ \ c_i\ ptq\quad u; \text{ for all } i$$

and all t ⁄ t.

Proof of Lemma D.3. Note that from (C.5) in Denition C.2 and Lemma D.2 we can conclude there exists a $_1$ ¡ 0 such that whenever } } 8 _ } K K } 8 $\quad$ $_1$,

(D.15) $\quad 1\ \ c_i\ ptq\ \ ⁄\ \ 1\ \ c_i ptq\ ⁄\ \ 1\ \ c_i\ ptq\quad :$

Based on (D.15), it suces to show that there exists a $_2$ ¡ 0 such that whenever } }8 _ }K K }8 $\qquad$ $_2$,

(D.16) $\quad tS_i\ \ ¡\ \ 1\ \ c_i\ ptq\ \ u\ €\ tS_i\ \ ¡\ \ 1\ \ c_i\ ptq\ \ 2u;$

and

(D.17) $$tS_i - 1\, c_i(t) \quad u \bullet tS_i - 1\, c_i(t) \quad 2u$$

which conclude the lemma by taking $_1 \_ _2$ and replacing with {2. In fact, since $t \nearrow t$ and $c_i(t)$ are bounded away from 1, we will show the more general statement: For a given $\underline{s} > 0$, there exists $\underline{\delta}(\underline{s}) > 0$ such that whenever $\|\|_8 \_ \|K\ K\|_8$,

(D.18) $$tS_i - su \in tS_i - s\, u$$

and

(D.19) $$tS_i - su \bullet tS_i - s \quad u$$

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

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

(D.20) $$P_{iL}(s) \cap P_{iR}(s) \in P_{iL}(s) \cap P_{iR}(s)$$

in light of Denition C.2. In particular, it suces to only consider the case where $\underline{s}$, since if $s \nearrow 0$, $tS_i - s\, u \quad tS_i \geq 0u$ becomes the whole underlying probability space which makes (D.20) trivially true. Now for each i, let

$$\sigma(i) : \sup\{s^1 \in [0,1] : P_{iL}(s) \cap P_{iR}(s) \in P_{iL}(s^1) \cap P_{iR}(s^1)\}.$$

By Denition C.2 it must be the case that

(D.21) $$\sigma(i) \quad 1 \quad _{iR}(\sigma(i)) \quad _{iL}(\sigma(i)) \nearrow 1 \quad _{iR}(s) \quad _{iL}(s) \quad s, \text{ and}$$

only one of the following possibilities can be true:

(i) $_{iL}(\sigma(i)) \quad _{iL}(s)$ and $_{iR}(s) \quad _{iR}(\sigma(i))$,
(ii) $_{iL}(\sigma(i)) \quad _{iL}(s)$ and $_{iR}(s) \quad _{iR}(\sigma(i))$,
(iii) $_{iL}(\sigma(i)) \quad _{iL}(s)$ and $_{iR}(s) \quad _{iR}(\sigma(i))$.

In light of the monotone properties in Denition C.2, it suces to show that

(D.22) $$s \quad \sigma(i) \quad ;$$

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 $\sigma(i)$ since the proof for the case of (ii) follows a parallel argument.

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

(D.23) $$P_{iL}(s) \cap P_{iR}(s) \in r(u_0; 1 \quad u_0 s) \text{ for all } i \in N \text{ and } s \geq \underline{s}:$$

Consider each $b_i(u) \quad b(u; _i; k_i)$ as a function in $(u; _i; k_i)$, and let

$$r_{;k} b_i(u) : \quad \frac{\partial}{\partial_i} b(u; _i; k_i) \quad ; \quad \frac{\partial}{\partial k_i} b(u; _i; k_i)$$

be the gradient of b_i with respect to $(_i; k_i)$ evaluated at u. Using the compactness of $r(u_0; 1 \quad u_0 s B)$ and Assumption 2(i) again, one can nd a constant $C(u_0) > 0$ such that the gradient bounds

(D.24) $$\|r_{;k} b_i(u)\|_1 \quad C(u_0) \text{ for all } i \text{ and for all } u \in \left[\frac{u_0}{2}; 1 \quad \frac{u_0}{2}\right] :$$

On the other hand, without loss of generality, with Lemma D.1pivq, let ~ ¡ 0 be a small enough constant such that

(D.25) $$pb_i^{-1}prm_i ; m_i \quad \tilde{}qq \quad \frac{}{2} \text{ for all } i:$$

By the mean-value theorem and the gradient bound (D.24), one can then nd ¡ 0 such that when } }8 _ }K K }8

(D.26) $$|b_ipuq \; b_i \; puq| \quad \tilde{} \text{ for all } u \text{ P } ru_O; 1 \; u_Os:$$

Since $_{iL}$pspiqq $_{iL}$psq P ru$_O$; 1u$_O$s, (D.26) and the last property in Denition C.2 suggest that

(D.27) $$b_i \quad _{iR}pspiqq \quad b \quad _i \quad _{iL}pspiqq \quad b_i \quad _{iL}psq \quad \tilde{} \; b_i \quad _{iR}psq \quad \tilde{}:$$

But since $_{iR}$psq is also in the interval ru$_O$; 1 u$_O$s, we must have

(D.28) $$b_i \quad _{iR}psq \quad ¡ \; b_i \quad _{iR}psq \quad \tilde{}:$$

Combining (D.27) and (D.28), we get that 2~ ¡ b$_i$ $_{iR}$pspiqq b $_i$ $_{iR}$psq which in light of the construction of ~ in (D.25) and convexity properties from Lemma C.2 gives that

$$_{iR}pspiqq \quad _{iR}psq \quad ;$$

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

$$p \; _{iL}ps \quad q; \; _{iR}ps \quad qq \text{ € } p \; _{iL}psq; \; _{iR}psqq:$$

We will alternatively dene

$$spiq : infts_1 \text{ P } p0; 1q : p \; _{iL}ps^{-1}q; \; _i^R ps^{-1}qq \text{ € } p \; _{iL}psq; \; _{iR}psqqu:$$

then show spiq s . 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 ¡ 0 and positive numbers 0 t̲ t min$_i$ c$_i^{-1}$pq, there exists pq ¡ 0 such that, for suciently large m,

(D.29) $$\sup_{\substack{\max p\}KK;\}\}q \\ t̲/t/t}} \quad \frac{1}{m}, ^m \; IpT_i / \; tq \; Gptq /$$

(D.30) $$\sup_{\substack{\max p\}KK;\}\}q \quad t̲/t/t}} \quad \frac{1}{m}, ^m p1 \; H_iqIpT_i / \; tq \; _OG_Optq / \; ;$$

(D.31) $$\sup_{\substack{\max p\}KK;\}\}q \\ t̲/t/t}} \quad \frac{1}{m}, ^m \; IpS_i ¥ \; 1 \; c_iptqq \; G_Optq /$$

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

$$G_i(t) = \frac{1}{m}\sum_{i=1}^{m} 1(T_i \le t),$$

where the subscript emphasizes that the T_i's are defined with an unspecified (p,q), to distinguish from G in Assumption 3. Let n be large enough such that $1/n < \underline{t}/2$ and consider $G^{-}(1/n), G^{-}(2/n), \ldots, G^{-}(n/n)$. If we define

$$n_1 := \min\{i : G^{-}(i/n) > \underline{t}\},$$

$$d_1 := |\{i : G^{-}(i/n) > \underline{t}\}|.$$

Define $t_1 := G^{-}(n_1/n), t_2 := G^{-}(n_1+1/n), \ldots, t_{d_1} := G^{-}(n_1+d_1-1/n)$, as well as $t_0 := \underline{t}$ and $t_{d+1} := t$ with $d := d_1 - 1$. Following the proof of the Glivenko-Cantelli theorem in Resnick (2019, p.224), we have

(D.32)

$$\sup_{t \le t} |G_i(t) - G(t)| \le \max_{v=0}^{d} |G_i(t_v) - G(t_v)| \vee |G_i(t_v) - G(t_v)| + 1/n \le \underline{t}/2 + t.$$

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

(D.33) $$|G_i(t_v) - G(t_v)| = m^{-1}\sum_{i=1}^{m} (1(T_i \le t_v) - 1(T \le t_v)) - R_v,$$

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

(D.34) $$\frac{1}{m}\sum_{i=1}^{m} (1(T_i \le t_v) - 1(T_i \le t_v)) \le \underbrace{m^{-1}\sum_{i=1}^{m} (1(S_i \le c_i(t_v) - \epsilon) - 1(S_i \le c_i(t_v)))}_{(A)} \vee \underbrace{m^{-1}\sum_{i=1}^{m} (1(S_i \le c_i(t_v) + \epsilon) - 1(S_i \le c_i(t_v)))}_{(B)}.$$

for $\max\{\epsilon, \epsilon^2\} < \underline{t}/2$. This is because $\sum_{i=1}^{m} 1(S_i \le c_i(t_v) - \epsilon) - 1(T_i \le t_v) \le \sum_{i=1}^{m} 1(T_i \le t_v) - 1(T \le t_v)$ if the latter term is greater than 0; likewise, $\sum_{i=1}^{m} 1(S_i \le c_i(t_v) + \epsilon) - 1(T_i \le t_v) \ge \sum_{i=1}^{m} 1(T_i \le t_v) - 1(T \le t_v)$ if the latter is less than 0.

We will rst develop a bound for term pAq. One have

$$pAq \;/\; E\Big[\frac{1}{m}\sum_{i1}^{m} rP\,pS_i \;i\; cpt_vq|X_iq\;P\,pS\;i\;cpt_vq\;_i\;|X_iqs\Big] \quad Q_v$$

$$E\Big[\frac{1}{m}\sum_{i1}^{m} P\;U_i\;P\;p\;\;pcpt_vqq;\;\;pcpt_vqq|X_i\Big]$$

$$P\,pU_i\;P\;p\;_{iL}pcpt_vq\;\;q;\;_{iR}pcpt_vq\;\;q|X_iq\;\;Q\;\;_v$$

(D.35) $\quad / \; C\underline{pt}q \quad Q_v$

where $Q_v \;\; o_{a:s:}p1q$ is a remainder term coming from the strong law of Gpq in Assumption 3. The second equality comes from Denition C.2. Note that the intervals

$$p\;_{iL}pcpt_vqq;\;\;_{iR}pcpt_vqq; i\;1;:::;m$$

can be equivalently represented as

$$p!_{iL}pt_vq;!_{iR}pt_vqq; i\;1;:::;m$$

by Denition C.1, which all belong to a compact sub-interval in p0;1q by Lemma D.1piq and the fact that $\underline{t}\;i\;0$. As such, Assumption 2piiq can be applied to give the last inequality (D.35). By realizing, from Denitions C.1 and C.2, that the event

$$tS_i \;i\; c_i\,pt_vq\;u$$

is equivalent to U_i belonging to an interval that is $p!_{iL}pt_vq;!_{iR}pt_vqq$ expanded by a further width, it is obvious that one can analogously develop the bound

(D.36) $\qquad pBq \;/\; C\underline{pt}q \quad Q_v$

for a constant $C\underline{pt}q$ and $Q_v \;\; o_{a:s:}p1q$. Combining (D.33), (D.35) and (D.36) gives (by appropriately adjusting)

(D.37) $\qquad |G_;pt_vq\;\;Gpt_vq| \;/\; \{2$ a.s.; for

suciently large m. A similar bound

(D.38) $\qquad |G_;pt_vq\;\;Gpt_vq| \;/\; \{2$ a.s.;

can be derived in much the same way with no diculty by rst writing

$$G_;pt_vq\;\;Gpt_vq\;\;m^{1}\Big[\;IpT_i\;\;\sum_{i1}^{m}\;\;t_vq\;\;Gpt_vq$$

using the continuity of Gpq, 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 rst writing

$$\sum_{i1}^{m} p1\;H_iqIpT_i \;/\; tq \qquad \sum_{H_i0} IpT_i \;/\; tq;$$

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

$$\frac{1}{m}\sum_i \mathbb{I}\{S_i \le 1 - c_i(t_v)q\} - P\{S_i \le 1 - c_i(t_v)q\} \le \frac{1}{m}\sum_i \mathbb{I}\{S_i \le 1 - c_i(t_v)q + \epsilon\} - P\{S_i \le 1 - c_i(t_v)q + \epsilon\}$$

$$+ \frac{1}{m}\sum_i \mathbb{I}\{S_i \le 1 - c_i(t_v)q + \epsilon\} - P\{S_i \le 1 - c_i(t_v)q + \epsilon\} - \left(\frac{1}{m}\sum_i \mathbb{I}\{S_i \le 1 - c_i(t_v)q\} - P\{S_i \le 1 - c_i(t_v)q\}\right)$$

with Lemma D.3 for t_v that are now quantiles of G_0, for sufficiently small $\epsilon > 0$ and $\max\{\cdot\}$. From Definition C.2, the events $\{S_i \le 1 - c_i(t_v)q\}$ have the form

$$\{U_i \in [\ell_{iL}(1 - c_i(t_v)q), \ell_{iR}(1 - c_i(t_v)q)]\},$$

and to show that the intervals $[\ell_{iL}(1 - c_i(t_v)q), \ell_{iR}(1 - c_i(t_v)q)]$ can be placed in a compact sub-interval of $(0,1)$ by Lemma D.1(iii) to apply Assumption 2(iii), one need to show that $c_1(t_v)q, \ldots, c_m(t_v)q$ are bounded away from 1. This is true because $c_i(t_v)q = c(t_i)q \le 1$ for all i by definition. The same proof rundown goes through, again, by realizing that $[\ell_{iL}(1 - c_i(t_v)q), \ell_{iR}(1 - c_i(t_v)q)]$ are just ϵ-expansion of $[\ell_{iL}(1 - c_i(t_v)q), \ell_{iR}(1 - c_i(t_v)q)]$ 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 $0 \le \underline{t} < t < \min_i c_i^{-1}(q)$, we have

$$\sup_{t \in [\underline{t}, t]} \left|\frac{1}{m}\sum_{i=1}^m \mathbb{I}\{T_i \le t\} - G(t)\right| \to 0, \quad \text{(D.39)}$$

$$\sup_{t \in [\underline{t}, t]} \left|\frac{1}{m}\sum_{i=1}^m (1 - H_i)\mathbb{I}\{T_i \le t\} - G_0(t)\right| \to 0, \quad \text{(D.40)}$$

$$\sup_{t \in [\underline{t}, t]} \left|\frac{1}{m}\sum_{i=1}^m \mathbb{I}\{S_i \le 1 - c_i(t)q\} - G_0(t)\right| \to 0, \quad \text{(D.41)}$$

almost surely.

Proof of Lemma D.5. Let $T_1, T_2, \ldots, T_m$, and define $\rho(q)$ and $K(\rho)$. We will first show that

$$\|\cdot\|_\infty - \|K - K\|_\infty \to 0 \text{ a.s.; which is} \quad \text{(D.42)}$$

a consequence of

$$\cdot \to \cdot \text{ and } \cdot \to \cdot \text{ a.s.:} \quad \text{(D.43)}$$

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

$$\text{(D.44)} \qquad |\log(1 - \pi_{li} - \pi_{ri}) + \pi_{li} h_{li}(U_i) + \pi_{ri} h_{ri}(U_i))| \leq |\log h_{X_i}(U_i)|$$

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

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

for all (π, θ, X_i) whenever $U_i \in [u_0, 1-u_0]$. Note that Lemma C.2 also implies that $h_{X_i}(U_i) \geq c$ for $U_i \in (0, u_0) \cup (1-u_0, 1)$.

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

$$\log h_{X_i}(U_i) \leq \log\left(1 + C + \frac{(1 - U_i)^{\gamma-1} U_i^{k-1}}{B_{\min}}\right)$$

$$\leq \log(2 + C) + \log\left(\tilde{C} \frac{(1 - U_i)^{\gamma-1} U_i^{k-1}}{(1-u_0)^{\gamma-1} u_0^{k-1}}\right) =: m_l(U_i);$$

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

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

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

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

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 $D_m = D_m(\omega) = \sup_{\underline{t} \leq t \leq \bar{t}} \left|\frac{1}{m} \sum_{i=1}^m I(\hat{T}_i \leq t) - G(t)\right|$, where ω denotes a point in the underlying probability space Ω. It suffices to show that for any $\epsilon > 0$, there exists a subspace $\Omega(\epsilon) \subset \Omega$ such that $P(\Omega(\epsilon)) = 1$ and $D_m(\omega) \leq \epsilon$ for sufficiently large m and every $\omega \in \Omega(\epsilon)$. By Lemma D.4, there exists Ω_1 with $P(\Omega_1) = 1$ such that (D.29) holds on Ω_1 for $\psi(\epsilon) > 0$. By (D.42), there exists Ω_2 with $P(\Omega_2) = 1$ such that $\|\hat{\zeta} - \zeta\|_\infty \vee \|\hat{K} - K\|_\infty \leq \psi(\epsilon)$ on Ω_2 for sufficiently large m. Take $\Omega(\epsilon) = \Omega_1 \cap \Omega_2$.

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 $\bar{t} \in (\underline{t}, \min_i c_i^{-1}(\epsilon))$,

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

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

(D.46) $$\limsup_{m\to\infty}\sup_{t\le t/t} FDP(t) \le \frac{G_0(m_0 t)}{\max\{\sum_{i=1}^m I(T_i \le t), 1\}}$$

$$\le \limsup_{m\to\infty}\sup_{t\le t/t} \frac{\sum_{i=1}^m (1-H_i) I(T_i \le t)}{\max\{\sum_{i=1}^m I(T_i \le t), 1\}} \Big/ \frac{G_0(t)}{G_0(t)}$$

$$\le \limsup_{m\to\infty} \frac{\max\{\sum_{i=1}^m I(T_i \le t), 1\}}{m} \Big/ \sup_{t\le t/t} \frac{G_0(t)}{G_0(t)} \le \alpha \quad \text{a.s.}$$

since $\limsup_{m\to\infty} \frac{1}{m}\max\{\sum_{i=1}^m I(T_i \le t), 1\} \ge G(t)$ almost surely, given that $G(t) > 0$.

On the other hand, it must be that $G_0(t) \ge \alpha G_0(t)$ considering that $P(S_i \ge 1 - c(t) \mid H_i = 0) \ge P(T_i \le t \mid H_i = 0)$, which, together with the last Glivenko-Cantelli statement in Lemma D.5, gives

$$\liminf_{m\to\infty} \inf_{t\le t/t} \left[\frac{\sum_{i=1}^m I(S_i \ge 1 - c(t))}{m} - G(t) \right] \frac{G_0(t)}{G_0(t)} \ge 0$$

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

$$\liminf_{m\to\infty} \inf_{t\le t/t} \left(FDP_{asymp}(t) - FDP(t) \right)$$

$$\ge \liminf_{m\to\infty} \inf_{t\le t/t} \left(\frac{\sum_{i=1}^m I(S_i \ge 1 - c(t))}{\max\{\sum_{i=1}^m I(T_i \le t), 1\}} - FDP(t) \right) \ge 0,$$

which is (D.45).

Towards nishing, we will establish that, almost surely,

(D.47) $$\liminf_m \hat{t}_{asymp}(\alpha) > 0 \quad \text{and} \quad \limsup_m \hat{t}_{asymp}(\alpha) \le \min_i c_i^{-1}(\alpha).$$

Fix $t_1 \in (t^*, \min_i c_i^{-1}(\alpha))$. By the denition of t^* in Assumption 3 it must be the case that

$$\frac{G_0(t_1)}{G(t_1)} > \frac{G_0(t^*)}{G(t^*)}$$

and we can let $\frac{G_0(t_1)}{G(t_1)} - 1 > 0$. For suciently large m, because of Lemma D.5 we can get that $\frac{\sum_i I(S_i \ge 1 - c(t_1))}{\max\{\sum_i I(T_i \le t_1), 1\}} \le \frac{G_0(t_1)}{G(t_1)}$ a.s., which implies that $FDP_{asymp}(t_1) > \alpha$ almost surely to give the "limsup" statement in (D.47). On the other hand, let $\frac{G_0(t_0)}{G(t_0)} - 1 > 0$ for t_0 in Assumption 3. Since $t_0 < t^*$ (by continuity of the "G" functions), Lemma D.5 also suggests that for large enough m, $\left| \frac{G_0(t_0)}{G(t_0)} FDP_{asymp}(t_0) \right| < \alpha$ almost surely, which implies $FDP_{asymp}(t_0) < \alpha$ almost surely and hence the "liminf" statement in (D.47).

Now given (D.47) is true, since $\hat{F}DP_{asymp}pt_{asymp}^{\wedge}pqq \nearrow$, by (D.45) it must be true that

$$\limsup_{m} F\,DP\,pt_{asymp}^{\wedge}pqq \nearrow \quad \text{a.s.}$$

By the reverse Fatou's lemma, this implies

$$\limsup_{m} ErF\,DP\,pt_{asymp}pqqs \nearrow E \quad \limsup_{m} F\,DP\,pt_{asymp}pqq \quad \nearrow ;$$

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 dene the notation required to apply their argument. First, for each $t \quad 0;1;:::$, let $V_t \quad \#ti : U_i \; P \; R_t$ and $H_i \quad 0u$ and $U_t \quad \#ti : U_i \; P \; A_t$ and $H_i \quad 0u$ which are respectively the numbers of true nulls in the rejection set and acceptance set at step t. Dene

$$m_i \quad IpU_i \yen 0{:}5qpU_iq \_ U_iq \quad IpU_i \quad 0{:}5qpU_iq \wedge U_iq$$

and

$$b_i \quad Ip0{:}25 \nearrow U_i \nearrow 0{:}75q$$

so that

$$U_i \quad b_iIpm_i \yen 0{:}5qp1{:}5 \; m_iq \quad Ipm_i \quad 0{:}5qp0{:}5 \; m_iqu \quad p1 \; b_iqm_i{:}$$

Also dene $C_t \quad ti : i \; P \; A_t \; Y \; R_t$ and $H_i \quad 0u$ to give

$$U_t \quad \sum_{iPC_t} b_i \text{ and } V_t \quad p1 \; b_iq \; |C_t| \; U_t{:}_{iPC_t}$$

If we set the initial sigma-algebra $G_1 \quad tpX_i; m_iq_{iPt1;:::;mu}; pb_iq_{i:H_i0u}$, then $P\,pb_i \; 1|G_1q \quad 0{:}5$ almost surely for a null i under the uniform null distribu-tion 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 simplied notation in Appendix C. The complete data log-likelihood for Model (3.6), treated as a function of t; u, has the form

$$\begin{aligned} lp;q \quad & \sum_{i1}^{m} H_{li}\, rpk_{li} \; 1q \log pU_iq \quad p_l \; 1q \log p1 \; U_iq \quad \log B pk_{li}; {}_lqs \\ & \sum_{i1}^{m} H_{ri}\, rpk_{ri} \; 1q \; og \; 1 \; U_iq \quad p_r \; 1q \log pU_iq \quad og \, Bp_r; k_{ri}qs \\ & \sum_{i1}^{m} p1 \; H_{li} \; H_{ri}q \log p1 \; {}_{li} \; {}_{ri}q \quad H_{li} \; logp_{li}q \quad H_{ri} \; logp_{ri}q \quad ; \end{aligned} \tag{F.1}$$

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 $p^{(0)}, \pi^{(0)}$ **while** $p^{(j)}, \pi^{(j)}$ not converged **do**

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

$$Q(p, \pi \mid p^{(j)}, \pi^{(j)}) = \sum_{i=1}^{m} \Big(w_{li}^{(j)} \log\big(\pi_{li} h_{li}(U_i)\big) + w_{ri}^{(j)} \log\big(\pi_{ri} h_{ri}(U_i)\big) + \big(1 - w_{li}^{(j)} - w_{ri}^{(j)}\big) \log\big(p_i(1-\pi_{li}-\pi_{ri})\big) \Big),$$

where

$$w_{li}^{(j)} = E_{p^{(j)}, \pi^{(j)}}[H_{li} \mid X_i, U_i] = \frac{\pi_{li}^{(j)} h_{li}^{(j)}(U_i)}{h^{(j)}(U_i, X_i)},$$

$$w_{ri}^{(j)} = E_{p^{(j)}, \pi^{(j)}}[H_{ri} \mid X_i, U_i] = \frac{\pi_{ri}^{(j)} h_{ri}^{(j)}(U_i)}{h^{(j)}(U_i, X_i)}.$$

M step: Compute $p^{(j+1)}, \pi^{(j+1)} = \arg\max_{p,\pi} Q(p, \pi \mid p^{(j)}, \pi^{(j)})$. **end**

Output: Estimated coefficients and

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 $(U_{t,i}, X_i)_{i \in A_t \cup R_t}$ (Appendix F.3) to estimate the parameters. Next, for each masked $i \in A_t \cup R_t$, evaluated assessor value $T_i^{(t)}$ at whichever U_i or $1-U_i$ 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 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 (iii) in Theorem 3.2, to give a different $s_{l,t+1}$ or $s_{r,t+1}$: If $U_j \leq 0.5$, $s_{r,t+1}$ will be updated from $s_{r,t}$ at the point X_j as $s_{r,t+1}(X_j) = U_j - \Delta U_j$, 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 $U_j \wedge (1 - U_j)$. As such, at the next step $t+1$, one of A_{t+1} or R_{t+1} will be shrunk

by exactly one element which is j. This is intuitive since if $\hat{F}^{DP}_{finite}(t) = i$ at step t, one would hope to reduce the size of 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 $(\hat{l}_{t,r}, \hat{r}_{t,l}, \hat{u})$ using the EM algorithm in Appendix F.3.

Construct $\{\hat{F}_{X_i}(u) : i \in A_t \cup R_t\}$ with (3.7) by setting the underlying parameters as the estimates in the prior step.

Find $j = \arg\max_{i \in A_t \cup R_t} |T_i|$ for $T = \hat{F}_{X_i}(U)$, where

$$U_i^{-1} = \begin{cases} U_i & \text{if } U_i < 0.5 \text{ or } U_i = \tilde{U}_i \\ 1 - U_i & \text{if } U_i \geq 0.5 \text{ or } U_i = \tilde{U}_i \end{cases}$$

if $U_j < 0.5$ then

  $s_{rt} = (1 - \hat{F}_{X_i}(\tilde{U}_i)) \cdot \mathbb{1}(T_i > 0) + s_{r,t-1} \cdot \mathbb{1}(T_i \leq 0)$; else

  $s_{lt} = \hat{F}_{X_i}(\tilde{U}_i) \cdot \mathbb{1}(T_i > 0) + s_{l,t-1} \cdot \mathbb{1}(T_i \leq 0)$; end

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

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 $D_{ti} = (X_i, U_{t,i})$ be the available data for i at step t of the finite-sample ZAP algorithm. To update from the parameters $(\theta^{(j)}, \eta^{(j)})$ at the j-th EM iteration, we need to compute the following quantities:

$$\begin{aligned} &E_{(\theta^{(j)},\eta^{(j)})}[H_{li} \mid D_{ti}], \quad E_{(\theta^{(j)},\eta^{(j)})}[H_{ri} \mid D_{ti}], \\ &E_{(\theta^{(j)},\eta^{(j)})}[H_{ri} \log(U_i) \mid D_{ti}], \quad E_{(\theta^{(j)},\eta^{(j)})}[H_{ri} \log(1 - U_i) \mid D_{ti}], \quad E_{(\theta^{(j)},\eta^{(j)})}[H_{li}] \\ &\log(U_i) \mid D_{ti}], \quad E_{(\theta^{(j)},\eta^{(j)})}[H_{li} \log(1 - U_i) \mid D_{ti}]. \end{aligned} \tag{F.2}$$

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 $l_i^{(j)}, \theta_i^{(j)}, h_{ri}^{(j)}, h_{li}^{(j)}$ and $h_{ri}^{(j)}$ be as defined in Section 3.2 evaluated at $(\theta^{(j)}, \eta^{(j)})$. We will have

$$\hat{H}_{li}^{(j)} = E_{(\theta^{(j)},\eta^{(j)})}[H_{li} \mid D_{ti}] = P_{(\theta^{(j)},\eta^{(j)})}(H_{li} = 1 \mid D_{ti}) = \frac{\theta_i^{(j)} h_{ri}^{(j)}(U_i)}{(1 - \theta_i^{(j)}) h_{li}^{(j)}(U_i) + \theta_i^{(j)} h_{ri}^{(j)}(U_i)} \cdot \mathbb{1}(X_i \leq l_i)$$

$$\hat{H}_{ri}^{(j)} = E_{(\theta^{(j)},\eta^{(j)})}[H_{ri} \mid D_{ti}] = P_{(\theta^{(j)},\eta^{(j)})}(H_{ri} = 1 \mid D_{ti}) = \frac{\theta_i^{(j)} h_{ri}^{(j)}(U_i)}{(1 - \theta_i^{(j)}) h_{X_i}^{(j)}(\tilde{U}_i) + \theta_i^{(j)} h_{X_i}^{(j)}(U_i)} \cdot \mathbb{1}(X_i \geq r_i)$$

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

$$y^{pjq}_{li;A} \quad E_{pjq;pjq}\, r_{li} \log p U_i q \,|\, D_t i; H_{li} \quad 1 s \quad \frac{h^{pjq} p U_i q \log p U_i q \quad h^{pjq} p U_i q \log p U_i q}{h^{pjq} p U_{li} q \quad h^{pjq}_{i} p \Theta' q}$$

$$y^{pjq}_{ri;A} \quad E_{pjq;pjq}\, r H_{ri} \log p U_i q \,|\, D_t i; H_{ri} \quad 1 s \quad \frac{h^{pjq} p U_i q \log p U_i q \quad h^{pjq}_{ri} p U_i q \log p U_i q}{h^{pjq}_{li} p U_i q \quad h_{li} p \Theta_i q}$$

$$y^{pjq}_{li;B} \quad E_{pjq;pjq}\, r_{li} \log p 1 \quad U_i q \,|\, D_t i; H_{li} \quad 1 s \quad \frac{h^{pjq} p U_i q \log p 1 \quad U_i q \quad h^{pjq} p U_i q \log p 1 \quad U_i q}{h^{pjq} p U_{li} q \quad h^{pjq}_{i} p \Theta' q}$$

$$y^{pjq}_{ri;B} \quad E_{pjq;pjq}\, r H_{ri} \log p 1 \quad U_i q \,|\, D_t i; H_{ri} \quad 1 s \quad \frac{h^{pjq} p U_i q \log p 1 \quad U_i q \quad h^{pjq}_{ri} p U_i q \log p 1 \quad U_i q}{h^{pjq}_{ri} p U_i q \quad h_{ri} p \Theta_i q}$$

then one can express

$$E_{p^{pjq};pjq q} r H_{li} \log p U_i q \,|\, D_t i s \; y_{li;A} H^{pjq q}; \hat{} _{li} \qquad E_{p^{pjq};pjq q} r H_{ri} \log p U_i q \,|\, D_t i s \; y_{ri;A} H^{pjq q}; \hat{} _{ri}$$

$$E_{p^{pjq};pjq q} r H_{li} \log p 1 U_i q \,|\, D_t i s \; y_{li;B} H^{pjq}_{li}; \hat{} \; pj \qquad E_{p^{pjq};pjq q} r H_{ri} \log p 1 U_i q \,|\, D_t i s \; y^{pjq} \; H^{pjq}_{ri;B} \hat{} _{ri}$$

Initialization. We now discuss how to specify values for $^{pOq}_{li}$, $^{pOq}_{ri}$, ^{pOq} and ^{pOq} to initialize the algorithm. Specifying ^{pOq} and ^{pOq} is easy: For , we only consider the left group $L \quad ti : U_i q \_ U_i \diagup 0{:}5u$, and t the left-leaning beta density $h_{li}pq$ to the data points $tU_i \hat{} U_i : q P Lu$ to obtain an estimate for as the initial value ^{pOq}, with a given value for (such as 4). Note that we t the model to the smaller point $U_i \hat{} U_i q$ instead of $U_i \_ U_i q$ for each $i P L$ with the goal of having a more \aggressive" left alternative distribution. The initial value $^{pOq}_{r}$ can be obtained similarly by considering the right group $R \quad ti : U_i q \hat{} U_i \; 0{:}5u$, the data $tU_i \_ U_i q \; i P Ru$ and $h_{ri}pq$.

To speciy $^{pOq}_{li}$, ^{pOq}, for each i, we rst dene

(F.3)

$$_{ri} \quad P\, pH_{ri} \quad 1 | X_i; U_i \; 0{:}5q; \qquad _{li} \quad P\, pH_{li} \quad 1 | X_i; U_i \; 0{:}5q; \qquad _{i} \quad P\, pU_i \; 0{:}5 | X_i q; _{li}$$

$$P\, pH_{li} \quad 1 | X_i; U_i \diagup 0{:}5q; \quad _{ri} \quad P\, pH_{ri} \quad 1 | X_i; U_i \diagup 0{:}5q; \quad _{i} \quad P\, pU_i \diagup 0{:}5 | X_i q$$

and note that, by denition, $_{ri} \yen$ and $_{li} \yen$. We will form estimates for $\hat{}_{ri}; \hat{}_{i}; \hat{}_{li}$ $; \hat{}_{i}$ and let $_{ri}^{q} \quad \hat{}_{ri} \hat{}_{i}$ and $^{pOq} \quad \hat{}_{li} \hat{}_{i}$ be conservative estimates for $_{ri}$ and $_{li}$. The estimates $\hat{}_{i}$ and $\hat{}_{i}$ can be obtained as predicted probabilities by tting a logistic regression on the indicator responses $D_i \quad IpU_i \; 0{:}5q$ with covariates X_i. For the rest of this section we will focus on the estimates $\hat{}_{ri}$, since the estimates $\hat{}_{li}$ can be obtain analogously.

Let $J_i \quad IpU_{\sigma,i}$ has one elementq. Then

$$E r J_i | X_i; D_i \quad 1 s \; P\, pJ_i \quad 1 | X_i; D_i \quad 1 q \yen p 1 \quad _{ri} \quad _{li} q \qquad \frac{0{:}5 \quad 2 \quad 1 \; s_{r;0} p X_i}{p \quad 0{:}5 \quad q q}$$

which is equivalent to

$$_{li} \yen E \; J_i X_i; D_i \tilde{} \; 1 \; ; \qquad _{ri}$$

where $\tilde{J}_i = 1_{\{0.5 \le 2(1-s_{r,0}(X_i))q\}}$ and

$$E(J_i \mid X_i, D_i = 1) = \underbrace{P(\tilde{U}_{0,i} \text{ has two elements} \mid X_i, D_i = 1)}_{1_{\{J_i=1\}}} + \underbrace{P(\tilde{U}_{0,i} \text{ has one element} \mid X_i, D_i = 1)}_{1_{\{J_i=1\}}} \frac{0.5}{0.5 \vee (1-s_{r,0}(X_i))q}$$

As the probability of $H_{li} = 1$ should be small under $U_i \le 0.5$, 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.

$$\pi_{ri} \le (1-q) 1_{\{J_i=1\}} + 1_{\{0.5 \le 2(1-s_{r,0}(X_i))q\}} \frac{0.5}{(1-s_{r,0}(X_i))q}$$

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

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

$$(1-w)q f_0(z) + w(1-q) f_{-\mu}(z) + w q f_{\mu_r}(z) \qquad (G.1)$$

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

$$\mu_l = -2.5, -2, -1.5, -1, -0.5; \quad \mu_r = 0.5, 1, 1.5, 2, 2.5;$$
$$w = 0.1, 0.15, 0.2; \quad = 0.5, 0.7, 0.9.$$

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 parametrizes the degree of asymmetry reected in the mixing probabilities w_l and w_r. For each set of 8000 z-values, the beta mixture model (3.6) for $(p_l, p_r) = (4, 4)$ is tted with regression intercepts only by an EM algorithm, and the resulting left and right model-based non-null probabilities $\hat{}_l$ and $\hat{}_r$ 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 $(\hat{}_l, \hat{}_r)$ 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 tting

Table 2. Estimated probabilities $\hat{\pi}_l$ and $\hat{\pi}_r$ based on the beta mixture (3.6) with $p_{l;r}q$ $p4;4q$ and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with 0:5. $\hat{\pi}_l$ and $\hat{\pi}_r$ are respectively the left and right entries in each cell.

$_l z_r$	0:5		1		1:5		2		2:5	
	w 0:1, $pw_l; w_r q$ p0:05; 0:05q									
-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, $pw_l; w_r q$ p0:075; 0:075q									
-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, $pw_l; w_r q$ p0:1; 0:1q									
-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 ^$_l$ and ^$_r$ based on the beta mixture (3.6) with p$_l$; $_r$q p4; 4q and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with 0:7. ^$_l$ and ^$_r$ are respectively the left and right entries in each cell.

$_l$z$_r$	0:5		1		1:5		2		2:5	
	w 0:1, pw$_l$; w$_r$q p0:03; 0:07q									
-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, pw$_l$; w$_r$q p0:045; 0:105q									
-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, pw$_l$; w$_r$q p0:06; 0:14q									
-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 ^$_l$ and ^$_r$ based on the beta mixture (3.6) with $p_l; _r q$ p4; 4q and regression intercepts only, for 8000 z-values generated by the normal mixture (G.1) with 0:9. ^$_l$ and ^$_r$ are respectively the left and right entries in each cell.

$_l z_r$	0:5		1		1:5		2		2:5	
	w 0:1, $pw_l; w_r q$ p0:01; 0:09q									
-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, $pw_l; w_r q$ p0:015; 0:135q									
-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, $pw_l; w_r q$ p0:02; 0:18q									
-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 ^{Z}
(c) SABHA: Structure adaptive BH procedure with 0:5, 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_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. Dierential expression analysis of RNA-Seq data. We shall rst briey discuss the importance of pre-ltering genes with excessively low read counts before applying FDR methodologies for dierential 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 dier by their underlying modelling assumptions, to produce test statistics that are re-scaled measures of dierential 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

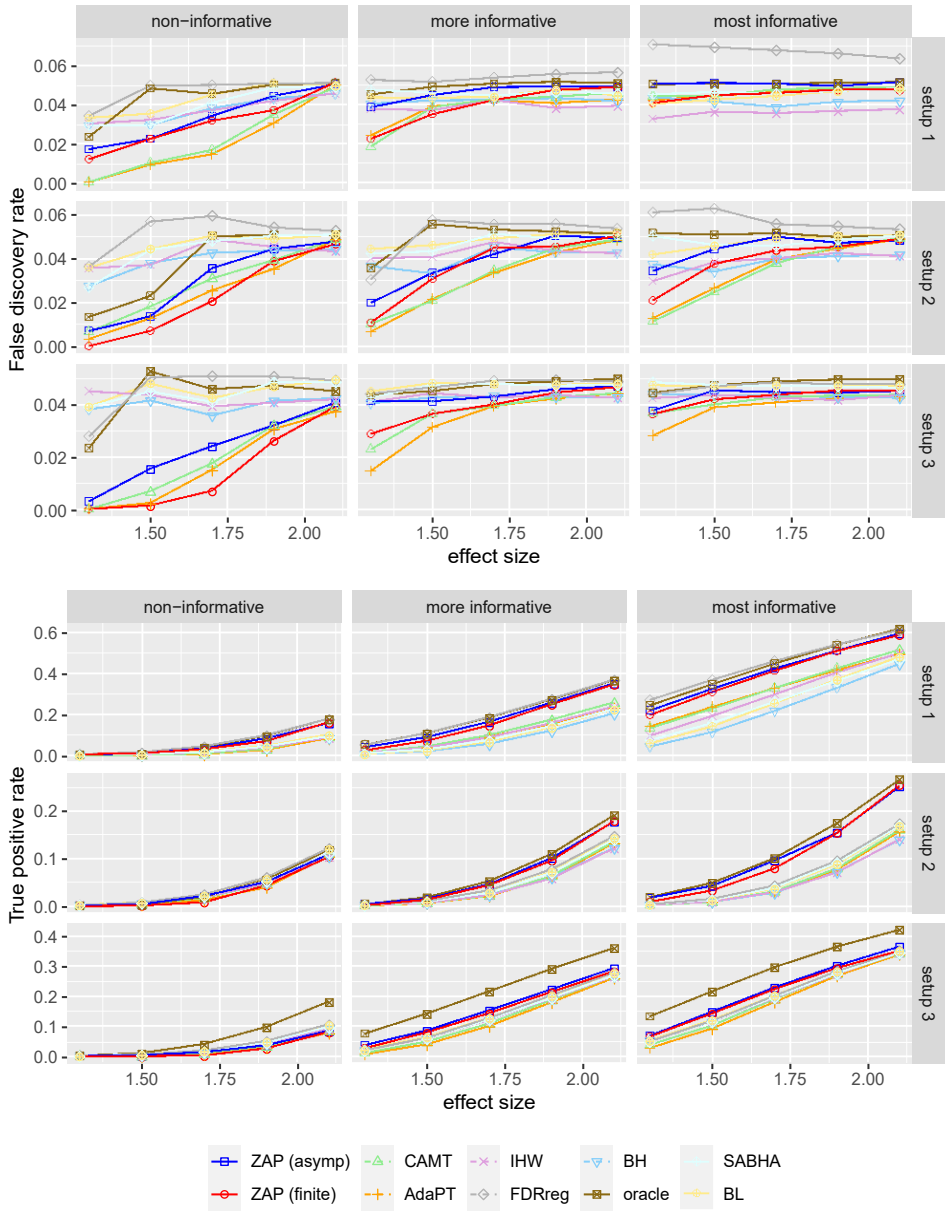

Figure G.1. FDR and TPR performances of an extended list of methods under Setup 1 - 3. All methods are applied at a targeted FDR level of 0:05. The x-axes show the values of . non-informative, more informative and most informative correspond to dierent values of from the smallest to the largest.

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

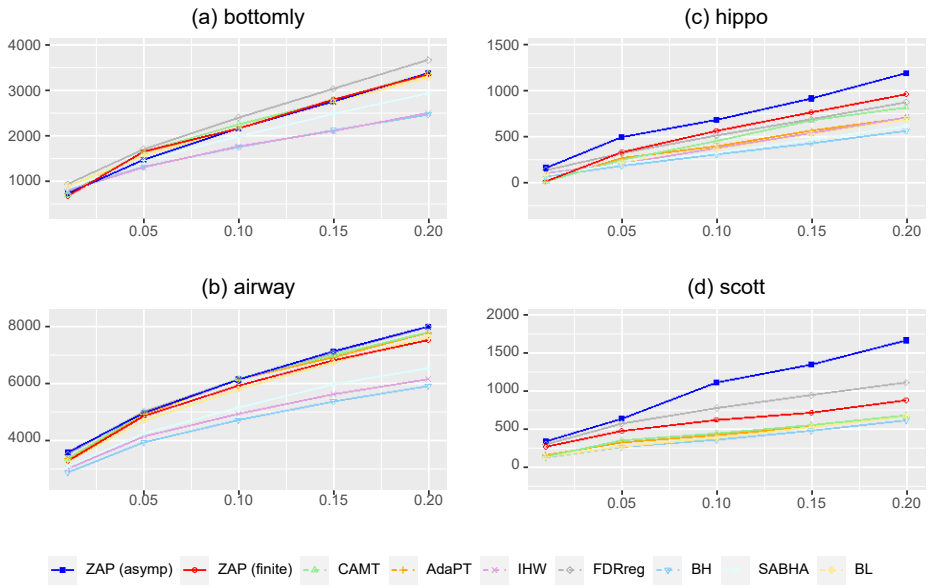

Figure G.2. (a)-(d) plot the numbers of rejections for different methods across datasets, against targeted FDR level at 0:01; 0:05; 0:1; 0:15; 0:2.

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 p0; 1q (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 spuri-ous 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 lter 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 ltering 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 pcq and Figure 4.2peq, the latter of which has its u-values produced by the ltered version of hippo dataset.

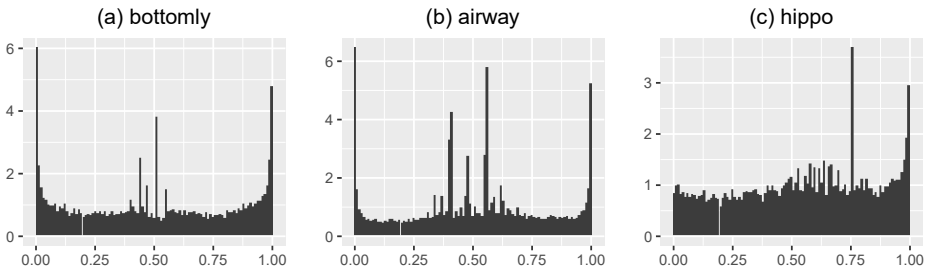

Figure G.3. Histograms of u-values for the original unltered versions of the three RNA-Seq datasets in the main text. In principle, one should only see at most two \spikes" on the two ends of the interval p0; 1q. Spikes not located close to 0 or 1 in any histogram result from genes with extremely low 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