AdaPT: An interactive procedure for multiple testing with side information

Lihua Lei and William Fithian

Department of Statistics, University of California, Berkeley, USA

E-mail: {lihua.lei, wfithian}@berkeley.edu

Summary. We consider the problem of multiple hypothesis testing with generic side information: for each hypothesis H_i we observe both a p-value p_i and some predictor x_i encoding contextual information about the hypothesis. For large-scale problems, adaptively focusing power on the more promising hypotheses (those more likely to yield discoveries) can lead to much more powerful multiple testing procedures. We propose a general iterative framework for this problem, called the Adaptive p-value Thresholding (AdaPT) procedure, which adaptively estimates a Bayes-optimal p-value rejection threshold and controls the false discovery rate (FDR) in finite samples. At each iteration of the procedure, the analyst proposes a rejection threshold and observes partially censored p-values, estimates the false discovery proportion (FDP) below the threshold, and proposes another threshold, until the estimated FDP is below α. Our procedure is adaptive in an unusually strong sense, permitting the analyst to use any statistical or machine learning method she chooses to estimate the optimal threshold, and to switch between different models at each iteration as information accrues. We demonstrate the favorable performance of AdaPT by comparing it to state-of-the-art methods in five real applications and two simulation studies.

Keywords: multiple testing, false discovery rate, p-value weighting, selective inference, adaptive inference, martingales

1. Introduction

1.1. Interactive data analysis

In classical statistics we assume that the question to be answered, and the analysis to be used in answering the question, are both fixed in advance of collecting the data. Many modern applications, however, involve extremely complex data sets that may be collected without any specific hypothesis in mind. Indeed, very often the express goal is to explore the data in search of insights we may not have expected to find. A central challenge in modern statistics is to provide scientists with methods that are flexible enough to allow for exploration, but that nevertheless provide statistical guarantees for the conclusions that are eventually reported.

Selective inference methods blend exploratory and confirmatory analysis by allowing a search over the space of potentially interesting questions, while still guaranteeing control of an appropriate Type I error rate such as a conditional error rate (e.g., Yekutieli, 2012; Lee et al., 2016; Fithian et al., 2014), familywise error rate (e.g., Tukey, 1994; Berk et al., 2013), or false discovery rate (e.g., Benjamini and Hochberg, 1995; Barber and Candès, 2015). However, most selective inference methods require that the selection algorithm be specified in advance, forcing a choice between either ignoring any difficult-to-formalize domain knowledge or sacrificing statistical validity guarantees.

Interactive data analysis methods relax the requirement of a pre-defined selection algorithm. Instead, they provide for an interactive analysis protocol between the analyst and the data, guaranteeing statistical validity as long as the protocol is followed. The two central questions in interactive data analysis are "what did the analyst know and when did she know it?" Previous methods for interactive data analysis involve randomization (Dwork et al., 2015; Tian et al., 2018) to control the analyst's access to the data at the time she decides what questions to ask.

This paper proposes an iterative, interactive method for multiple testing in the presence of *side information* about the hypotheses. We restrict the analyst's knowledge by partially

censoring all p-values smaller than a currently-proposed rejection threshold, and guarantee finite-sample FDR control by applying a version of the optional-stopping argument pioneered by Storey et al. (2004) and extended in Barber and Candès (2015); G'Sell et al. (2016); Li and Barber (2016a); Lei and Fithian (2016); Barber and Candès (2016).

1.2. Multiple testing with side information

In many areas of modern applied statistics, from genetics and neuroimaging to online advertising and finance, researchers routinely test thousands or millions or hypotheses at a time. For large-scale testing problems, perhaps the most celebrated multiple testing procedure of the modern era is the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Given n hypotheses and a p-value for each one, the BH procedure returns a list of rejections or "discoveries." If R is the number of total rejections and V is the number of false rejections (rejections of true null hypotheses), the BH procedure controls the *false discovery rate* (FDR), defined as

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

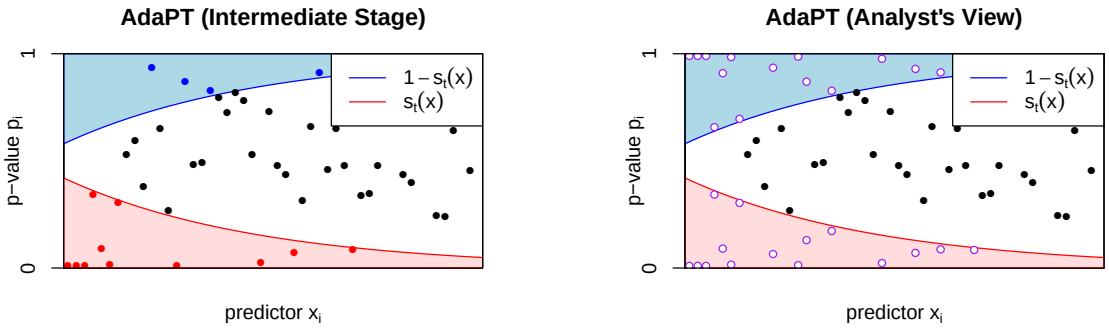
at a user-specified target level α. The random variable $V/\max\{R,1\}$ is called the *false discovery proportion* (FDP).

The BH procedure is nearly optimal when the null hypotheses are exchangeable *a priori*, and nearly all true. In other settings, however, the power can be improved, sometimes dramatically, by applying prior knowledge or by learning from the data. For example, adaptive FDR-controlling procedures can gain in power by estimating the overall proportion of true nulls (Storey, 2002), applying priors to increase power using p-value weights (Benjamini and Hochberg, 1997; Genovese et al., 2006; Dobriban et al., 2015; Dobriban, 2016), grouping similar null hypotheses and estimating the true null proportion within each group (Hu et al., 2012), or exploiting a prior ordering to focus power on more "promising" hypotheses near the top of the ordering (Barber and Candès, 2015; G'Sell et al., 2016; Li and Barber, 2016a; Lei and Fithian, 2016).

In most large-scale testing problems, the null hypotheses do not comprise an undifferentiated list; rather, each hypothesis is associated with rich contextual information that could potentially help to inform our testing procedures. For example, Li and Barber (2016a) test for differential expression of 22,283 genes between a treatment and control condition for a breast cancer drug, with side information in the form of an ordering of genes from most to least "promising" using auxiliary data collected at larger dosages. Multiple testing procedures that exploit the ordering can reject hundreds of hypotheses while the BH procedure (which does not exploit the ordering) rejects none.

More generally, prior information could arise in more complex ways. For example, consider testing for association of 400,000 single-nucleotide polymorphisms (SNPs) with each of 40 related diseases. If gene-regulatory relationships are known, then we might expect SNPs near related genes to be associated (or not) with related diseases, but without knowing ahead of time which gene-disease pairs are promising. In a similar vein, Fortney et al. (2015) used prior knowledge of each SNP's associations with age-related diseases to focus their search for SNPs associated with longevity, leading to novel discoveries. Inspired by examples like this, Ignatiadis et al. (2016) and Li and Barber (2016b) have recently proposed a more general problem setting where, for each hypothesis H_i, $i \in [n]$ we observe not only a p-value $p_i \in [0,1]$ but also a predictor x_i lying in some generic space $\mathcal{X}$. Unlike p_i, x_i carries only indirect information about the hypothesis: it is meant to capture some side information that might bear on H_i's likelihood to be false, or on the power of p_i under the alternative, but the nature of this relationship is not fully known ahead of time and must be learned from the data.

In other situations, the "predictor" information could simply represent a measure of sample size or overall signal for testing the ith hypothesis, which could be informative about the power of the ith test to distinguish the alternative from the null. For example, if each p_i concerns a test for association between the ith SNP and a disease, then the overall prevalence of that SNP



(a) $A_t = 4$ and $R_t = 11$ are the numbers of blue and red points respectively, leading to $\widehat{\text{FDP}} = (1+4)/11 \approx 0.45$. If $\widehat{\text{FDP}} \leq \alpha$, we stop and reject the red points; otherwise we choose a new threshold $s_{t+1} \preceq s_t$ and continue.

(b) Information available to the analyst when choosing $s_{t+1}(x)$ (A_t and R_t are also known). Each red and blue point is reflected across $p = 0.5$, leaving the analyst to impute which are the true p-values and which are the mirror images.

Fig. 1: Illustration of one step of the AdaPT procedure with a univariate predictor.

(in the combined treatment and control groups) can be used as prior information. Or, if p_i arises from a two-sample t-test, we could use the pooled variance, the sample variance ignoring the group labels, as prior information; see e.g. (Bourgon et al., 2010; Ignatiadis et al., 2016).

1.3. AdaPT: a framework for FDR control

This paper presents a new framework for FDR control with generic side information, which we call *adaptive p-value thresholding* or AdaPT for short. Our method proceeds iteratively: at each step $t = 0, 1, \ldots$, the analyst proposes a rejection threshold $s_t(x)$ and computes an estimator $\widehat{\text{FDP}}_t$ for the false discovery proportion for this threshold. If $\widehat{\text{FDP}}_t \leq \alpha$, she stops and rejects every H_i for which $p_i \leq s_t(x_i)$. Otherwise, she proposes a more stringent threshold $s_{t+1} \preceq s_t$ and moves on to the next iteration, where the notation $a \preceq b$ means $a(x) \leq b(x)$ for all $x \in \mathcal{X}$.

The estimator $\widehat{\text{FDP}}_t$ is computed by comparing the number R_t of rejections to the number A_t of p-values for which $p_i \geq 1 - s_t(x_i)$:

$$R_t = |\{i : \ p_i \leq s_t(x_i)\}|, \qquad A_t = |\{i : \ p_i \geq 1 - s_t(x_i)\}|, \qquad \text{and} \quad \widehat{\text{FDP}}_t = \frac{1 + A_t}{R_t \vee 1}.$$

The estimate $\widehat{\text{FDP}}_t$ is also used by Lei and Fithian (2016) and Arias-Castro and Chen (2016). Figure 1a illustrates the way $s_t(x)$ and $1 - s_t(x)$ partition the data into three regions; A_t is the number of points in the upper blue region and R_t is the number in the lower red region.

At each step t, the analyst can choose the next threshold $s_{t+1}(x)$ however she chooses, with only two constraints. First, $s_{t+1} \preceq s_t$ as stated before. Second, the large and small p-values (the ones contributing to A_t and R_t) are partially masked. Specifically, at step t the analyst is allowed to observe A_t and R_t, as well as the entire sequence $(x_i, \tilde{p}_{t,i})_{i=1}^n$, where

$$\tilde{p}_{t,i} = \begin{cases} p_i & s_t(x_i) < p_i < 1 - s_t(x_i) \\ \{p_i, \ 1 - p_i\} & \text{otherwise.} \end{cases} \tag{2}$$

Thus, if $p_i = 0.01 \leq s_t(x_i)$ then at step t the analyst knows only that p_i is either 0.01 or 0.99, but if $s_{t+1}(x_i) < 0.01$ then p_i is revealed at step $t + 1$ as 0.01. Figure 1b illustrates what the analyst can see: each red and blue point from Figure 1a is shown along with its mirror image reflected across the midline $p = 0.5$.

We show in Section 3 that, in a generic two-groups empirical Bayes model, an ideal choice for $s_t(x)$ would be a level surface of the *local false discovery rate* (fdr), as a function of x and p:

$$\mathrm{fdr}(p \mid x) = \mathbb{P}(H_i \text{ is null} \mid p_i = p, x_i = x).$$

Formally, $\mathrm{fdr}(p \mid x)$ is unidentifiable from the data but, under reasonable assumptions, we can use a good proxy based on the conditional density of the p-value given the covariate, $f(p \mid x)$ (note however that our method controls FDR without any empirical Bayes assumptions).

In each step information is gradually revealed to the analyst as the threshold shrinks and more p-values are unmasked. Our procedure is adaptive in an unusually strong sense: provided that the two constraints are met, the analyst may apply any method she wants to select $s_{t+1}(x)$, consulting her own hunches or the intuition of domain experts, and can even switch between different methods as information accrues. Moreover, the analyst is under no obligation to describe, or even to fully understand, her update rule for choosing $s_{t+1}(x)$. In this sense, we say our method is fully *interactive* — the analyst's behavior is arbitrary as long as she abides by a certain protocol for interacting with the algorithm.

While the partial masking of p-values obscures just enough information from the analyst to control the FDR, in many cases it does not seriously impact the ability of the analyst to learn the optimal threshold surface $s(x)$. This is because, by the time the algorithm is close to stopping, the vast majority of p-values have already been revealed, and many of the ones that remain masked are so minuscule as to leave little doubt about whether p_i is large or small. As we show in numerous simulation and real data experiments in Section 5, the fdr estimates based on masked data typically converge to the full-data estimates well before the algorithm stops.

The AdaPT procedure controls FDR at level α in finite samples provided that the null p-values are uniform, or mirror-conservative as defined in Section 2.1, and independent conditional on the non-null p-values. The proof relies on a pairwise exchangeability argument similar to the argument in Barber and Candès (2015).

Algorithm 1 summarizes the AdaPT procedure, using the generic sub-routine UPDATE to represent whatever process the analyst uses to select $s_{t+1}(x)$. Note that $s_{t+1}(x)$ is a random function that is measurable to $\mathcal{F}_t$. Sections 3–4 discuss recommendations for a good UPDATE routine. It is worth mentioning that AdaPT reduces to Barber-Candès method, inspired by Barber and Candès (2015) and proposed by Arias-Castro and Chen (2016), when $s_t(x)$ is a constant function for every t.

Algorithm 1 AdaPT

Input: predictors and p-values $(x_i, p_i)_{i\in[n]}$, initialization s_0, target FDR level α

Procedure:

for $t = 0, 1, \ldots$ **do**
  $\widehat{\mathrm{FDP}}_t \leftarrow \frac{1+A_t}{R_t \vee 1}$;
  if $\widehat{\mathrm{FDP}}_t \le \alpha$ **then**
    Reject $\{H_i : p_i \le s_t(x_i)\}$;
    Return s_t;
  end if
  $s_{t+1} \leftarrow \mathrm{UPDATE}((x_i, \tilde{p}_{t,i})_{i\in[n]}, A_t, R_t, s_t)$;
end for

1.4. Related work

In recent work Ignatiadis et al. (2016) propose a different method *independent hypothesis weighting* (IHW) for multiple testing with side information. They first bin the predictors into groups $g_1, \ldots, g_K$, and then apply the weighted-BH procedure at level α with piecewise-constant weights; i.e., if $x_i \in g_k$, then $w_i = w(g_k)$. The weights $w(g_1), \ldots, w(g_K)$ are chosen to maximize the number of rejections. This proposal is similar in spirit to the AdaPT procedure since it

attempts to find optimal weights, but it is a bit more limited: first, binning the data may be difficult if the predictor space $\mathcal{X}$ is multivariate or more complex; and second, their method is only guaranteed to control FDR asymptotically, as the number of bins stays fixed and the number of hypotheses in each bin grows to infinity. As a result, we must trust that n is large enough to support however many bins we have chosen to use. By contrast, AdaPT can use any machine-learning method to estimate $\hat{f}(p \mid x)$, and we can "overfit away" without fear of compromising finite-sample FDR control (though overfitting can of course reduce our power if our fdr estimates are too noisy). Another method is proposed by Du et al. (2014) when the covariate is an auxiliary univariate p-value derived by prior information. However, similar to Ignatiadis et al. (2016), it only controls FDR asymptotically under the fairly strong conditions that the p-values are symmetrically distributed under the null and bounded by $\frac{1}{2}$ under the alternative.

Perhaps the procedure most closely related to ours is the *structure-adaptive BH algorithm* or SABHA (Li and Barber, 2016b). SABHA first censors the p-values below at a fixed level τ ($\tau = 0.5$ in their simulations), leading to censored p-values $p_i\mathbf{1}\{p_i > \tau\}$. Using these, they can estimate $\pi_1(x)$, defined as $P(H_i \text{ is non-null} \mid x_i = x)$, as a function of x, then apply the weighted BH procedure of Genovese et al. (2006) with weights $\hat{\pi}_1(x_i)^{-1}$, at a corrected FDR level $\tilde{\alpha} = C\alpha$ (where $C < 1$ depends on the Rademacher complexity of the estimator $\hat{\pi}_1^{-1}$). We notice that this type of censoring is also employed in a variant of IHW (Ignatiadis and Huber, 2017), which guarantees the FDR control in finite samples.

As the first procedure to provably control the finite-sample FDR using generic feature information, SABHA represents a major step forward. However, AdaPT has several important advantages: First, even if $\hat{\pi}_1(x)$ estimates $\pi_1(x)$ consistently, the weights $\pi_1(x)^{-1}$ are not Bayes optimal as we show in Section 3; by contrast, our method estimates a Bayes optimal threshold. Second, the correction factor C makes the method conservative and restricts the available estimators $\hat{\pi}_1^{-1}$ to those with provably low Rademacher complexity. Third, AdaPT can use more information for learning: in later stages we will typically have $s_t(x_i) \ll 0.5$ and the masked p-values $\tilde{p}_{t,i}$ may be much more informative than $p_i\mathbf{1}\{p_i > 0.5\}$, especially since our goal is to estimate $f(p \mid x)$ for small values of p.

Finally, we remark that there is a literature on very different approaches for incorporating covariates into multiple testing problems; see e.g. Lewinger et al. (2007); Ferkingstad et al. (2008); Lawyer et al. (2009); Zablocki et al. (2014). Unlike our method (and IHW and SABHA), these approaches hinge on the correct specification of the model and might lose the statistical guarantee if the proposed model deviates from the ground truth. By contrast, our method (and IHW and SABHA) rely only on validity of p-values (see assumptions of Theorem 1 in next Section) and guarantee FDR control even when employing a misspecified model.

1.5. Outline

Section 2 defines the AdaPT procedure more formally and gives our main result: if the null p-values are independent and mirror-conservative (defined below), AdaPT controls FDR at level α in finite samples. Section 3 explains why selection of $s_{t+1}(x)$ will typically operate by first estimating the conditional density $f(p \mid x)$ as a function of x, and Section 4 gives practical suggestions for update rules. Section 5 illustrates the AdaPT procedure's power on five real datasets and two simulated datasets, and Section 6 concludes. The programs to replicate all our experiments can be obtained from `https://github.com/lihualei71/adaptPaper/`. Our `R` package `adaptMT` can be found in `https://github.com/lihualei71/adaptMT/`.

2. The AdaPT procedure

2.1. Notation and assumptions

Let $[n]$ denote the set $\{1, \ldots, n\}$. For each hypothesis H_i, $i \in [n]$ we observe $x_i \in \mathcal{X}$ and $p_i \in [0, 1]$. Let $\mathcal{H}_0$ denote the set of true null hypotheses. We will assume throughout that $(p_i)_{i\in\mathcal{H}_0}$ are mutually independent, and independent of $(p_i)_{i\notin\mathcal{H}_0}$ (see Section 6 for a discussion

of how we might relax the independence assumption). Finally, for each $i \in \mathcal{H}_0$, we assume that p_i is either uniform or mirror-conservative in a sense we will define shortly.

Let $\mathcal{F}_t$ for $t = 0, 1, \ldots$ represent the filtration generated by all information available to the user at step t:

$$\mathcal{F}_t = \sigma\left((x_i, \tilde{p}_{t,i})_{i=1}^n, A_t, R_t\right).$$

We similarly define an initial σ-field with all p-values masked, $\mathcal{F}_{-1} = \sigma\left((x_i, \{p_i, 1-p_i\})_{i=1}^n\right)$. The p-value masking is equivalent to requiring that $s_{t+1} \in \mathcal{F}_t$. (For simplicity we have implicitly ruled out the possibility that the analyst uses a randomized rule to update the threshold, but this restriction could be easily removed.) The two constraints $s_{t+1} \preceq s_t$ and $s_{t+1} \in \mathcal{F}_t$ ensure that $(\mathcal{F}_t)_{t=-1,0,1,\ldots}$ is a filtration; i.e., the information in $\mathcal{F}_t$ only grows from t to $t+1$:

LEMMA 1. *For all $t \geq -1$, $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$.*

PROOF. We use induction on u to show that $\mathcal{F}_u \subseteq \mathcal{F}_t$ for any $u \leq t$. The conclusion is trivial for $u = -1$ since $\{p_i, 1-p_i\}$ is always computable from $p_{t,i}$ (masked p-values can always be computed from masked or unmasked ones).

For $u \geq 0$, note that, by the inductive assumption, $s_u \in \mathcal{F}_{u-1} \subseteq \mathcal{F}_t$. As a result, we can compute $p_{u,i}$ which depends only on $p_{t,i}$ and $s_u(x_i)$. Furthermore,

$$R_u = R_t + \#\{i:\ p_{t,i} \in (s_t(x_i), s_u(x_i)]\}, \quad A_u = A_t + \#\{i:\ p_{t,i} \in [1-s_u(x_i), 1-s_t(x_i))\},$$

completing the proof.

To avoid trivialities we assume that the analyst always reveals at least one censored p-value in each step of the algorithm, since there is no reason ever to update the threshold surface in a way that reveals no new information. Thus, the stopping time $\hat{t} \leq n$ almost surely.

In many common settings, null p-values are conservative but not necessarily exactly uniform. For example, p-values from permutation tests are discrete, and p-values for composite null hypotheses are often conservative if the true value of the parameter lies in the interior of the null.

Our method does not require uniformity, but the standard definition of conservatism — that $\mathbb{P}_{H_i}(p_i \leq a) \leq a$ for all $0 \leq a \leq 1$ — is *not* enough to guarantee FDR control. Instead, we say that a p-value p_i is *mirror-conservative* if

$$\mathbb{P}_{H_i}(p_i \in [a_1, a_2]) \leq \mathbb{P}_{H_i}(p_i \in [1-a_2, 1-a_1]), \quad \text{for all } 0 \leq a_1 \leq a_2 \leq 0.5. \tag{3}$$

If p_i is discrete, (3) means $p_i = 1-a$ is at least as likely as $p_i = a$ for $a \leq 0.5$; if p_i has a continuous density, it means the density is at least as large at $1-a$ as at a. Mirror-conservatism is not a consequence of conservatism (take $p_i = 0.1 + 0.9B$ where $B \sim \text{Bernoulli}(0.9)$), and neither does it imply conservatism (take $p_i = B$). Any null distribution with an increasing density is evidently both conservative and mirror-conservative.

Permutation p-values are mirror-conservative, as are p-values for one-sided tests of univariate parameters with monotone likelihood ratio (with discrete p-values randomized to be uniform at the boundary between the null and alternative). See Appendix B.1 for proofs of these claims.

2.2. FDR control

We are now prepared to prove our main result: the AdaPT procedure controls FDR in finite samples. The proof relies on a similar optional stopping argument as the one presented in Lei and Fithian (2016) and Barber and Candès (2016) (themselves modifications of arguments in Storey et al. (2004) and Barber and Candès (2015)). Let V_t and U_t denote the numbers of *null* $p_i \leq s_t(x_i)$ and *null* $p_i \geq 1-s_t(x_i)$, respectively. If the null p-values are uniform then, no matter how we choose $s_t(x)$ at each step, we will always have $V_t \approx U_t$ and $\widehat{\text{FDP}}_t > \frac{U_t}{R_t \vee 1} \approx \frac{V_t}{R_t \vee 1}$.

LEMMA 2. *Suppose that, conditionally on the σ-field $\mathcal{G}_{-1}$, $b_1, \ldots, b_n$ are independent Bernoulli random variables with $\mathbb{P}(b_i = 1 \mid \mathcal{G}_{-1}) = \rho_i \geq \rho > 0$, almost surely. Also suppose that $[n] \supseteq \mathcal{C}_0 \supseteq \mathcal{C}_1 \supseteq \cdots$, with each subset $\mathcal{C}_{t+1}$ measurable with respect to*

$$\mathcal{G}_t = \sigma\left(\mathcal{G}_{-1}, \mathcal{C}_t, (b_i)_{i \notin \mathcal{C}_t}, \sum_{i \in \mathcal{C}_t} b_i\right).$$

If $\hat{t}$ is an almost-surely finite stopping time with respect to the filtration $(\mathcal{G}_t)_{t \geq 0}$, then

$$\mathbb{E}\left[\frac{1 + |\mathcal{C}_{\hat{t}}|}{1 + \sum_{i \in \mathcal{C}_{\hat{t}}} b_i} \mid \mathcal{G}_{-1}\right] \leq \rho^{-1}.$$

Our Lemma 2 generalizes Lemma 1 in Barber and Candès (2016) and uses a very similar technical argument. The proof is given in the appendix. Using Lemma 2, we can give our main result:

THEOREM 1. *Assume that the null p-values are independent of each other and of the non-null p-values, and the null p-values are uniform or mirror-conservative. Then the AdaPT procedure controls the FDR at level α, conditional on $\mathcal{F}_{-1}$ and also marginally.*

PROOF. Let $\hat{t}$ denote the step at which we stop and reject. Then

$$\mathrm{FDP}_{\hat{t}} \;=\; \frac{V_{\hat{t}}}{R_{\hat{t}} \vee 1} \;=\; \frac{1 + U_{\hat{t}}}{R_{\hat{t}} \vee 1} \cdot \frac{V_{\hat{t}}}{1 + U_{\hat{t}}} \;\leq\; \alpha \frac{V_{\hat{t}}}{1 + U_{\hat{t}}},$$

where the last step follows from the stopping condition that $\widehat{\mathrm{FDP}}_{\hat{t}} \leq \alpha$, and the fact that $U_t \leq A_t$. We will finish the proof by establishing that $\mathbb{E}[V_{\hat{t}}/(1 + U_{\hat{t}})] \leq 1$, using Lemma 2.

Let $m_i = \min\{p_i, 1 - p_i\}$ and $b_i = \mathbf{1}\{p_i \geq 0.5\}$, so $p_i = b_i(1 - m_i) + (1 - b_i)m_i$. Then knowing b_i and m_i is equivalent to knowing p_i. Let $\mathcal{C}_t = \{i \in \mathcal{H}_0 : \; p_i \notin (s_t(x_i), 1 - s_t(x_i))\}$, representing the null p-values that are *not* visible to the analyst at time t. Then,

$$U_t = \sum_{i \in \mathcal{C}_t} b_i, \quad \text{and} \quad V_t = \sum_{i \in \mathcal{C}_t} (1 - b_i) = |\mathcal{C}_t| - U_t.$$

Further, define the σ-fields

$$\mathcal{G}_{-1} = \sigma\left((x_i, m_i)_{i=1}^n, \; (b_i)_{i \notin \mathcal{H}_0}\right), \quad \text{and} \quad \mathcal{G}_t = \sigma\left(\mathcal{G}_{-1}, \mathcal{C}_t, (b_i)_{i \notin \mathcal{C}_t}, U_t\right).$$

The assumptions of independence and mirror-conservatism guarantee $\mathbb{P}\left(b_i = 1 \mid \mathcal{G}_{-1}\right) \geq 0.5$ almost surely for each $i \in \mathcal{H}_0$, with the b_i conditionally independent.

Next, note that $\mathcal{F}_t \subseteq \mathcal{G}_t$ because $p_i \in \mathcal{G}_t$ for each $p_i \in (s_t(x_i), 1 - s_t(x_i))$, and

$$A_t = U_t + \left|\{i \notin \mathcal{H}_0 : \; p_i \geq 1 - s_t(x_i)\}\right|,$$

and $R_t \in \mathcal{G}_t$ by a similar argument. It follows that $\hat{t} = \min\{t : \; \widehat{\mathrm{FDP}}_t \leq \alpha\}$ is a stopping time with respect to $\mathcal{G}_t$; furthermore, $\mathcal{C}_{t+1} \in \mathcal{F}_t \subseteq \mathcal{G}_t$ by assumption.

As a result, conditional on $\mathcal{G}_{-1}$, we can apply Lemma 2 to obtain

$$\mathbb{E}[\mathrm{FDP} \mid \mathcal{G}_{-1}] \;\leq\; \alpha\, \mathbb{E}\left[\frac{V_{\hat{t}}}{1 + U_{\hat{t}}} \mid \mathcal{G}_{-1}\right] \;=\; \alpha\, \mathbb{E}\left[\frac{1 + |\mathcal{C}_{\hat{t}}|}{1 + U_{\hat{t}}} - 1 \mid \mathcal{G}_{-1}\right] \;\leq\; \alpha\,(2 - 1) \;=\; \alpha.$$

Note that $\mathcal{F}_{-1} \subset \mathcal{G}_{-1}$. The proof is completed by applying the tower property of conditional expectation.

The main technical point of departure for our method is that the optional stopping argument is not merely a technical device to prove FDR control for a fixed algorithm like the BH, Storey-BH, or Knockoff+ procedures. Instead, we push the optional-stopping argument to its limit, allowing the analyst to interact with the data in a much more flexible and adaptive way. Sections 6.2–6.3 further investigate the connection to knockoffs.

3. A Guideline To Choose Thresholding Rules

Although the AdaPT procedure controls FDR no matter how we update the threshold, its power depends on the quality of the updates. This section concerns the question of what thresholds we would choose if we had perfect knowledge of the data-generating distribution, with Section 4 discussing suggestions for learning optimal thresholds from the data. To establish a guideline for threshold update, we consider a conditional two-groups model as the *working model*. As we will see, under mild conditions, the Bayes-optimal rejection thresholds are the level surfaces of the *local false discovery rate* (fdr), defined as the probability that a hypothesis is null conditional on its p-value. The local FDR was first discussed by Efron et al. (2001); see also Efron (2007). A similar result is obtained by Storey (2007) under a different framework.

3.1. The two-groups model and local false discovery rate

To begin, we assume a *two-groups model* conditional on the predictors x_i. Letting $H_i = 0$ if the ith null is true and $H_i = 1$ otherwise, we assume:

$$\begin{aligned} H_i \mid x_i &\sim \text{Bernoulli}(\pi_1(x_i)) \\ p_i \mid H_i, x_i &\sim \begin{cases} f_0(p \mid x_i) & \text{if } H_i = 0 \\ f_1(p \mid x_i) & \text{if } H_i = 1 \end{cases}. \end{aligned}$$

In addition, we assume that (x_i, H_i, p_i) are independent for $i \in [n]$. Unless otherwise stated we will assume for simplicity that both f_0 and f_1 are continuous densities, with $f_0(p \mid x) \equiv 1$ (null p-values are uniform) and $f_1(p \mid x)$ non-increasing in p (smaller p-values imply stronger evidence against the null). Furthermore, define the conditional mixture density

$$f(p \mid x) \;=\; (1 - \pi_1(x))\, f_0(p \mid x) + \pi_1(x) f_1(p \mid x) \;=\; 1 - \pi_1(x) + \pi_1(x) f_1(p \mid x),$$

and the conditional local false discovery rate

$$\text{fdr}(p \mid x) \;=\; \mathbb{P}(H_i \text{ is null } \mid x_i = x, p_i = p) \;=\; \frac{1 - \pi_1(x)}{f(p \mid x)}.$$

Note that we never observe H_i directly. Thus, while f is identifiable from the data, π_1 and f_1 are not: for example, $\pi_1 = 0.5, f_1(p \mid x) = 2(1-p)$ and $\pi_1 = 1, f_1(p \mid x) = 1.5 - p$ result in exactly the same mixture density. Unless $f_1(p \mid x)$ is known *a priori*, we can make the conservative identifying assumption that

$$1 - \pi_1(x) = \inf_{p \in [0,1]} f(p \mid x) = f(1 \mid x),$$

attributing as many observations as possible to the null hypothesis. This approximation is very good when $\text{fdr}(1 \mid x) \approx 1$, which is reasonable in many settings. Thus, any estimate $\hat{f}$ of the mixture density translates to a conservative estimate $\widehat{\text{fdr}}(p \mid x) = \hat{f}(1 \mid x)/\hat{f}(p \mid x)$.

3.2. Optimal thresholds under the two-groups model

Let ν be a probability measure on $\mathcal{X}$ and define a random variable $X \sim \nu$. Similar to Sun et al. (2015), for any thresholding rule $s(x)$, we define the global FDR as

$$\text{FDR}(s; \nu) = \mathbb{P}(H = 0 \mid H \text{ is rejected}) = \mathbb{P}(H = 0 \mid P \le s(X))$$

where H and P are a hypothesis and p-value distributed according to the two-groups model. The power is defined in a similar fashion as

$$\text{Pow}(s; \nu) = \mathbb{P}(H \text{ is rejected} \mid H = 1) = \mathbb{P}(P \le s(X) \mid H = 1).$$

Sun et al. (2015) formulates a compound decision-theoretic framework by defining a Bayesian-type loss function. Instead, we propose a Neyman-Pearson type framework, i.e.

$$\max_s \ \mathrm{Pow}(s;\nu) \quad \text{s.t. } \mathrm{FDR}(s;\nu) \le \alpha. \tag{4}$$

Next, define

$$Q_0(s) = \mathbb{P}(P \le s(X), H = 0) = \int_{\mathcal{X}} F_0(s(x)|x)(1-\pi_1(x))\nu(dx)$$
$$Q_1(s) = \mathbb{P}(P \le s(X), H = 1) = \int_{\mathcal{X}} F_1(s(x)|x)\pi_1(x)\nu(dx),$$

where F_0 and F_1 are the cumulative distribution functions under the null and alternative. We can simplify (4) as

$$\max_s \frac{Q_1(s)}{\mathbb{P}(H=1)} \quad \text{s.t. } \frac{Q_0(s)}{Q_0(s)+Q_1(s)} \le \alpha \tag{5}$$
$$\iff \min_s \ -Q_1(s) \qquad \text{s.t. } -\alpha Q_1(s) + (1-\alpha)Q_0(s) \le 0 \tag{6}$$
$$\iff \min_s \int_{\mathcal{X}} -F_1(s(x)|x)\pi_1(x)\nu(dx)$$
$$\text{s.t. } \int_{\mathcal{X}} \Big\{ -\alpha F_1(s(x)|x)\pi_1(x) + (1-\alpha)F_0(s(x)|x)(1-\pi_1(x)) \Big\}\nu(dx) \le 0. \tag{7}$$

The corresponding Lagrangian function can be written as

$$L(s;\lambda) = \int_{\mathcal{X}} \Big\{ -(1+\lambda\alpha)F_1(s(x)|x)\pi_1(x) + \lambda(1-\alpha)F_0(s(x)|x)(1-\pi_1(x)) \Big\}\nu(dx). \tag{8}$$

Let s^* be the optimum, then the Karush-Kuhn-Tucker (KKT) condition (under regularity conditions) implies that

$$(1+\lambda\alpha)f_1(s^*(x)|x)\pi_1(x) = \lambda(1-\alpha)f_0(s^*(x)|x)(1-\pi_1(x))$$
$$\implies \mathrm{fdr}(s^*(x)|x) = \frac{1+\lambda\alpha}{1+\lambda}. \tag{9}$$

In other words, the optimal thresholding rules are level surfaces of local FDR. Theorem 2 formalizes the above derivation by clarifying the regularity conditions.

THEOREM 2. *Assume that*

(a) $f_1(p \mid x_i)$ *is continuously non-increasing and* $f_0(p \mid x_i)$ *is continuously non-decreasing and uniformly bounded away from* ∞*;*

(b) ν *is a discrete measure supported on* $\{x_1, \ldots, x_n\}$ *with* $\nu(\{x_i : \mathrm{fdr}(0 \mid x_i) < \alpha, f(0 \mid x_i) > 0\}) > 0$.

Then (4) *has at least a solution, and all solutions are level surfaces of* $\mathrm{fdr}(p \mid x)$.

In practice, any conservative null distribution (stochastically dominated by $U([0,1])$) with positive density at zero satisfies condition (a). The monotonicity of f_1 is also valid since smaller p-values imply stronger evidence against null. In condition (b), the assumption on the support is reasonable since we treat $\{x_i : i \in [n]\}$ as fixed and hence only the quantities associated with these values are of interest. We believe it can be relaxed to more general measures and will not discuss it due to the technical complication. In contrast, the second requirement is necessary since it implies the feasibility of the problem. If the local FDR is above α almost everywhere, no thresholding rule is able to control FDR at α. As mentioned above, we can set s as the level surfaces of $\widehat{\mathrm{fdr}}(p \mid x) = \hat{f}(1 \mid x)/\hat{f}(p \mid x)$ given some estimator $\hat{f}(p \mid x)$. The next section discusses estimation of $\hat{f}(p \mid x)$.

4. Implementation

Having shown that level surfaces of the local FDR are optimal under the two-groups model, we now turn to estimation of $\mathrm{fdr}(p \mid x)$, which boils down to estimation of the conditional density $f(p \mid x)$. This section discusses a flexible framework for conditional density estimation that can perform favorably when no domain-specific expertise can be brought to bear.

More generally, we should model the data using as much domain-specific expertise as possible. We emphasize once more that, no matter how misspecified our model is, no matter how misguided our priors are (if we use a Bayesian method), no matter how we select a model or tuning parameter, or how much that selection biases our resulting estimate of local FDR, the AdaPT procedure nevertheless controls global FDR. Thus, there is every reason to be relatively aggressive in choosing a modeling strategy.

4.1. Conditional density estimation via the expecation maximization algorithm

Generically, we can model the conditional density by a parametric family where we assume null p-values are uniform distributed, i.e. $f_0(p \mid x_i) \equiv 1$, and each non-null p-value has a density in the following exponential family, indexed by a univariate parameter η_i:

$$f_1(p \mid x_i) = h(p; \eta_i) \triangleq e^{\eta_i g(p) - B(\eta_i)}. \tag{10}$$

Note that η_i and $g(p)$ can be vectors but we focus on the scalar case for simplicity. Let

$$y_i = g(p_i), \quad \mu_i = B'(\eta_i). \tag{11}$$

Using the standard argument, (10) implies that

$$\mathbb{E}_{\eta_i}[y_i] = \mathbb{E}_{\eta_i}[g(p_i)] = B'(\eta_i) = \mu_i, \tag{12}$$

where $\mathbb{E}_{\eta_i}$ denotes the expectation under $h(\cdot; \eta_i)$. If g is not almost-everywhere constant, then $B''(\eta) = \mathrm{Var}_\eta(y_i) > 0$ and B' is bijective. Then there is a one-to-one mapping from μ_i to η_i, denoted by $\eta_i = \eta(\mu_i)$ as convention. In fact, $\eta(\cdot) = (B')^{-1}(\cdot)$. Then (10) can be reparametrized using μ_i,

$$h(p; \mu_i) = e^{\eta(\mu_i) g(p) - A(\mu_i)}, \tag{13}$$

where $A(\cdot) = B(\eta(\cdot))$ and we abuse the notation $h(p; \cdot)$. As we will see, it is more convenient to use the mean parametrization (13).

Given (13), it is left to model $\pi_{1i} \triangleq \pi_1(x_i)$ and μ_i (or η_i equivalently). In this article we consider the following generalized linear model where $\phi_\pi(x), \phi_\mu(x)$ denote two featurization and ζ denotes a link function:

$$\begin{aligned} H_i \mid x_i &\sim \mathrm{Bernoulli}(\pi_{1i}), \qquad \text{with } \log \frac{\pi_{1i}}{1-\pi_{1i}} = \theta' \phi_\pi(x_i), \text{ and} \\ p_i \mid x_i, H_i &\sim \begin{cases} h(p; \mu_i) & \text{if } H_i = 1 \\ 1 & \text{if } H_i = 0 \end{cases}, \quad \text{with } \zeta(\mu_i) = \beta' \phi_\mu(x_i). \end{aligned} \tag{14}$$

In particular, $\zeta(\cdot) = \eta(\cdot)$ gives the canonical link function. For instance, when $g(p) = -\log p, \eta(\mu) = -\frac{1}{\mu} + 1$ and $A(\mu) = \log \mu$,

$$f(p|x) = \pi_{1i} h(p; \mu_i) + (1 - \pi_{1i}) = \pi_{1i} \cdot \frac{1}{\mu_i} p^{\frac{1}{\mu_i} - 1} + (1 - \pi_{1i}). \tag{15}$$

This yields a beta-mixture model on the conditional density, which has been considered in literature, e.g. Parker and Rothenberg (1988); Allison et al. (2002); Pounds and Morris (2003); Markitsis and Lai (2010).

The fully-observed log-likelihood for the model (14) is

$$\ell(\theta, \beta; p, H, x) = \sum_{i=1}^{n} \left\{ H_i \theta' \phi_\pi(x_i) - \log\left(1 + e^{-\theta' \phi_\pi(x_i)}\right) \right\} + \sum_{i=1}^{n} H_i \{ y_i \cdot \eta \circ \zeta^{-1}(\beta' \phi_\mu(x_i)) - A \circ \zeta^{-1}(\beta' \phi_\mu(x_i)) \} \tag{16}$$

Because some values of y_i and all values of H_i are unknown, we can use the expectation maximization (EM) algorithm to maximize the partially observed log-likelihood. To simplify estimation, we will proceed as though A_t and R_t are missing, so that the (y_i, H_i) pairs are mutually independent given the predictors. That is, at step t of the AdaPT procedure we attempt to maximize the likelihood of the data $D_t = (x_i, \tilde{p}_{t,i})_{i \in [n]}$ and treating s_t as fixed.

Recall that $b_i = I(p_i \geq 0.5)$. There are four possible values of (b_i, H_i), with each pair conditionally independent given D_t, and whose probabilities can be efficiently computed for any values of θ and β. Let $r = 0, 1, \ldots$ index stages of the EM algorithm (recall t is fixed for the duration of the EM algorithm). For the E-step we compute the expectation of the log-likelihood,

$$\mathbb{E}_{\hat{\theta}^{(r-1)}, \hat{\beta}^{(r-1)}} \left[\ell(\theta, \beta; y, H, x) | D_t \right],$$

which amounts to computing the following quantities:

$$\widehat{H}_i^{(r)} = \mathbb{E}_{\hat{\theta}^{(r-1)}, \hat{\beta}^{(r-1)}}[H_i \mid D_t], \quad \text{and} \tag{17}$$

$$\hat{y}_i^{(r,1)} = \mathbb{E}_{\hat{\theta}^{(r-1)}, \hat{\beta}^{(r-1)}}[y_i H_i \mid D_t,\ H_i = 1] / \widehat{H}_i^{(r)}, \tag{18}$$

where $\hat{\theta}^{(r)}$ and $\hat{\beta}^{(r)}$ denote the current coefficient estimates. We derive the exact formula for (17) and (18) in Appendix A.1. For the M-step, we set

$$\begin{aligned} \hat{\theta}^{(r)}, \hat{\beta}^{(r)} &= \arg\max_{\beta, \theta}\ \mathbb{E}_{\hat{\theta}^{(r-1)}, \hat{\beta}^{(r-1)}} \left[\ell(\theta, \beta; y, H, x) \mid D_t \right] \\ &= \arg\max_{\beta, \theta}\ \sum_{i=1}^{n} \widehat{H}_i^{(r)} \theta' \phi_\pi(x_i) - \log\left(1 + e^{-\theta' \phi_\pi(x_i)}\right) \\ &\qquad + \sum_{i=1}^{n} \widehat{H}_i^{(r)} \cdot \left(\hat{y}_i^{(r,1)} \cdot \eta \circ \zeta^{-1}(\beta' \phi_\mu(x_i)) - A \circ \zeta^{-1}(\beta' \phi_\mu(x_i)) \right). \end{aligned} \tag{19}$$

The optimization above splits into two separate optimization problems, a logistic regression with predictors $\phi_\pi(x_i)$ and fractional responses $\widehat{H}_i^{(r)}$, and a GLM with predictors $\phi_\mu(x_i)$, responses $\hat{y}_i^{(r,1)}$, and weights $\widehat{H}_i^{(r)}$. Each of these GLM problems can be solved efficiently using the `glm` function in `R` (e.g. Dobson and Barnett (2008)). For $r = 0$, we can initialize $\hat{\theta}^{(0)}$ and $\hat{\beta}^{(0)}$ by a simple method with details discussed in Appendix A.2. Algorithm 2 formalizes the EM algorithm using R pseudocode. The family argument for estimating $\hat{\beta}^{(r)}$ depends on the form of exponential family (13). For example, (19) yields a Gamma GLM in the beta-mixture model (15).

The GLM model (14) provides the starting point for an extremely flexible and extensible modeling framework. More generally, we could replace the fitting procedure in M-step by penalized GLM (`glmnet` package), generalized additive model (`gam` or `mgcv` package), or generalized boosting regression (`gbm` package). Furthermore, noting that

$$\pi_1(x) = \mathbb{E}[H \mid x], \quad \mu(x) = \mathbb{E}[y \mid x, H = 1],$$

one can even fit them directly using any nonparametric method, such as random forest or neural networks, that targets on estimating conditional mean.

Algorithm 2 EM algorithm to estimate $\pi_1(\cdot)$ and $\mu(\cdot)$ based on $D_t = (x_i, \tilde{p}_{t,i})_{i\in[n]}$

Input: data D_t, number of iterations m, initialization $\hat{\theta}^{(0)}, \hat{\beta}^{(0)}$;

for $r = 1, 2, \ldots, m$ **do**

(E-*step*):

$\widehat{H}_i^{(r)} \leftarrow \mathbb{E}_{\hat{\theta}^{(r-1)},\hat{\beta}^{(r-1)}}[H_i \mid D_t], \quad i \in [n];$

$\hat{y}_i^{(r,1)} \leftarrow \mathbb{E}_{\hat{\theta}^{(r-1)},\hat{\beta}^{(r-1)}}[y_i H_i \mid D_t, H_i = 1]/\widehat{H}_i^{(r)}, \quad i \in [n];$

(M-*step*):

$\hat{\theta}^{(r)} \leftarrow \texttt{glm}\left(\widehat{H}^{(r)} \sim \phi_\pi(x),\ \texttt{family = binomial}\right);$

$\hat{\beta}^{(r)} \leftarrow \texttt{glm}\left(\hat{y}^{(r,1)} \sim \phi_\mu(x),\ \texttt{family = ...(link} = \zeta\texttt{)},\ \texttt{weights = } \widehat{H}^{(r)}\right);$

end for

Output: $\hat{\pi}_1(x) = \left(1 + e^{-\phi_\pi(x)'\hat{\theta}^{(m)}}\right)^{-1}, \ \hat{\mu}(x) = \zeta^{-1}\left(\phi_\mu(x)'\hat{\beta}^{(m)}\right).$

4.2. Selecting featurization

Suppose we are given a finite set of candidate featurization $\{(\phi_{\pi,j}(x), \phi_{\mu,j}(x)) : j = 1, \ldots, M\}$. For instance for univariate x, $\phi_{\pi,j}(x)$ and $\phi_{\mu,j}(x)$ could be spline bases with certain numbers of equi-spaced knots; for multivariate x, $\phi_{\pi,j}(x)$ and $\phi_{\mu,j}(x)$ could be subsets of covariates contained in x. At step t, one is permitted to fit a model for each featurization, using arbitrary methods (e.g., GLM, penalized GLM, etc.), based on $((\phi_{\pi,j}(x_i), \phi_{\mu,j}(x_i), \tilde{p}_{t,i})_{i=1}^n)$. Let $\hat{\pi}_1^{(j)} = (\hat{\pi}_{11}^{(j)}, \ldots, \hat{\pi}_{1n}^{(j)})$ and $\hat{\mu}^{(j)} = (\hat{\mu}_1^{(j)}, \ldots, \hat{\mu}_n^{(j)})$ denote the resulting fitted values. The full log-likelihood, assuming H_i is known, for the GLM model (14) based on $(\phi_{\pi,j}(x), \phi_{\mu,j}(x))$ can be written as

$$\ell_j(\pi_1, \mu) = \sum_{i=1}^n (H_i \log \pi_{1i}^{(j)} + (1 - H_i)\log(1 - \pi_{1i}^{(j)})) + \sum_{i=1}^n H_i \log h(p_i; \mu_i^{(j)}).$$

Though ℓ_j is not computable, we can replace it by

$$\tilde{\ell}_j \triangleq \mathbb{E}_{\hat{\pi}_1^{(j)},\hat{\mu}^{(j)}}[\ell_j(\pi_1, \mu)].$$

This is precisely the objective of M-step and hence is directly computed from the EM algorithm.

Based on $\{\tilde{\ell}_j\}_{j=1}^M$, we can use any information criterion for featurization selection. Our implementation uses BIC as default, defined as

$$\mathrm{BIC}_j = \log n \cdot (\mathrm{df}_{\pi,j} + \mathrm{df}_{\mu,j}) - 2\tilde{\ell}_j$$

where $\mathrm{df}_{\pi,j}$ (resp. $\mathrm{df}_{\mu,j}$) is the degree of freedom of $\phi_{\pi,j}$ (resp. $\phi_{\mu,j}$). For instance, $\mathrm{df}_{\pi,j}$ is the number of knots plus 1 (for the intercept) when $\phi_{\pi,j}$ is the spline basis; $\mathrm{df}_{\pi,j}$ is the number of selected covariates plus 1 (for the intercept) when $\phi_{\pi,j}$ is a sparse subset of x.

Alternatively, the user can also apply cross-validation to select the featurization. Specifically, at step t the data is divided into K folds. For k-th fold, the expected log-likelihood $\tilde{\ell}_{jk}$ is computed by taking the k-th fold as the holdout set and fitting the parameters on other folds. The selection is then based on $\tilde{\ell}_j = \sum_{k=1}^K \tilde{\ell}_{jk}$.

We emphasize that any of above selection procedures can be performed in any intermediate step of AdaPT. If the featurization selection can be computed efficiently, we suggest applying it in every step. Otherwise we suggest performing it only at the first step, in which $s(x) = s_0(x)$, and keeping the selected featurization for all later steps.

4.3. Updating the threshold

Theorem 2 suggests that our updated threshold s_{t+1} should approximate a level surface of $\widehat{\mathrm{fdr}}(p \mid x)$. For the model (14), level surfaces of the local FDR are given by

$$c = \frac{f(1|x)}{f(s(x)|x)} = \frac{\pi_1(x)h(1; \mu(x)) + 1 - \pi_1(x)}{\pi_1(x)h(s(x); \mu(x)) + 1 - \pi_1(x)}. \tag{20}$$

For various widely-used exponential families in the form (13), $h(p;\mu)$ is decreasing with respect to p, in which case,

$$s(x;c) = f^{-1}\left(\frac{h(1;\mu(x))}{c} + \frac{1-\pi_1(x)}{\pi_1(x)}\frac{1-c}{c}; \mu(x)\right) \tag{21}$$

Given a chosen local FDR level c, we can evolve s_t by

$$s_{t+1}(x) = \min\{s_t(x), s(x;c)\}, \tag{22}$$

where the minimum is taken to meet the requirement that $s_{t+1}(x) \le s_t(x)$. Note that a higher level surface (larger c) will typically give a higher $\widehat{\mathrm{FDP}}_t$ and vice versa. Unless computational efficiency is at a premium, it is better to force the procedure to be patient since more information can be gained after each update and the learning step can be more accurate. In other words, we shall choose a large c such that $s_{t+1}(x)$ only deviates from $s_t(x)$ slightly.

In this article we propose a simple procedure to achieve this: it chooses c such that exactly one partially-masked p-value is revealed based on $s_{t+1}(x)$ defined in (22). The choice of c can be computed in the following way

(a) Estimate local FDR for each $p'_{t,i}$ as

$$\mathrm{fdr}_{t,i} = \frac{f(1|x_i)}{f(p'_{t,i}|x_i)} = \frac{\hat\pi_{1i}\cdot h(1;\hat\mu_i) + 1 - \hat\pi_{1i}}{\hat\pi_{1i}\cdot h(p'_{t,i};\hat\mu_i) + 1 - \hat\pi_{1i}}, \tag{23}$$

where $p'_{t,i}$ is the minimum element in $\tilde p_{t,i}$ (i.e., $\tilde p_{t,i} = p'_{t,i}$ for revealed p-values and $\tilde p_{t,i} = \{p'_{t,i}, 1-p'_{t,i}\}$ for masked p-values.)

(b) Set c as the largest value of $\mathrm{lfdr}_{t,i}$ among all *partially masked p-values*. (Strictly speaking, c should be slightly smaller than $\max_i \mathrm{lfdr}_{t,i}$. In implementation we subtract 10^{-15} from it.)

As a consequence, this choice of c is measurable with respect to $\mathcal{F}_t$ and hence a permissible operation in AdaPT .

4.4. Other Issues

Initial thresholds. As shown in Algorithm 1, AdaPT starts from some curve $s_0(x)$ and then slowly update it. If the hypotheses are not ordered, then we can simply set $s_0(x) \equiv s_{0,1}$ with $s_{0,1} \le 0.5$. A larger $s_{0,1}$ is conceptually preferred since the procedure is more patient. We found that $s_{0,1} = 0.45$ is a consistently good choice.

Computation efficiency. The model update (Algorithm 2) is the most computationally costly component. To save computation, we recommend not updating the model at every step. In our implementation, the default is to update the model every $\lceil n/20 \rceil$ steps.

q-Values. Rather than specify α in advance, some researchers might prefer to see a list of discoveries for each of a range of α values. Rather than return a single list for a single α, we can alternatively run the algorithm once and output *q-values* for every hypothesis (Storey, 2002; Storey and Tibshirani, 2003), defined as the minimum value of α for which the hypothesis would be rejected.

Let $\hat t_\alpha = \min\{t:\ \widehat{\mathrm{FDP}}_t \le \alpha\}$ and

$$t_i^* = \min\{t:\ s_t(x_i) < p_i < 1 - s_t(x_i)\},$$

the time at which p_i is revealed. We then see that

$$\begin{aligned} H_i \text{ rejected at level } \alpha &\iff p_i \le s_{\hat t_\alpha}(x_i) \\ &\iff \hat t_\alpha < t_i^* \\ &\iff \min_{t<t_i^*} \widehat{\mathrm{FDP}}_t \le \alpha \end{aligned}$$

As a result, $q_i = \min_{t<t_i^*} \widehat{\mathrm{FDP}}_t$ is a valid q-value for hypothesis i.

5. Experiments

5.1. Gene/Drug response data: an illustrating example

To illustrate the power of the AdaPT procedure, we apply it to the GEOquery gene-dosage data (Davis and Meltzer, 2007), which has been analyzed repeatedly as a benchmark for ordered testing procedures Li and Barber (2016a); Lei and Fithian (2016); Li and Barber (2016b). We use Algorithm 2 with a beta-mixture model (15) for the E-step (see Appendix A.1.1 for details) and a Gamma GLM with canonical link function for the M-step. This dataset consists of gene expression measurements for $n = 22283$ genes, in response to estrogen treatments in breast cancer cells for five groups of patients, with different dosage levels and 5 trials in each. The task is to identify the genes responding to a low dosage. The p-values p_i for gene i is obtained by a one-sided permutation test which evaluates evidence for a change in gene expression level between the control group (placebo) and the low-dose group. $\{p_i : i \in [n]\}$ are then ordered according to permutation t-statistics comparing the control and low-dose data, pooled, against data from a higher dosage (with genes that appear to have a strong response at higher dosages placed earlier in the list).

We consider two orderings: first, a stronger (more informative) ordering based on a comparison to the highest dosage; and second, a weaker (less informative) ordering based on a comparison to a medium dosage. Let $\sigma_S(i)$ and $\sigma_W(i)$ denote respectively the permutations of $i \in [n]$ given by the stronger and weaker orderings. Further details on these two orderings can be found in Li and Barber (2016a) and Li and Barber (2016b). We write the p-values, thus reordered, as $p_i^S = p_{\sigma_S(i)}$ and $p_i^W = p_{\sigma_W(i)}$. Once the data are reordered, we can apply either a method that ignores the ordering altogether, or an ordered testing procedure, or a testing procedure that uses generic side information, using the index of the reordered p-values as a univariate predictor.

We compare AdaPT against twelve other methods :

(a) SeqStep with parameter $C = 2$ (Barber and Candès, 2015);

(b) ForwardStop (G'Sell et al., 2016);

(c) the accumulation test with the HingeExp function and parameter $C = 2$ (Li and Barber, 2016a);

(d) Adaptive SeqStep with $s = q$ and $\lambda = 1 - q$ (Lei and Fithian, 2016);

(e) BH procedure (Benjamini and Hochberg, 1995);

(f) Storey's BH procedure with threshold $\lambda = 0.5$ (Storey et al., 2004);

(g) Barber-Candès method (Barber and Candès, 2015; Arias-Castro and Chen, 2016);

(h) SABHA with $\tau = 0.5, \epsilon = 0.1$ and the stepwise constant weights, monotone taking values in $\{\epsilon, 1\}$ (see section 4.1 of Li and Barber (2016b));

(i) SABHA with $\tau = 0.5, \epsilon = 0.1$ and the monotone weights, taking values in $[\epsilon, 1]$ (see section 4.1 of Li and Barber (2016b));

(j) Independent Hypothesis Weighting (IHW) with number of bins and folds set as default (Ignatiadis et al., 2016);

(k) an oracle version of IHW with the number of bins determined by maximizing the number of rejections;

(l) an oracle version of Independent Filtering (IF) with the cutoff determined by maximizing the number of rejections (Bourgon et al., 2010).

Note that the last two methods do not guarantee FDR control because the optimal parameter is selected; and both versions of SABHA control FDR at level 1.134α (Lemma 1 of Li and Barber (2016b)) when the target level is α. Despite the potential anti-conservativeness of these methods, we do not make correction in order to compare their best possible performance to AdaPT . Figure 2 shows the number of discoveries with different target FDR levels. We only show the range of α from 0.01 to 0.3 since it is rare to allow FDR to be above 0.3 in practice. We use different featurization for estimating $\pi(x)$ and $\mu(x)$, selected from the combination of all spline basis with $6-15$ equi-quantile knots via BIC criterion at the initial step and kept the same afterwards; see Section 4.2.

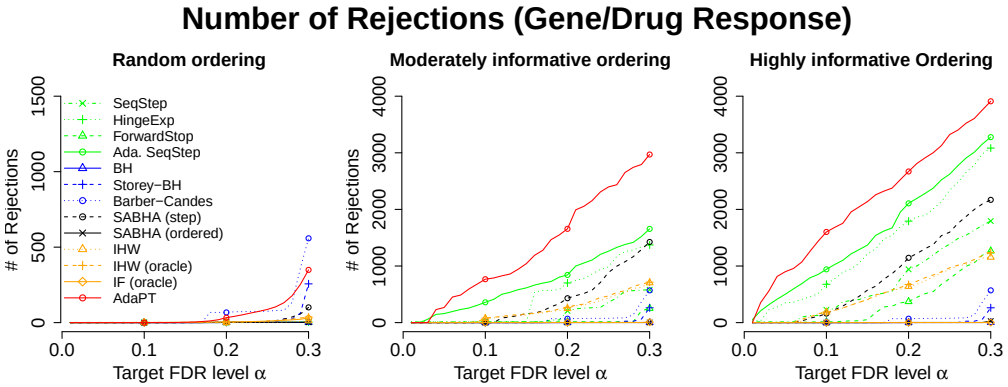


Fig. 2: Number of discoveries, in gene/drug response dataset, by each method at a range of target FDR levels α from 0.01 to 0.30. Each panel plots the results for an ordering, ranging from random ordering to highly informative.

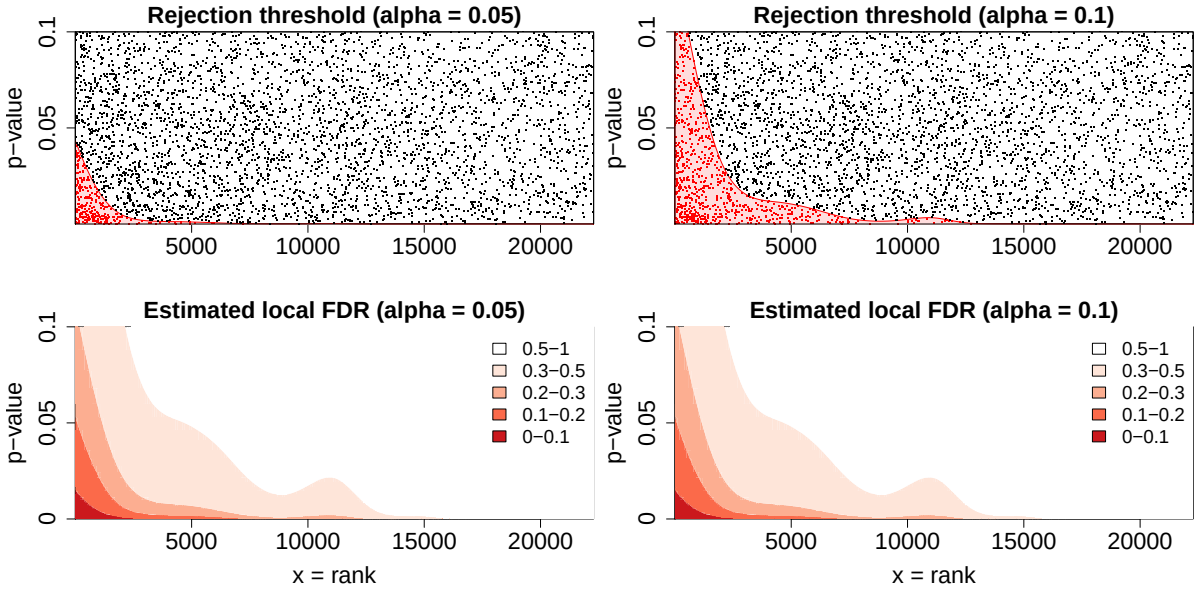


Fig. 3: Results for gene/drug response data with moderately informative ordering of p-values, i.e. $\{p_i^W\}$, with $\alpha = 0.05$ (left) and $\alpha = 0.1$ (right): (top) the dots represent the p-values and the red dots are rejected ones. The red curve is the thresholding rule $s(x)$; (bottom) the contour plots of estimated local FDR.

The right two panels of Figure 2 correspond to the weaker and the strong orderings, and show that AdaPT significantly outperforms all other methods for all target FDR levels. One might doubt whether the power gain is driven by overfitting. To check this, we also apply AdaPT, as well as all other methods, on the same set of p-values with a random ordering. We

repeat it using 100 random seeds and report the average number of rejections in the left panel of Figure 2. In this case, the number of rejections drop dramatically and the power is almost the same as Barber-Candès method, the non-adaptive version of AdaPT. This provides strong evidence against overfitting.

To illustrate how AdaPT exploits the covariate to improve the power, we plot the thresholding rules and estimated signal strength for p-values with moderately informative ordering and p-values with highly informative ordering, respectively in Figure 3 and Figure 4. It can be seen from the bottom panels that the evidence to be non-null has an obvious decreasing trend when the ordering is used. Moreover, the highly informative ordering indeed sorts the p-values better than the moderately informative ordering. For the former, the thresholding rule is fairly monotone while it has a small bump at $i \approx 5000$ for the latter. In both cases, most discoveries are from the first 5000 genes in the list.

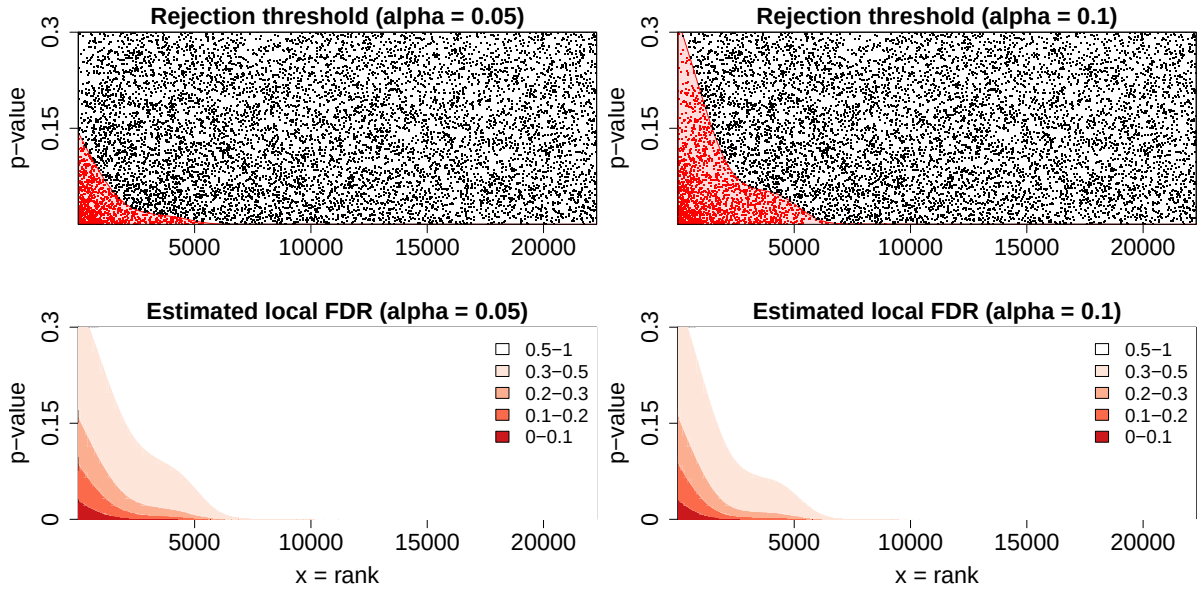


Fig. 4: Results for gene/drug response data with highly informative ordering of p-values, i.e. $\{p_i^S\}$, with $\alpha = 0.05$ (left) and $\alpha = 0.1$ (right): (top) the dots represent the p-values and the red dots are rejected ones. The red curve is the thresholding rule $s(x)$; (bottom) the contour plots of estimated local FDR.

Finally, we measure the information loss caused by partial masking: We first estimate local FDR using the set of (unmasked) p-values and the covariates, denoted by $\mathrm{lfdr}^*(x)$. It can be regarded as the best possible estimate given the algorithm. Let $\mathrm{lfdr}_t(x)$ denote the estimate of local FDR at step t (based on partially masked p-values). Then we measure the information loss by the correlation of $\{\mathrm{lfdr}^*(x_i)\}_{i=1}^n$ and $\{\mathrm{lfdr}_t(x_i)\}_{i=1}^n$. The results are shown in Figure 5 where the x-axis corresponds to the target FDR, in a reverse order ranging from 0.5 to 0.01, and y-axis corresponds to the correlation at the step where $\widehat{\mathrm{FDP}}$ first drops below the target FDR. As expected from the discussion in Subsection 1.3, the information loss is quite small and even negligible after the target FDR drops to the "practical" regime (e.g. below 0.2), where the correlation between $\{\mathrm{lfdr}^*(x_i)\}_{i=1}^n$ and $\{\mathrm{lfdr}_t(x_i)\}_{i=1}^n$ is almost 1. The pattern is even more significant in other data examples in the next Subsection. This provides a strong evidence that AdaPT allows efficient data exploration under comparatively limited information loss.

In summary, these plots show a strong data adaptivity of AdaPT , which can also learn the local structure of data while controlling FDR. Moreover, it provides a quantitative way, by estimated signal strength, to evaluate the quality of ordering, which is the major concern in ordered testing problems (Li and Barber, 2016a; Lei and Fithian, 2016; Li and Barber, 2016b).

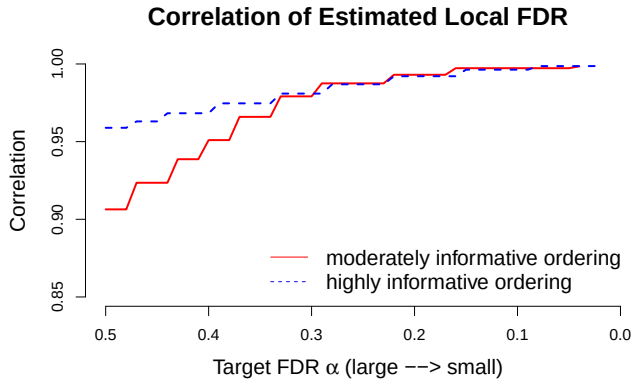


Fig. 5: Correlation of $\{\text{lfdr}^*(x_i)\}_{i=1}^n$ and $\{\text{lfdr}_t(x_i)\}_{i=1}^n$ for gene/drug-response dosage dataset under original, moderately informative and highly informative orderings. The x-axis corresponds to the target FDR, in a reverse order ranging from 0.5 to 0.01, and y-axis corresponds to the correlation at the step where $\widehat{\text{FDP}}$ first drops below the target FDR.

5.2. Simulation studies

• Example 1: a two-dimensional case

We generate the covariates x_i's from an equi-spaced 50×50 grid in the area $[-100, 100] \times [-100, 100]$. We generate p-values i.i.d. from a one-sided normal test, i.e.

$$p_i = 1 - \Phi(z_i), \quad \text{and} \quad z_i \sim N(\mu, 1), \tag{24}$$

where Φ is the cdf of $N(0, 1)$. For $i \in \mathcal{H}_0$ we set $\mu = 0$ and for $i \notin \mathcal{H}_0$ we set $\mu = 2$. Figure 6 below shows three types of $\mathcal{H}_0$ that we conduct tests on.

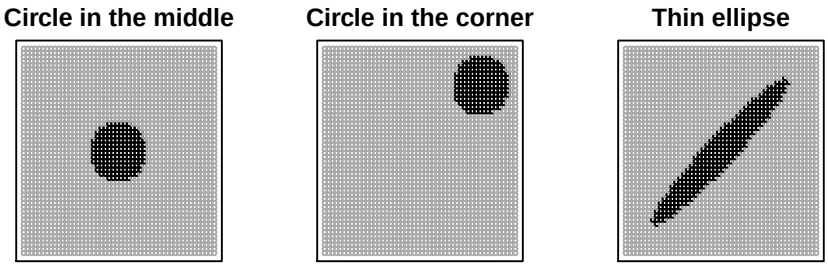


Fig. 6: The above panels display the underlying ground truth for three cases in Example 1. Each point represents a hypothesis (2500 in total) with gray ones being nulls and black ones being non-nulls.

In this case, it is not clear how to apply non-adaptive ordered testing procedures or Independent Filter. Thus we compare AdaPT only with Storey's BH method, Barber-Candés method, IHW using the default automatic parameter tuning procedure and SABHA using 2-dim low total variation weights (see Section 4.3 of Li and Barber (2016b)). For AdaPT , we fit two-dimensional Generalized Additive Models in M-step, using `R` package `mgcv` with the knots selected automatically in every step by GCV criterion. For each procedure and a given level α, let $\mathcal{R}_\alpha$ be the set of rejected hypotheses with a target FDR level α. Then we calculate the FDP and the power as

$$\text{FDP}(\alpha) = \frac{|\mathcal{R}_\alpha \cap \mathcal{H}_0|}{|\mathcal{R}_\alpha|}, \quad \text{power}(\alpha) = \frac{|\mathcal{R}_\alpha \cap \mathcal{H}_0^c|}{|\mathcal{H}_0^c|}. \tag{25}$$

We repeat the above procedure for on 100 fresh simulated datasets and calculate the average of FDP(α) and power(α) as the measure of FDR and power. The results are shown in Figure 7. It is clearly seen that AdaPT controls FDR as other methods while achieving a significantly higher power.

To see why AdaPT gains power, we plot the estimated local FDR in Figure 8 for the first case, at the initial step, the step where $\widehat{\text{FDP}}$ is first below 0.3 and the step where $\widehat{\text{FDP}}$ is first below 0.1. As shown in the real examples, the fitted local FDR identifies the non-nulls quite accurately even at the initial step where most p-values are partially-masked. The estimates become very stable and informative after reaching the practical regime of α's.

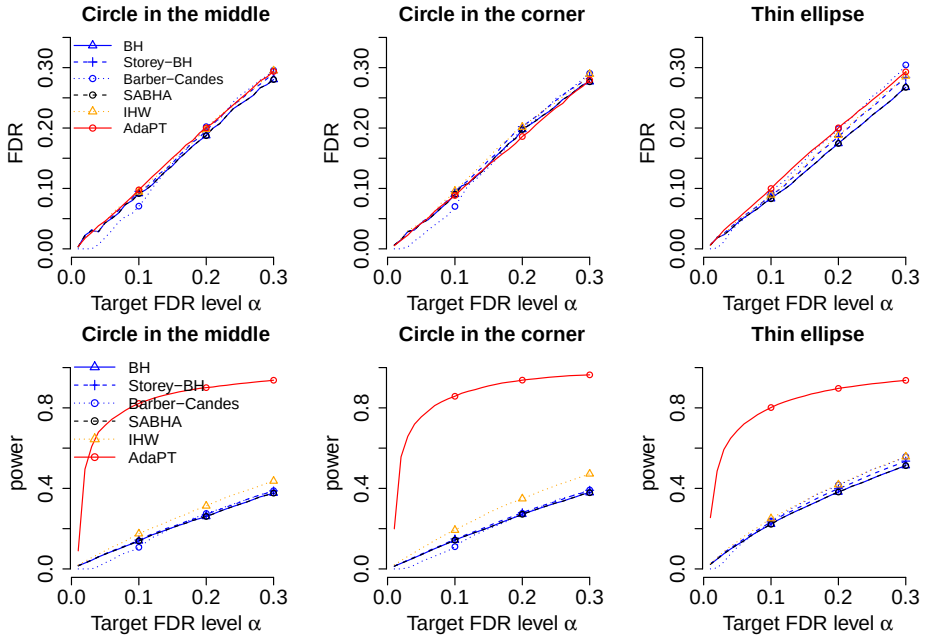


Fig. 7: FDR and power with $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$ in Example 1.

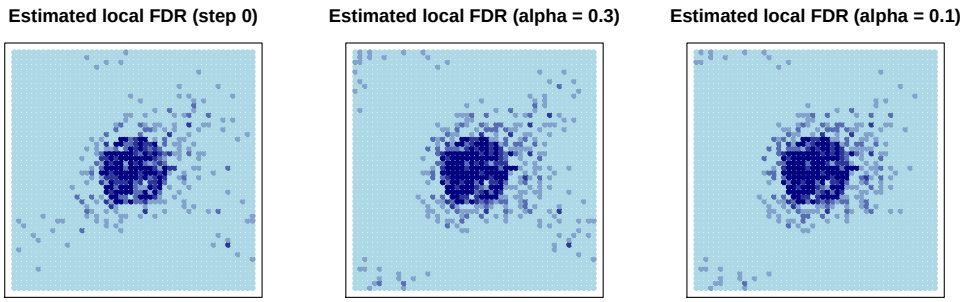


Fig. 8: Estimated local FDR in the first case of Example 1 at the initial step (left), with the target FDR level 0.3 (middle) and with the target FDR level 0.1 (right). The dark color marks the hypotheses with low local FDR and vice versa.

• Example 2: a 100-dimensional case

We generate $x_i \in \mathbb{R}^d$ with $d = 100$ and

$$\{x_{ij} : i \in [n], j \in [d]\} \overset{i.i.d.}{\sim} U([0,1]).$$

Then we generate p-values from a varying-coefficient two group beta-mixture model (15) with π_{1i} and μ_i are specified as a logistic model and a truncated linear model, respectively, i.e.,

$$\log\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) = \theta_0 + x_i^T\theta, \quad \mu_i = \max\{x_i^T\beta, 1\}, \quad \beta, \theta \in \mathbb{R}^d.$$

In this case, we choose θ and β as highly sparse vectors with only two non-zero entries:

$$\theta = (3, 3, 0, \ldots, 0)^T, \quad \beta = (2, 2, 0, \ldots, 0)^T$$

and θ_0 is chosen so that $\frac{1}{n}\sum_{i=1}^n \pi_{1i} = 0.3$. In this case, $\mathbb{E}(-\log p_i) = \mu_i$ under the alternative. Figure 9 shows the histograms of π_{1i}'s and μ_i's.

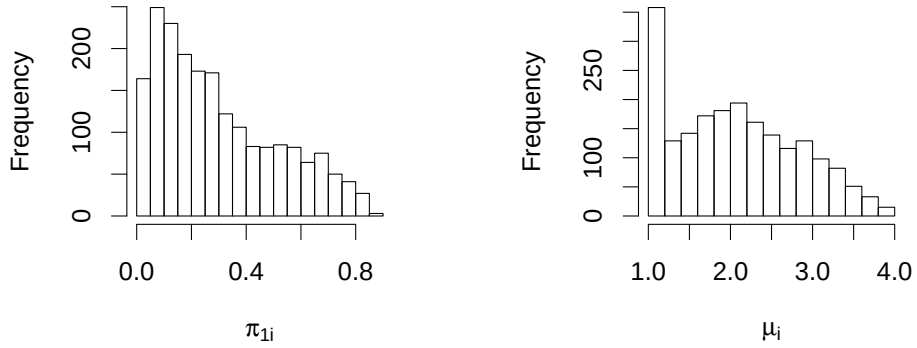


Fig. 9: Distributions of π_{1i}'s and μ_i's in Example 2.

In this case, it is not clear how to apply non-adaptive ordered testing procedures or Independent Filter or adaptive procedures like IHW and SABHA. Thus we compare AdaPT only with BH method, Storey's BH method and Barber-Candés method. For AdaPT , we fit L_1-regularized GLMs in M-step (See Appendix A for details), using `R` package `glmnet` with the penalty level selected automatically in every step by cross validation. Further we run AdaPT by fitting an "oracle" GLM in M-steps where only the first two covariates are involved.

As in Example 1, we estimate the FDR and the power using 100 replications. The results are plotted in Figure 10. It is clearly seen that both AdaPT 's control FDR as other methods while achieving a higher power. Not surprisingly, compare to AdaPT with L_1-regularized GLMs, AdaPT with "oracle" GLMs has a higher power. Nevertheless, this example shows the unprecedented ability of AdaPT to improve power by squeezing information from a large set of noisy features.

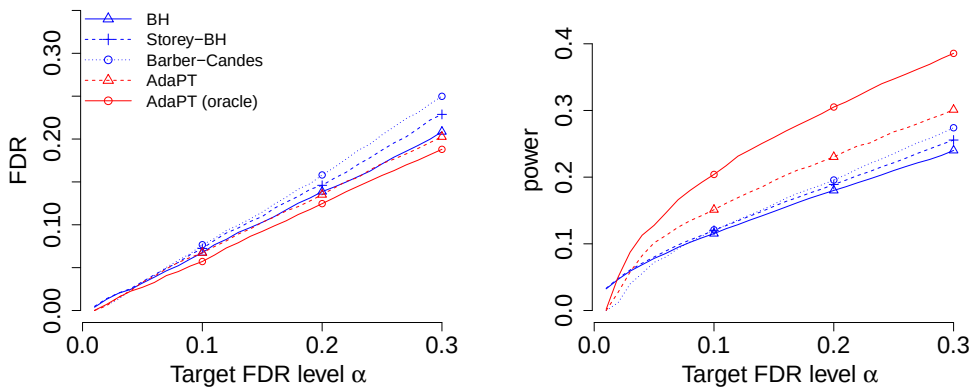


Fig. 10: FDR and power with $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$ in Example 2.

5.3. Other exemplary applications

In this Subsection, we examine the performance of AdaPT on four more real datasets, which are analyzed in other papers exploiting adaptive FDR control methods, e.g. Bourgon et al. (2010); Ignatiadis et al. (2016). In all cases, we start with a brief introduction of the dataset and show the plots on the number of rejections, as Figure 2, path of information loss, as Figure 5, and threshold curve and level curves of estimated local FDR with target FDR 0.1, as Figure 3 and Figure 4. We use the same settings for AdaPT as in the gene dosage dataset: performing model selection at the initial step with candidate featurization being all combinations of spline basis with $6 \sim 15$ equi-quantile knots on $\pi(x)$ and $\mu(x)$; and fixing the selected model in subsequent updates.

• Bottomly data

This dataset is an RNA-Seq dataset targeting on detecting the differential expression on two mouse strains, C57BL/6J (B6) and DBA/2J (D2), collected by Bottomly et al. (2011), available on ReCount repository (Frazee et al., 2011), and analyzed by Ignatiadis et al. (2016) using IHW. It consists of gene expression measurements for $n = 13932$ genes. Following Ignatiadis et al. (2016), we analyze the data using `DEseq2` package (Love et al., 2014) and use the logarithm of normalized count (averaged across all samples plus 1) as the univariate covariate for each gene. The results are plotted in Figure 11. It is clearly seen that AdaPT produces significantly more discoveries than all other methods and the information loss is almost negligible (with correlation consistently above 0.985). Furthermore, we observe the same pattern that AdaPT prioritizes the genes with higher mean normalized means.

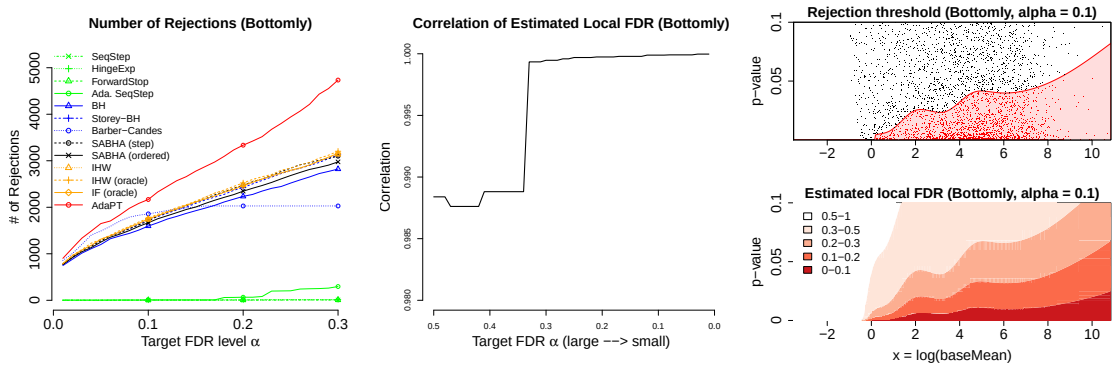


Fig. 11: Results for Bottomly dataset: (left) number of rejections; (middle) path of information loss; (right) threshold curve and level curves of estimated local FDR when $\alpha = 0.1$.

• Airway data

This dataset is an RNA-Seq dataset targeting on identifying the differentially expressed genes in airway smooth muscle cell lines in response to dexamethasone, collected by Himes et al. (2014) and available in `R` package `airway`. It is analyzed in the vignette of `IHW` package using IHW method Ignatiadis et al. (2016). As in the vignette and the previous example, we analyze the data using `DEseq2` package (Love et al., 2014) and use the logarithm of normalized count as the univariate covariate for each gene. The results are plotted in Figure 12. Again, AdaPT produces significantly more discoveries than all other methods.

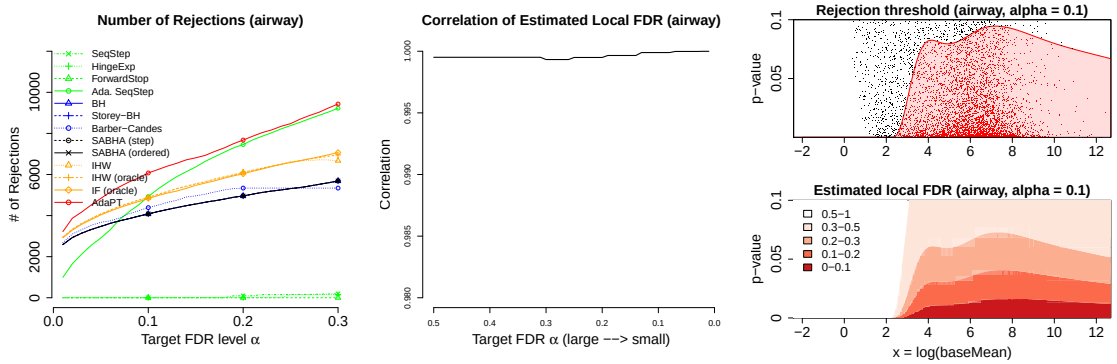


Fig. 12: Results for Airway dataset: (left) number of rejections; (middle) path of information loss; (right) threshold curve and level curves of estimated local FDR when $\alpha = 0.1$.

- **Pasilla data**

This dataset is also an RNA-Seq dataset targeting on detecting genes that are differentially expressed between the normal and Pasilla-knockdown conditions, collected by Brooks et al. (2011) and available in `R` package `pasilla` (Huber and Reyes, 2016). It is analyzed in the vignette of `genefilter` package (Gentleman et al., 2016) using independent filtering method Bourgon et al. (2010). As in the vignette, we analyze the data using `DEseq` package (Anders and Huber, 2010) and use the logarithm of normalized count as the univariate covariate for each gene. The results are plotted in Figure 13. It is clear that we arrive at the same conclusion that AdaPT is more powerful than all other methods.

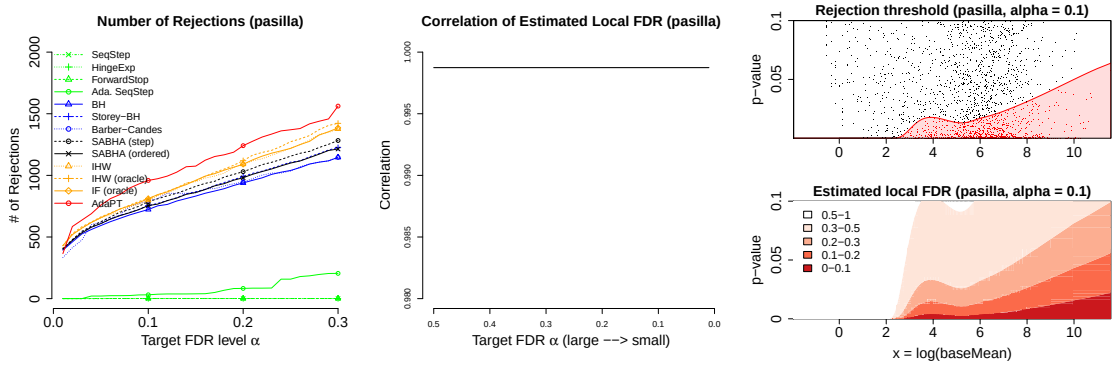


Fig. 13: Results for Pasilla dataset: (left) number of rejections; (middle) path of information loss; (right) threshold curve and level curves of estimated local FDR when $\alpha = 0.1$.

- **Yeast proteins data**

This dataset is a proteomics dataset, collected by Dephoure and Gygi (2012) and available in `R` package `IHWpaper`, that provides temporal abundance profiles for 2666 yeast proteins from a quantitative mass-spectrometry (SILAC) experiment. The goal is to identify the differential protein abundance in yeast cells treated with rapamycin and DMSO. It is analyzed in Ignatiadis et al. (2016) using IHW method. As in Dephoure and Gygi (2012) and Ignatiadis et al. (2016), we calculate the p-values using Welch's t-test and use as the univariate covariate the logarithm of total number of peptides that were quantified across all samples for each gene. The results are plotted in Figure 14. In this case, AdaPT has a similar performance to Barber-Candés method and Storey's BH method. However, it still outperforms all other methods. Furthermore, AdaPT learns the monotone pattern of the local FDR, which coincides with the heuristic.

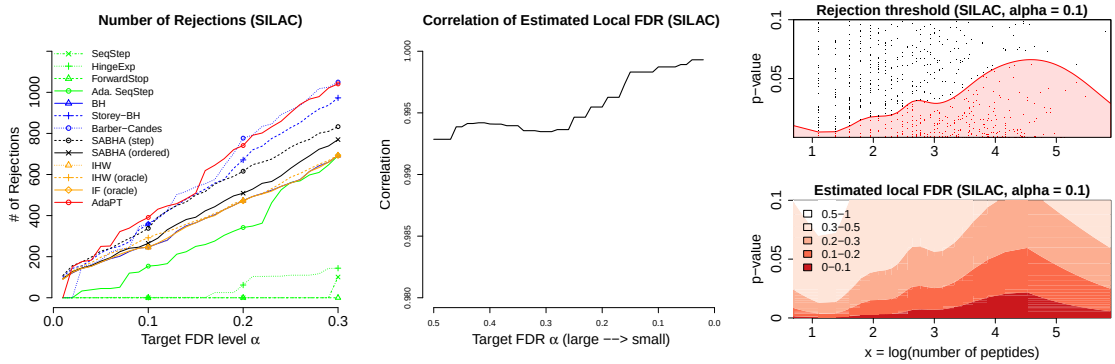


Fig. 14: Results for yeast proteins dataset: (left) number of rejections; (middle) path of information loss; (right) threshold curve and level curves of estimated local FDR when $\alpha = 0.1$.

6. Discussion

We have proposed the AdaPT procedure, a general iterative framework for multiple testing with side information. Using partially masked p-values, we estimate a family of optimal and increasingly stringent rejection thresholds, which are level surfaces of the local FDR. We then monitor an estimator of FDP to decide which threshold to use, updating our estimates as we unmask more p-values and gain more information.

Our method is interactive in that it allows the analyst to use an arbitrary method for estimating the local FDR, and to consult her intuition to change models at any iteration, even after observing most of the data. No matter what the analyst does or how badly she overfits the data, FDR is still controlled at the advertised level (though power could be adversely affected by overfitting).We show using various experiments that AdaPT can give consistently significant power improvements over current state-of-the-art methods.

6.1. AdaPT *without thresholds*

Although we state AdaPT as a procedure that interactively updates a covariate-variant threshold curve, the thresholds are not essential. In fact, Algorithm 1 can be modified as follows in the absence of $s(x)$.

Algorithm 3 AdaPT without thresholds

Input: predictors and p-values $(x_i, p_i)_{i\in[n]}$, target FDR level α.

Procedure:

1: Initialize $\mathcal{R}_0 = [n]$;
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad R_t \leftarrow \#\left\{i \in \mathcal{R}_t : p_i \le \frac{1}{2}\right\}; A_t \leftarrow \#\left\{i \in \mathcal{R}_t : p_i > \frac{1}{2}\right\}$;
4: $\quad \widehat{\mathrm{FDP}}_t \leftarrow \frac{1+A_t}{R_t \vee 1}$;
5: $\quad$ **if** $\widehat{\mathrm{FDP}}_t \le \alpha$ **then**
6: $\quad\quad$ Reject $\mathcal{R}_t$;
7: $\quad$ **end if**
8: $\quad \mathcal{R}_{t+1} \leftarrow \textsc{update}\left(\left(x_i, \tilde{p}_{t,i}\right)_{i\in[n]}, \mathcal{R}_t\right)$;
9: **end for**

Rephrasing Algorithm 3: we start from partially masking all p-values, yielding a "candidate rejection set" $\mathcal{R}_0 = [n]$, then apply arbitrary method to update $\mathcal{R}_t$ directly. The FDP estimator (line 4) is defined in an essentially identical way as Algorithm 1. It is easy to see that Algorithm 1 is a special case of Algorithm 3. Perhaps strikingly, the proof of FDR control carries through to this general case without any modification.

It is not hard to see that our implementation in Section 4 can be reformulated in a more simple and straightforward way: in each step we estimate local FDR for each partially-masked p-values and peel off δ-proportion of them with highest estimated local FDR.

In principle, we can define any "score" that measures how "promising" each hypothesis is or how "likely" each hypothesis is non-null. A simple workflow based on Algorithm 3 is to peel off the hypotheses with least favorable "scores" and proceed with refitted "scores" by exploiting the revealed p-values. Heuristically, the most statistical meaningful "score" is local FDR, which is directly associated with our purpose. However, it arguably allows the framework of AdaPT to be more general and flexible. For instance, we recently exploited this idea and develop a general framework for controlling FDR under structural constraints. We refer the readers to Lei et al. (2017) for more thoughts in this vein.

6.2. Extension to dependent data using knockoffs

It would also be interesting to attempt to relax our restriction that the p-values must be independent. In the absence of some modification, our AdaPT procedure does not control FDR in finite samples for dependent p-values. In particular, there is a danger of "overfitting" to local random effects shared by nearby hypotheses: to the AdaPT procedure, such random effects are treated as signal to discover.

It could be interesting to pursue a hybrid method using ideas from AdaPT and Knockoff+ procedures in the case where the p-values arise from regression coefficients or other multivariate Gaussian test statistics. Suppose that we observe feature matrix $X \in \mathbb{R}^{n\times d}$ and response vector $y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, and we wish to test hypotheses $H_j : \beta_j = 0$ for $j = 1, \dots, d$. The key step in Barber and Candès (2015) is to compute another matrix $\widetilde{X} \in \mathbb{R}^{n\times d}$ with $\widetilde{X}'\widetilde{X} = X'X$ and $\widetilde{X}'X = X'X - D$, for some diagonal $D \in \mathbb{R}^{d\times d}$ with positive entries; this can be done provided that $n \geq 2d$ and X has full column rank.

If we define $v = X'y$ and $\tilde{v} = \widetilde{X}'y$, then we have

$$\begin{bmatrix} v+\tilde{v} \\ v-\tilde{v} \end{bmatrix} \sim \mathcal{N}_d \left(\begin{bmatrix} (2X'X - D)\beta \\ D\beta \end{bmatrix}, \sigma^2 \begin{bmatrix} 4X'X - 2D & 0 \\ 0 & 2D \end{bmatrix} \right).$$

As a result $((v_j, \tilde{v}_j))_{j\in\mathcal{H}_0}$ are independent exchangeable pairs, conditional on $(v_j)_{j\notin\mathcal{H}_0}$. Let $\mathcal{F}_{-1} = \sigma((\{v_j, \tilde{v}_j\})_{j=1}^d)$. The knockoff filter directly uses these exchangeable pairs by constructing knockoff statistics $w(X, y) \in \mathbb{R}^d$. The sufficiency and antisymmetry conditions together imply that each $|w_j|$ is $\mathcal{F}_{-1}$-measurable and that, conditional on $\mathcal{F}_{-1}$, $b_j = 1 - \mathrm{sgn}(w_j)$ is a mirror-conservative "binary p-value:" that is $(b_j)_{j\in\mathcal{H}_0}$ are i.i.d. Bern(1/2) independently of $\mathcal{F}_{-1}$ and $(b_j)_{j\notin\mathcal{H}_0}$. Using $|w_j|$ as a "predictor" (along with any other predictors for feature j that we might have at hand) and b_j as the p-value, the AdaPT procedure is immediately applicable.

Note that $\min\{b_j, 1 - b_j\} = 0$ for every j; hence, at each step it matters only where the rejection threshold surface is above zero or not. If q_t is the tth smallest value of $(|w_j|)_{j=1}^d$, the Knockoff+ filter corresponds to using the thresholds $s_t(|w_j|) = 0.5 \cdot 1\{|w_j| \geq q_t\}$. More generally, we can use AdaPT and interactively change the threshold we use.

If σ^2 is known, we can proceed more directly by constructing z-statistics and two-tailed p-values:

$$z_j = \frac{v_j - \tilde{v}_j}{\sqrt{2d_j\sigma^2}} \sim \mathcal{N}\left(\frac{2\beta_j}{\sqrt{2d_j\sigma^2}}, 1\right); \quad p_j = 2\min\{\Phi(z_j), 1 - \Phi(z_j)\}.$$

In that case $(p_j)_{j\in\mathcal{H}_0}$ are i.i.d. uniform p-values conditional on $(p_j)_{j\notin\mathcal{H}_0}$ and $v + \tilde{v}$ (not on $\mathcal{F}_{-1}$ above). Once again, we can immediately apply AdaPT using $v + \tilde{v}$ as a "predictor." While it is not fully clear *a priori* just how we should use $v + \tilde{v}$ as a predictor, this represents an interesting avenue for future work.

6.3. Connection to knockoffs in the orthogonal design case

Focusing on the case of orthogonal design further illuminates the relationship between AdaPT and the Knockoff+ procedure. Suppose that $X \in \mathbb{R}^{n\times d}$ has orthonormal columns, and that

$d \geq 2n$. In that case Barber and Candès (2015) suggest using the knockoff matrix $\widetilde{X}$ of d more orthonormal columns which are also orthogonal to the columns of X. Then $X'y \sim \mathcal{N}_d(\beta, \sigma^2 I_d)$ while $\widetilde{X}'y \sim \mathcal{N}_d(0, \sigma^2 I_d)$, independently.

In this case, using the LASSO, forward stepwise regression, or virtually any other model selection path procedure on the design matrix $[X\widetilde{X}]$ is identical to selecting variables in decreasing order of absolute value of $|X_j'y|$ and $|\widetilde{X}_j'y|$; or equivalently, in increasing order of the two-tailed p-values $p_j = 2 - 2\Phi(|X_j'y|/\sigma)$ and $p_j^* = 2 - 2\Phi(|\widetilde{X}_j'y|/\sigma)$ (this is true whether or not σ^2 is known). As a result, if we operationalize the Knockoff+ procedure using e.g. LASSO, we would reject hypotheses H_j for which $\min\{p_j, p_j^*\}$ is small *and* $p_j < p_j^*$. By contrast, if we were to implement AdaPT with a constant threshold in each step, we would reject hypotheses H_j for which $\min\{p_j, 1-p_j\}$ is small *and* $p_j < 1 - p_j$. Hence, the pairwise exchangeability of $(p_j, 1-p_j)$ is playing the same role as the i.i.d. pair (p_j, p_j^*) in knockoffs.

The two most salient differences between AdaPT and Knockoff+ in this case are that:

(a) AdaPT allows for iterative interaction between the analyst and data, allowing the analyst to update her local FDR estimates as information accrues. By contrast, the knockoff filter as described in Barber and Candès (2015) does not allow for such interaction (though it could, and this is a potentially interesting avenue for extending knockoffs).

(b) Unlike Knockoff+, AdaPT introduces no extra randomness into the problem. This is because AdaPT uses pairwise exchangeability of p_i with the "mirror image" p-value $1 - p_i$ instead of the independent "knockoff" p-value $p_i^* \sim U[0,1]$. Thus, as a statistical procedure AdaPT respects the sufficiency principle: for any (non-randomized) choice of UPDATE subroutine, the AdaPT result is a deterministic function of the original data.

6.4. Extension: estimating local FDR

In addition to returning a list of rejections that is guaranteed to control the global FDR, most implementations of AdaPT will also return estimates, for each rejected hypothesis, of the local FDR,

$$\widehat{\mathrm{fdr}}(p_i \mid x_i) = \widehat{\mathbb{P}}(H_i \text{ is null} \mid x_i, p_i).$$

If we have reasonably high confidence in the model we have used to produce these estimates, they may provide the best summary of evidence against the individual hypothesis H_i. By contrast, the significance level for global FDR only summarizes the strength of evidence against the entire list of rejections, taken as a whole. Indeed, it is possible to construct pathological examples where $\mathrm{fdr}(p_i \mid x_i) = 1$ for some of the rejected H_i, despite controlling FDR at some level $\alpha \ll 1$. Even apart from such perversities, it will typically be the case that $\widehat{\mathrm{fdr}}(p_i \mid x_i) > \alpha$ for many of the rejected hypotheses.

Despite their more favorable interpretation, however, the local FDR estimates produced by AdaPT rely on much stronger assumptions than the global FDR control guarantee — namely, that the two-groups model, as well as our specifications for $\pi_1(x)$ and $f_1(p \mid x)$, must be correct. Instead of using the parametric estimates $\widehat{\mathrm{fdr}}(p_i \mid x_i)$, we could estimate the local FDR in a moving window of w steps of the AdaPT algorithm:

$$\widehat{\mathrm{fdp}}_{t,w} = \frac{A_t - A_{t+w}}{1 \vee (R_t - R_{t+w})}, \quad \text{or } \widehat{\mathrm{fdp}}^+_{t,w} = \frac{1 + A_t - A_{t+w}}{1 \vee (R_t - R_{t+w})}.$$

Note that if we take an infinitely large window, we obtain $\widehat{\mathrm{fdp}}^+_{t,\infty} = \widehat{\mathrm{FDP}}_t$; thus, these estimators adaptively estimate the false discovery proportion for p-values revealed in the next w steps of the algorithm, in much the same way that $\widehat{\mathrm{FDP}}_t$ estimates the false discovery proportion for *all* remaining p-values. It would be interesting to investigate, in future work, what error-control guarantees we might be able to derive by using these estimators.

Acknowledgments

The authors thank Jim Pitman, Ruth Heller, Aaditya Ramdas, and Stefan Wager for helpful discussions.

References

Allison, D. B., G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis 39*(1), 1–20.

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome biology 11*(10), R106.

Arias-Castro, E. and S. Chen (2016). Distribution-free multiple testing. *arXiv preprint arXiv:1604.07520*.

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

Barber, R. F. and E. J. Candès (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*.

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

Benjamini, Y. and Y. Hochberg (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics 24*(3), 407–418.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics 41*(2), 802–837.

Bottomly, D., N. A. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann (2011). Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using rna-seq and microarrays. *PloS one 6*(3), e17820.

Bourgon, R., R. Gentleman, and W. Huber (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences 107*(21), 9546–9551.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Brooks, A. N., L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, and B. R. Graveley (2011). Conservation of an rna regulatory map between drosophila and mammals. *Genome research 21*(2), 193–202.

Davis, S. and P. S. Meltzer (2007). GEOquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics 23*(14), 1846–1847.

Dephoure, N. and S. P. Gygi (2012). Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Science signaling 5*(217), rs2.

Dobriban, E. (2016). A general convex framework for multiple testing with prior information. *arXiv preprint arXiv:1603.05334*.

Dobriban, E., K. Fortney, S. K. Kim, and A. B. Owen (2015). Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika 102*(4), 753–766.

Dobson, A. J. and A. Barnett (2008). *An introduction to generalized linear models*. CRC press.

Du, L., C. Zhang, et al. (2014). Single-index modulated multiple testing. *The Annals of Statistics 42*(4), 1262–1311.

Dwork, C., V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth (2015). Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 117–126. ACM.

Efron, B. (2007). Size, power and false discovery rates. *The Annals of Statistics*, 1351–1377.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association 96*(456), 1151–1160.

Ferkingstad, E., A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong (2008). Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 714–735.

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Fortney, K., E. Dobriban, P. Garagnani, C. Pirazzini, D. Monti, D. Mari, G. Atzmon, N. Barzilai, C. Franceschi, A. B. Owen, et al. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genet 11*(12), e1005728.

Frazee, A. C., B. Langmead, and J. T. Leek (2011). Recount: a multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC bioinformatics 12*(1), 449.

Genovese, C. R., K. Roeder, and L. Wasserman (2006). False discovery control with p-value weighting. *Biometrika 93*(3), 509–524.

Gentleman, R., V. Carey, W. Huber, and F. Hahne (2016). *genefilter: genefilter: methods for filtering genes from high-throughput experiments*. R package version 1.54.2.

Geyer, C. J. and G. D. Meeden (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, 358–366.

G'Sell, M. G., S. Wager, A. Chouldechova, and R. Tibshirani (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(2), 423–444.

Hemerik, J. and J. Goeman (2014). Exact testing with random permutations. *arXiv preprint arXiv:1411.7565*.

Himes, B. E., X. Jiang, P. Wagner, R. Hu, Q. Wang, B. Klanderman, R. M. Whitaker, Q. Duan, J. Lasky-Su, C. Nikolos, et al. (2014). Rna-seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PloS one 9*(6), e99625.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 169–192.

Hu, J. X., H. Zhao, and H. H. Zhou (2012). False discovery rate control with groups. *Journal of the American Statistical Association 105*(491), 1215–1227.

Huber, W. and A. Reyes (2016). *pasilla: Data package with per-exon and per-gene read counts of RNA-seq samples of Pasilla knock-down by Brooks et al., Genome Research 2011.* R package version 0.12.0.

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

Ignatiadis, N., B. Klaus, J. B. Zaugg, and W. Huber (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods 13*(7), 577–580.

Lawyer, G., E. Ferkingstad, R. Nesvåg, K. Varnäs, and I. Agartz (2009). Local and covariate-modulated false discovery rates applied in neuroimaging. *NeuroImage 47*(1), 213–219.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics 44*(3), 907–927.

Lehmann, E. and J. P. Romano (2005). *Testing statistical hypotheses*. New York:. Springer.

Lei, L. and W. Fithian (2016). Power of ordered hypothesis testing. In *ICML*.

Lei, L., A. Ramdas, and W. Fithian (2017). STAR: A general interactive framework for fdr control under structural constraints. *arXiv preprint arXiv:1710.02776*.

Lewinger, J. P., D. V. Conti, J. W. Baurley, T. J. Triche, and D. C. Thomas (2007). Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic epidemiology 31*(8), 871–882.

Li, A. and R. F. Barber (2016a). Accumulation tests for FDR control in ordered hypothesis testing. *Journal of the American Statistical Association 112*(just-accepted), 1–38.

Li, A. and R. F. Barber (2016b). Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*.

Love, M. I., S. Anders, and W. Huber (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology 15*(12), 550.

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

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

Pounds, S. and S. W. Morris (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.

Slater, M. (1950). Lagrange multipliers revisited, cowles commis. Technical report, sion Discussion Paper, Mathematics.

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. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(3), 347–368.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*(16), 9440–9445.

Sun, W., B. J. Reich, T. Tony Cai, M. Guindani, and A. Schwartzman (2015). False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(1), 59–83.

Tian, X., J. Taylor, et al. (2018). Selective inference with a randomized response. *The Annals of Statistics 46*(2), 679–710.

Tukey, J. W. (1994). *The collected works of John W. Tukey: Multiple comparisons, 1948-1983*, Volume 8. Chapman & Hall/CRC.

Yekutieli, D. (2012). Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), 515–541.

Zablocki, R. W., A. J. Schork, R. A. Levine, O. A. Andreassen, A. M. Dale, and W. K. Thompson (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics 30*(15), 2098–2104.

A. EM algorithm details

A.1. Derivation of E-step

To fill in the details of the algorithm, we are left to calculate the imputed values $\hat{H}_i^{(r)}$ and $\hat{y}_i^{(r,1)}$ given the parameters θ and β. Denote π_i and μ_i by

$$\pi_i = \left(1 + e^{-\phi_\pi(x_i)'\theta}\right)^{-1}, \quad \mu_i = \zeta^{-1}\left(\beta'\phi_\mu(x_i)\right)$$

We distinguish two cases: for revealed p-values, $\tilde{p}_{t,i}$ is a singleton; for masked p-values, $\tilde{p}_{t,i}$ is a two-elements set. For clarity, we let $p'_{t,i}$ denote the minimum element of $\tilde{p}_{t,i}$, i.e. $\tilde{p}_{t,i} = p'_{t,i}$ in the former case and $\tilde{p}_{t,i} = \{p'_{t,i}, 1 - p'_{t,i}\}$ in the latter case.

- For revealed p-values,

$$\hat{H}_i^{(r)} = \mathbb{P}(H_i = 1 | p_i = \tilde{p}_{t,i}) = \frac{\pi_i \cdot h(\tilde{p}_{t,i}; \mu_i)}{\pi_i \cdot h(\tilde{p}_{t,i}; \mu_i) + 1 - \pi_i}, \tag{26}$$

  and

$$\hat{y}_i^{(r,1)} = \mathbb{E}(y_i H_i | \tilde{p}_{t,i}) / \hat{H}_i^{(r)} = y_i. \tag{27}$$

- For masked p-values,

$$\mathbb{P}(p_i = p'_{t,i} | \tilde{p}_{t,i}, H_i = 1) = \frac{h(p'_{t,i}; \mu_i)}{h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)} \tag{28}$$

  and

$$\mathbb{P}(p_i = 1 - p'_{t,i} | \tilde{p}_{t,i}, H_i = 1) = \frac{h(1 - p'_{t,i}; \mu_i)}{h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)} \tag{29}$$

  By Bayes' formula, we can also derive the conditional distribution of H_i given $\tilde{p}_{t,i}$:

$$\hat{H}_i^{(r)} = \mathbb{P}(H_i = 1 | \tilde{p}_{t,i}) = \frac{\pi_i \cdot \left(h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)\right)}{\pi_i \cdot \left(h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)\right) + 2(1 - \pi_i)}. \tag{30}$$

  As a consequence of (28) - (30),

$$\begin{aligned}
\hat{y}_i^{(r,1)} =& \frac{1}{\hat{H}_i^{(r)}} \cdot \{g(p'_{t,i}) \cdot \mathbb{P}(H_i = 1, p_i = p'_{t,i} | \tilde{p}_{t,i}) + g(1 - p'_{t,i}) \cdot \mathbb{P}(H_i = 1, p_i = 1 - p'_{t,i} | \tilde{p}_{t,i})\} \\
=& \frac{1}{\hat{H}_i^{(r)}} \cdot \frac{\pi_i \cdot \left(h(p'_{t,i}; \mu_i) \cdot g(p'_{t,i}) + h(1 - p'_{t,i}; \mu_i) \cdot g(1 - p'_{t,i})\right)}{\pi_i \cdot \left(h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)\right) + 2(1 - \pi_i)} \\
=& \frac{h(p'_{t,i}; \mu_i) \cdot g(p'_{t,i}) + h(1 - p'_{t,i}; \mu_i) \cdot g(1 - p'_{t,i})}{h(p'_{t,i}; \mu_i) + h(1 - p'_{t,i}; \mu_i)}.
\end{aligned} \tag{31}$$

A.1.1. Beta-mixture model
Consider the beta-mixture model (15), where

$$h(p;\mu) = \frac{1}{\mu} \cdot p^{\frac{1}{\mu}-1}.$$

Plug it into our general results, we obtain that

- for revealed p-values,

$$\hat{H}_i^{(r)} = \frac{\pi_i \cdot \frac{1}{\mu_i} \tilde{p}_{t,i}^{\frac{1}{\mu_i}-1}}{\pi_i \cdot \frac{1}{\mu_i} \tilde{p}_{t,i}^{\frac{1}{\mu_i}-1} + 1 - \pi_i}, \quad \hat{y}_i^{(r)} = y_i; \tag{32}$$

- for masked p-values,

$$\begin{aligned}
\hat{H}_i^{(r)} &= \frac{\pi_i \cdot \frac{1}{\mu_i} \left(p_{t,i}'^{\frac{1}{\mu_i}-1} + (1-p'_{t,i})^{\frac{1}{\mu_i}-1} \right)}{\pi_i \cdot \frac{1}{\mu_i} \left(p_{t,i}'^{\frac{1}{\mu_i}-1} + (1-p'_{t,i})^{\frac{1}{\mu_i}-1} \right) + 2(1-\pi_i)}, \\
\hat{y}_i^{(r)} &= \frac{p_{t,i}'^{\frac{1}{\mu_i}-1}(-\log p'_{t,i}) + (1-p'_{t,i})^{\frac{1}{\mu_i}-1}(-\log(1-p'_{t,i}))}{p_{t,i}'^{\frac{1}{\mu_i}-1} + (1-p'_{t,i})^{\frac{1}{\mu_i}-1}}.
\end{aligned} \tag{33}$$

A.1.2. Gaussian-mixture model
Suppose the p-values are derived from a one-sided z-test by transforming a normal random variable, the most natural transformation is $g(p) = \Phi^{-1}(1-p)$. By (13),

$$h(p;\mu) = \exp\left\{ \mu \cdot g(p) - \frac{1}{2}\mu^2 \right\}.$$

Plug it into our general results, we obtain that

- for revealed p-values,

$$\hat{H}_i^{(r)} = \mathbb{P}(H_i = 1 | p_i = \tilde{p}_{t,i}) = \frac{\pi_i \cdot e^{\mu_i y_i}}{\pi_i \cdot e^{\mu_i y_i} + (1-\pi_i) e^{\mu_i^2/2}}, \quad \hat{y}_i^{(r)} = y_i; \tag{34}$$

- for masked p-values,

$$\hat{H}_i^{(r)} = \frac{\pi_i \cdot \cosh(\mu_i y_i)}{\pi_i \cdot \cosh(\mu_i y_i) + (1-\pi_i) e^{\mu_i^2/2}}, \quad \hat{y}_i^{(r)} = |y_i| \cdot \tanh(\mu_i |y_i|). \tag{35}$$

A.2. Initialization

Another important issue is the initialization. The formulae of $\hat{H}_i^{(r)}$ and $\hat{y}_i^{(r,1)}$ requires estimates of π_{1i} and μ_i. In step 0 when no information can be obtained, we propose a simple method by imputing random guess as follows. First we obtain an initial guess of π_{1i}. Let $J_i = I(\tilde{p}_{t,i}$ contains two elements), then we observe that

$$\begin{aligned}
&\mathbb{E} J_i = \mathbb{P}(p_i \notin [s_0(x_i), 1-s_0(x_i)]) \geq (1-\pi_{1i})(1-2s_0(x_i)) \\
&\implies \pi_{1i} \geq \mathbb{E}\left(1 - \frac{J_i}{1-2s_0(x_i)}\right).
\end{aligned} \tag{36}$$

Let

$$\tilde{J}_i = 1 - \frac{J_i}{1-2s_0(x_i)} \tag{37}$$

and then we fit a logistic regression on $\tilde{J}_i$ with covariates $\phi_\pi(x_i)$, denoted by $\hat{\pi}_{1i}$ and then truncate $\hat{\pi}_{1i}$ at 0 and 1 to obtain an initial guess of π_{1i}, i.e.

$$\pi_{1i}^{(0)} = (\hat{\pi}_{1i} \vee 0) \wedge 1. \tag{38}$$

(36) implies that $\pi_{1i}^{(0)}$ is a conservative estimate of π_{1i}. This is preferred to an anti-conservative estimate since the latter might cause over-fitting.

Then we obtain an initial guess of $\mu(x_i)$ by imputing p_i's. If $p_i \in [s_0(x_i), 1 - s_0(x_i)]$, then $\tilde{p}_{t,i} = p_i$ and hence we can use it directly. Otherwise, we only know that $p_i \in \tilde{p}_{t,i} = \{p'_{t,i}, 1-p'_{t,i}\}$. If p_i is null, then it should be uniform on $\{p'_{t,i}, 1 - p'_{t,i}\}$; if p_i is non-null, then it should more likely to be $p'_{t,i}$ since $p'_{t,i} < 1 - p'_{t,i}$. Thus, we impute p_i by $p'_{t,i}$, and fit an unweighted GLM on $g(p'_{t,i})$ with covariates $\phi(x_i)$ and inverse link to obtain an initial guess of μ_i.

A.3. Other issues

In Algorithm 2, we fit $\mu(x)$ using a weighted GLM, corresponding to the the M-step. However, the weighted step is sensitive to the weights $\hat{H}_r$ derived from the E-step, which relies on the assumption that null p-values are uniformly distributed on $[0, 1]$. In practice, null p-values might be super-uniform or only asymptotically uniform. In this case, the weights generated by the E-step might lead to abnormal estimates for $\mu(x)$. For this reason, we modify the weighted step in M-steps for fitting $\mu(x)$ into an unweighted step and find that this choice leads to consistently good performance in all experiments shown in Section 5.

B. Technical Proofs

Proof (**Proof of Theorem 2**). Assume $f_0(p \mid x_i) \le M$. Let $\eta_i = \nu(\{x_i\})$ and $s = (s_1, \ldots, s_n) \triangleq (s(x_1), \ldots, s(x_n))$, then the objective function of (7)

$$\int_{\mathcal{X}} -F_1(s(x)|x)\pi_1(x)\nu(dx) = -\sum_{i=1}^{n} \eta_i F_1(s_i|x_i)\pi_1(x_i)$$

is a convex function of s by condition (i) and the constraint function

$$\begin{aligned} g(s) &\triangleq \int_{\mathcal{X}} \Big\{ -\alpha F_1(s(x)|x)\pi_1(x) + (1-\alpha)F_0(s(x)|x)(1-\pi_1(x)) \Big\}\nu(dx) \\ &= \sum_{i=1}^{n} \eta_i \left(-\alpha F_1(s_i|x_i)\pi_1(x_i) + (1-\alpha)F_0(s_i|x_i)(1-\pi_1(x_i))\right) \end{aligned}$$

is also a convex function of s by condition (i). To establish the necessity of the KKT condition, it is left to prove the Slater's condition (Slater 1950; Boyd and Vandenberghe 2004, Chap. 5), i.e. there exists a $\bar{s}$, such that for any $s \in B(\bar{s}, \delta)$ for some $\delta > 0$, the constraint inequality holds, i.e. $g(s) \le 0$, and $g(\bar{s}) < 0$. By condition (ii), WLOG we assume $\mathrm{fdr}(0|x_1) < \alpha$ with $\nu(\{x_1\}) = \eta_1 > 0$. Fix any $\epsilon > 0$ and denote by $\omega(\epsilon)$ by the maximum modulus of continuity of $f_1(p \mid x_1)$ and $f_0(p \mid x_0)$ at point 0, i.e.

$$\omega(\epsilon) = \max\left\{ \sup_{p \le \epsilon} f_1(p \mid x_1), \sup_{p \le \epsilon} f_0(p \mid x_1) \right\}.$$

By assumption (i), we know that

$$\lim_{\epsilon \to 0} \omega(\epsilon) = 0.$$

Let $\bar{s} \in \mathbb{R}^n$ with

$$\bar{s}_1 = 2\epsilon, \bar{s}_2 = \ldots = \bar{s}_n = \epsilon \cdot (2M)^{-1}\eta_1\Delta$$

where

$$\Delta = f(0|x_1)\left(\alpha(1-\mathrm{fdr}(0|x_1)) - (1-\alpha)\mathrm{fdr}(0|x_1)\right) > 0.$$

Let $\delta = \min\{\bar{s}_2, \epsilon\}$. We will show that for any $s \in B(\bar{s}, \delta)$, $g(s) < 0$. In fact, for any $s \in B(\bar{s}, \delta)$, we have

$$s_1 \in [\epsilon, 3\epsilon], \quad s_i \le \epsilon \cdot M^{-1}\eta_1\Delta.$$

Recalling that M is an upper bound for the null density. WLOG, we assume that $M \ge \eta_1\Delta$ in which case $s_i \le \epsilon$ for all $i \ge 2$. Note that $g(0) = 0$. By mean-value theorem, there exists $\tilde{s}_1 \in [0, 3\epsilon], \tilde{s}_2, \ldots, \tilde{s}_n \in [0, \epsilon]$, such that

$$\begin{aligned}
g(s) &= s_1\eta_1\left(-\alpha f_1(\tilde{s}_1|x_1)\pi_1(x_1) + (1-\alpha)f_0(\tilde{s}_1|x_1)(1-\pi_1(x_1))\right) \\
&\quad + \sum_{i=2}^{n} s_i\eta_i\left(-\alpha f_1(\tilde{s}_i|x_i)\pi_1(x_i) + (1-\alpha)f_0(\tilde{s}_i|x_i)(1-\pi_1(x_i))\right) \\
&\le s_1\eta_1\left(-\alpha f_1(0|x_1)\pi_1(x_1) + (1-\alpha)f_0(0|x_1)(1-\pi_1(x_1))\right) + s_1\eta_1\omega(\epsilon) \\
&\quad + \sum_{i=2}^{n} s_i\eta_i\left(-\alpha f_1(\tilde{s}_i|x_i)\pi_1(x_i) + (1-\alpha)f_0(\tilde{s}_i|x_i)(1-\pi_1(x_i))\right) \\
&= -s_1\eta_1\Delta + s_1\eta_1\omega(\epsilon) + \sum_{i=2}^{n} s_i\eta_i\left(-\alpha f_1(\tilde{s}_i|x_i)\pi_1(x_i) + (1-\alpha)f_0(\tilde{s}_i|x_i)(1-\pi_1(x_i))\right) \\
&\le -s_1\eta_1\Delta + s_1\eta_1\omega(\epsilon) + \sum_{i=2}^{n} s_i\eta_i \cdot M \\
&\le -\epsilon\eta_1\Delta + s_1\eta_1\omega(\epsilon) + \epsilon\eta_1\Delta\sum_{i=2}^{n}\eta_i \\
&\le -\epsilon\eta_1\Delta + O(\epsilon\omega(\epsilon)) + \epsilon\eta_1\Delta(1-\eta_1) \\
&\le -\epsilon\eta_1^2\Delta + o(\epsilon).
\end{aligned}$$

Thus, for sufficiently small ϵ, $g(s) < 0$ for all $s \in B(\bar{s}, \delta)$ and hence the Slater's condition is satisfied.

Proof (**Proof of Lemma 2**). We assume $\rho < 1$ (otherwise the result is trivial). Following Barber and Candès (2016), we introduce the random set $\mathcal{A} \subseteq [n]$ with

$$\mathbb{P}(i \in \mathcal{A} \mid \mathcal{G}_{-1}) = \frac{1-\rho_i}{1-\rho},$$

conditionally independent for $i \in [n]$, and construct conditionally i.i.d. Bernoulli variables $q_1, \ldots, q_n$, independent of $\mathcal{A}$, with $\mathbb{P}(q_i = 1 \mid \mathcal{G}_{-1}) = \rho$. Then we can define

$$\tilde{b}_i = q_i\mathbf{1}\{i \in \mathcal{A}\} + \mathbf{1}\{i \notin \mathcal{A}\}, \tag{39}$$

which by construction gives $\mathbb{P}(\tilde{b}_i = 1 \mid \mathcal{G}_{-1}) = \rho_i$ almost surely. Furthermore, noticing that

$$\begin{aligned}
&\mathbb{P}(\tilde{b}_i = 0, \tilde{b}_j = 0|\mathcal{G}_{-1}) = \mathbb{P}(i \in \mathcal{A}, j \in \mathcal{A}, q_i = 0, q_j = 0|\mathcal{G}_{-1}) \\
=&\mathbb{P}(i \in \mathcal{A}, q_i = 0|\mathcal{G}_{-1})\mathbb{P}(j \in \mathcal{A}, q_j = 0|\mathcal{G}_{-1}) \\
=&\mathbb{P}(\tilde{b}_i = 0|\mathcal{G}_{-1})\mathbb{P}(\tilde{b}_j = 0|\mathcal{G}_{-1}),
\end{aligned}$$

we conclude that the $\tilde{b}_i$ are conditionally independent given $\mathcal{G}_{-1}$. As a consequence, given $\mathcal{G}_{-1}$,

$$(\tilde{b}_1, \ldots, \tilde{b}_n) \overset{d}{=} (b_1, \ldots, b_n).$$

In the following proof, we will use (39) to represent b_i's.

To ensure that $\mathcal{C}_t$ decreases by at most a single element in each step, we introduce intermediate steps: for integers $t \geq 0$, $1 \leq i \leq n$ define

$$\mathcal{C}_{t+i/n} = \mathcal{C}_{t+1} \cup \{j \leq n - i : \ j \in \mathcal{C}_t\}.$$

Next, define the augmented filtration

$$\mathcal{G}_t^{\mathcal{A}} = \sigma\left(\mathcal{G}_{-1}, \mathcal{A}, \mathcal{C}_t, (b_i)_{i \notin \mathcal{C}_t \cap \mathcal{A}}, \sum_{i \in \mathcal{C}_t \cap \mathcal{A}} b_i\right) \supseteq \mathcal{G}_t,$$

for both integer and fractional values of t. Note $\mathcal{C}_{t+1/n}$ is measurable with respect to $\mathcal{C}_t$. In addition we define

$$U_t^{\mathcal{A}} = \sum_{i \in \mathcal{C}_t \cap \mathcal{A}} b_i, \qquad V_t^{\mathcal{A}} = \sum_{i \in \mathcal{C}_t \cap \mathcal{A}} 1 - b_i, \quad \text{and} \quad Z_t^{\mathcal{A}} = \frac{1 + |\mathcal{C}_t \cap \mathcal{A}|}{1 + U_t^{\mathcal{A}}}.$$

Recall the definition of U_t (Section 2.1) and b_i (defined above in (39)), for any t,

$$\frac{1 + |\mathcal{C}_t|}{1 + U_t} = \frac{1 + |\mathcal{C}_t \cap \mathcal{A}| + |\mathcal{C}_t \cap \mathcal{A}^c|}{1 + U_t^{\mathcal{A}} + |\mathcal{C}_t \cap \mathcal{A}^c|} \leq \frac{1 + |\mathcal{C}_t \cap \mathcal{A}|}{1 + U_t^{\mathcal{A}}} = Z_t^{\mathcal{A}}.$$

Finally, we observe that $(b_i)_{i \in \mathcal{C}_t \cap \mathcal{A}} = (q_i)_{i \in \mathcal{C}_t \cap \mathcal{A}}$ are exchangeable with respect to $\mathcal{G}_t^{\mathcal{A}}$, with the random vector distributed uniformly over configurations summing to $U_t^{\mathcal{A}}$.

There are three cases:

(i) if $\mathcal{C}_{t+1/n} \cap \mathcal{A} = \mathcal{C}_t \cap \mathcal{A}$ then

$$\mathbb{E}[Z_{t+1/n}^{\mathcal{A}} \mid \mathcal{G}_t^{\mathcal{A}}] = Z_t^{\mathcal{A}};$$

(ii) if $U_t^{\mathcal{A}} = 0$ but $\mathcal{C}_{t+1/n} \cap \mathcal{A} \subsetneq \mathcal{C}_t \cap \mathcal{A}$ then

$$Z_{t+1/n}^{\mathcal{A}} = 1 + |\mathcal{C}_{t+1/n} \cap \mathcal{A}| \leq |\mathcal{C}_t \cap \mathcal{A}| = Z_t^{\mathcal{A}} - 1 \leq Z_t^{\mathcal{A}};$$

(iii) otherwise, $U_t^{\mathcal{A}} > 0$ and $\mathcal{C}_t \setminus \mathcal{C}_{t+1/n} = \{j\}$ for some $j \in \mathcal{A}$. The exchangeability of $b_i = q_i$ implies that

$$P\left(b_j = 1 \mid \mathcal{G}_t^{\mathcal{A}}\right) = \frac{U_t^{\mathcal{A}}}{U_t^{\mathcal{A}} + V_t^{\mathcal{A}}}.$$

Then

$$\begin{aligned}\mathbb{E}\left[Z_{t+1/n}^{\mathcal{A}} \mid \mathcal{G}_t^{\mathcal{A}}\right] &= \frac{U_t^{\mathcal{A}} + V_t^{\mathcal{A}}}{1 + U_t^{\mathcal{A}}} \cdot \frac{V_t^{\mathcal{A}}}{U_t^{\mathcal{A}} + V_t^{\mathcal{A}}} + \frac{U_t^{\mathcal{A}} + V_t^{\mathcal{A}}}{U_t^{\mathcal{A}}} \cdot \frac{U_t^{\mathcal{A}}}{U_t^{\mathcal{A}} + V_t^{\mathcal{A}}} \\ &= \frac{V_t^{\mathcal{A}}}{1 + U_t^{\mathcal{A}}} + 1 \ = Z_t^{\mathcal{A}}.\end{aligned}$$

In all three cases, the conditional expectation of $Z_{t+1/n}^{\mathcal{A}}$ is smaller than $Z_t^{\mathcal{A}}$; thus, $Z_t^{\mathcal{A}}$ is a supermartingale with respect to the filtration $\mathcal{G}_t^{\mathcal{A}}$. Because $\hat{t}$ is also a stopping time with respect to the filtration $(\mathcal{G}_t^{\mathcal{A}})_{t=0,1/n,2/n,\ldots}$ (but one which can only take integer values), for any $\mathcal{A} \in [n]$, we have

$$\mathbb{E}\left[\frac{1 + |\mathcal{C}_{\hat{t}}|}{1 + \sum_{i \in \mathcal{C}_{\hat{t}}} b_i} \middle| \mathcal{G}_{-1}, \mathcal{A}\right] \leq \mathbb{E}\left[Z_{\hat{t}}^{\mathcal{A}} \mid \mathcal{G}_{-1}, \mathcal{A}\right] \leq \mathbb{E}\left[Z_0^{\mathcal{A}} \mid \mathcal{G}_{-1}, \mathcal{A}\right] = \mathbb{E}\left[Z_0^{\mathcal{A}} \mid \mathcal{G}_{-1}\right]. \tag{40}$$

Let $m = |\mathcal{C}_0|$ and assume $\mathcal{C}_0 = \{1, \ldots, m\}$ WLOG. Using the representation (39),

$$\begin{aligned}
\mathbb{E}\left[Z_0^{\mathcal{A}} \mid \mathcal{G}_{-1}\right] &= \mathbb{E}\left[\frac{1+m}{1+\sum_{i=1}^m (q_i I(i \in \mathcal{A}) + I(i \notin \mathcal{A}))} \mid \mathcal{G}_{-1}\right] \\
&\leq \mathbb{E}\left[\frac{1+m}{1+\sum_{i=1}^m q_i} \mid \mathcal{G}_{-1}\right] = \mathbb{E}\left[\frac{1+m}{1+\sum_{i=1}^m q_i}\right] \\
&= \sum_{k=0}^m \frac{1+m}{1+k} \cdot \binom{m}{k} \rho^k (1-\rho)^{m-k} \\
&= \sum_{k=0}^m \binom{m+1}{k+1} \rho^k (1-\rho)^{m-k} \\
&= \rho^{-1} \cdot \sum_{k=0}^m \binom{m+1}{k+1} \rho^{k+1} (1-\rho)^{m+1-(k+1)} \\
&= \rho^{-1}\left(1 - (1-\rho)^{m+1}\right) \leq \rho^{-1}.
\end{aligned}$$

Marginalizing over $\mathcal{A}$ in (40), we obtain the result.

B.1. Mirror-conservatism

In this subsection we provide two important examples that produce mirror-conservative p-values. The first example is the permutation test (e.g. Hoeffding (1952)). Typically we assume that under the null hypothesis, the test statistic $T(X)$, where X is a short-handed notation for observed data, is invariant in distribution under a finite group of transformations $\mathcal{G}$, i.e.

$$T(X) \overset{d}{=} T(gX), \forall g \in \mathcal{G}.$$

When $|\mathcal{G}|$ is small, one can compute a discrete p-value by $R/|\mathcal{G}|$, where R is the rank of $T(X)$ in the set $\{T(gX) : g \in \mathcal{G}\}$. When $|\mathcal{G}|$ is large, Hemerik and Goeman (2014) proposes sampling a subset $\mathcal{G}' = \{g_1, g_2, \ldots, g_m\}$ with $g_1 = \mathrm{Id}$ and $g_2, \ldots, g_m$ being a simple random sample (without replacement) from $\mathcal{G}$ and calculate the p-value based on $\mathcal{G}'$. In both cases, it can be proved that the p-value is uniformly distributed on an equi-spaced grid $\{\frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m}\}$ under the null hypothesis, where m is the number of replicates. Then for any $0 < a_1 \leq a_2 \leq \frac{1}{2}$,

$$P(p \in [a_1, a_2]) = \lceil m a_2 \rceil - (\lfloor m a_1 \rfloor - 1)_+$$

where $(u)_+$ denotes $\max\{u, 0\}$, and

$$P(p \in [1-a_2, 1-a_1]) = \lceil m(1-a_1) \rceil - \lfloor m(1-a_2) \rfloor + 1 = \lceil m a_2 \rceil - (\lfloor m a_1 \rfloor - 1).$$

As a result we conclude that

$$P(p \in [a_1, a_2]) \leq P(p \in [1-a_2, 1-a_1])$$

and hence p is mirror-conservative.

The second example is the one-sided test for distributions with monotone likelihood ratio, which is ubiquitous in practice. Specifically, let θ be the univariate parameter of interest and $p_\theta(x)$ be a family of densities with respect to some carrier measure μ. $p_\theta(x)$ is said to have monotone likelihood ratio with respect to some real-value function $T(x)$ if for any $\theta < \theta'$, $p_\theta \not\equiv p_{\theta'}$ and the ratio $p_{\theta'}(x)/p_\theta(x)$ is a nondecreasing function of $T(x)$. For testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, it is well-known that there exists a Uniformly Most Powerful (UMP) test (Lehmann and Romano, 2005), with the following decision function:

$$\phi(x) = \begin{cases} 1 & (T(x) > C) \\ \gamma & (T(x) = C) \\ 0 & (T(x) < C) \end{cases}$$

where (γ, C) is the solution of

$$P_{\theta_0}(T(X) > C) + \gamma P_{\theta_0}(T(X) = C) = \alpha.$$

Write $T(X)$ as T and $P_{\theta_0}(T(X) \geq t)$ as $G_0(t)$ for short. Then the induced p-value can be written as

$$p = G_0(T^+) + U(G_0(T) - G_0(T^+)), \quad U \sim U([0,1]), \tag{41}$$

where $G_0(t^+) = \lim_{t \downarrow t^+} G_0(t)$. (41) is termed as fuzzy p-values by Geyer and Meeden (2005).

PROPOSITION 1. *Let $p_\theta(x)$ be a family of densities (w.r.t the carrier measure μ) that has monotone likelihood ratio w.r.t. $T(x)$. Then the p-value defined in* (41) *is mirror-conservative.*

PROOF. Since $p_\theta(x)$ has monotone likelihood ratio, there exists a non-decreasing function $g_\theta(t)$ for each $\theta \leq \theta_0$, such that

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = g_\theta(T(x)).$$

Let ν be a measure such that for any event $A \subset \mathbb{R}$,

$$\nu(A) = \int I(T(x) \in A) \cdot p_{\theta_0}(x) \mu(dx).$$

Then for any event $A \subset \mathbb{R}$,

$$P_\theta(T(X) \in A) = \int g_\theta(T(x)) \cdot I(T(x) \in A) \cdot p_{\theta_0}(x) d\mu = \int g_\theta(t) \nu(dt). \tag{42}$$

Note that the above argument can be easily proved by standard approximation argument in measure theory that starts from indicator functions $g_\theta(t) = I(t \in A')$, extends the result to simple step functions and finally pushes it to the limit. Let ω be the product measure of ν and the Lebesgue measure on $[0,1]$. Then for any event $B \subset \mathbb{R}^2$,

$$P_\theta((T(X), U) \in B) = \int_B g_\theta(t) \omega(dt, du). \tag{43}$$

Note that $g_{\theta_0}(t) \equiv 1$ by definition. This implies that

$$\omega(B) = P_{\theta_0}((T(X), U) \in B). \tag{44}$$

Let $H(\cdot, \cdot)$ be the transformation such that $p = H(T, U)$. Then for any z,

$$\{(t,u) : G_0(t) < z\} \subset H^{-1}([0,z)) \subset H^{-1}([0,z]) \subset \{(t,u) : G_0(t) \leq z\}.$$

As a result, for any $0 < z_1 < z_2 < 1$,

$$H^{-1}([0,z_1]) \subset \{(t,u) : G_0(t) \leq z_1\}, \quad H^{-1}([z_2,1]) \subset \{(t,u) : G_0(t) \geq z_2\},$$

and hence there exists $t(z_1, z_2)$ such that

$$t_1 \leq t(z_1, z_2) \leq t_2, \quad \forall (t_1, u_1) \in H^{-1}([0,z_1]), (t_2, u_2) \in H^{-1}([z_2,1]). \tag{45}$$

Given $0 \leq a_1 \leq a_2 < 0.5$, let $A_1 = [a_1, a_2]$ and $A_2 = [1 - a_2, 1 - a_1]$. Then (45) and (43), together with the monotonicity of g_θ, imply that

$$P_\theta(p \in A_1) = \int_{H^{-1}(A_1)} g_\theta(t) \omega(dt, du) \leq \omega(H^{-1}(A_1)) \cdot g_\theta(t(a_2, 1 - a_2)),$$

and

$$P_\theta(p \in A_2) = \int_{H^{-1}(A_1)} g_\theta(t) \omega(dt, du) \geq \omega(H^{-1}(A_2)) \cdot g_\theta(t(a_2, 1 - a_2)).$$

Recalling (44), we obtain that

$$\frac{P_\theta(p \in A_1)}{P_\theta(p \in A_2)} \leq \frac{P_{\theta_0}(p \in A_1)}{P_{\theta_0}(p \in A_2)}. \tag{46}$$

It is left to prove that $\frac{P_{\theta_0}(p \in A_1)}{P_{\theta_0}(p \in A_2)} = 1$. In fact, we can prove that

$$p \sim U([0,1]) \quad \text{when } \theta = \theta_0. \tag{47}$$

Fix any $z \in (0,1)$, let

$$G_0^{-1}(z) = \sup\{t : G_0(t) \geq z\}.$$

For clarity we write u for $G_0^{-1}(z)$. Now we prove (47) in two cases:

- if u is a continuity point of G_0, i.e.

  $$G_0(u) = G_0\left(u^+\right).$$

  Since G_0 is left-continuous, we must have

  $$G_0(u) = G_0\left(u^+\right) = z.$$

  Then

  $$P_{\theta_0}(p \leq z) = P_{\theta_0}(T(X) \geq u) = G_0(u).$$

- if u is an atom of G_0, i.e.

  $$G_0(u) > G_0\left(u^+\right).$$

  By definition,

  $$G_0(u) \geq z \quad \text{and} \quad z \geq G_0\left(u^+\right).$$

  Then

  $$\begin{aligned} P_{\theta_0}(p \leq z) &= P_{\theta_0}\left(T(X) > u^+\right) + P_{\theta_0}(T(X) = u) \cdot \frac{z - G_0(u^+)}{G_0(u) - G_0(u^+)} \\ &= G_0(u^+) + (G_0(u) - G_0(u^+)) \cdot \frac{z - G_0(u^+)}{G_0(u) - G_0(u^+)} \\ &= z. \end{aligned}$$

Therefore we prove (47). By (46), we conclude that for any $\theta \leq \theta_0$,

$$\frac{P_\theta(p \in A_1)}{P_\theta(p \in A_2)} \leq 1$$

which implies the mirror-conservativeness.