ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 4, *pp.* 649–679

AdaPT: an interactive procedure for multiple testing with side information

Lihua Lei and William Fithian

University of California, Berkeley, USA

[Received January 2017. Revised March 2018]

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, the adaptive p-value thresholding procedure which we call AdaPT, which adaptively estimates a Bayes optimal p-value rejection threshold and controls the false discovery rate 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 below the threshold and proposes another threshold, until the estimated false discovery proportion 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 favourable performance of AdaPT by comparing it with state of the art methods in five real applications and two simulation studies.

Keywords: Adaptive inference; False discovery rate; Martingales; Multiple testing; p-value weighting; Selective inference

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 that we may not have expected to find. A central challenge in modern statistics is to provide scientists with methods that are sufficiently flexible to enable 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) and Fithian *et al.* (2014)), familywise error rate (e.g. Tukey (1994) and Berk *et al.* (2013)) or false discovery rate (FDR) (e.g. Benjamini and Hochberg (1995) and Barber and Candès (2015)). However, most selective inference methods require that the selection algorithm be specified in

Address for correspondence: Lihua Lei, Department of Statistics, University of California at Berkeley, 397 Evans Hall, Berkeley, CA 94720, USA.
E-mail: lihua.lei@berkeley.edu

 1369–7412/18/80649

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 predefined 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 and Taylor, 2018) to control the analyst's access to the data at the time that 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 that are smaller than a currently proposed rejection threshold, and we guarantee finite sample FDR control by applying a version of the optional stopping argument that was pioneered by Storey *et al.* (2004) and extended in Barber and Candès (2015, 2016), Grazier G'Sell *et al.* (2016), Li and Barber (2017) and Lei and Fithian (2016).

1.2. Multiple testing with side information

In many areas of modern applied statistics, from genetics and neuroimaging to on-line advertising and finance, researchers routinely test thousands or millions of 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 FDR, which is 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 by 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.*, 2010), or exploiting a prior ordering to focus power on more 'promising' hypotheses near the top of the ordering (Barber and Candès, 2015; Grazier G'Sell *et al.*, 2016; Li and Barber, 2017; 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 (2017) tested 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' by using auxiliary data collected at larger dosages. Multiple-testing procedures that exploit the ordering can reject hundreds of hypotheses whereas 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 that are associated with longevity, leading to novel discoveries. Inspired by examples like this, Ignatiadis *et al*. (2016) and Li and Barber (2016) 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 (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 for example Bourgon *et al*. (2010) and Ignatiadis *et al*. (2016).

1.3. AdaPT: a framework for false discovery rate control

This paper presents a new framework for FDR control with generic side information, which we call AdaPT for *adaptive p-value thresholding*. Our method proceeds iteratively: at each step $t = 0, 1, \ldots$, the analyst proposes a rejection threshold $s_t(x)$ and computes an estimator $\widehat{\mathrm{FDP}}_t$ for the false discovery proportion for this threshold. If $\widehat{\mathrm{FDP}}_t \leqslant \alpha$, she stops and rejects every H_i for which $p_i \leqslant s_t(x_i)$. 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 that $a(x) \leqslant b(x)$ for all $x \in \mathcal{X}$.

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

$$R_t = |\{i : \ p_i \leqslant s_t(x_i)\}|,$$
$$A_t = |\{i : \ p_i \geqslant 1 - s_t(x_i)\}|$$

and

$$\widehat{\mathrm{FDP}}_t = \frac{1 + A_t}{R_t \vee 1}.$$

The estimate $\widehat{\mathrm{FDP}}_t$ was also used by Lei and Fithian (2016) and Arias-Castro and Chen (2017). Fig. 1(a) illustrates the way that $s_t(x)$ and $1 - s_t(x)$ partition the data into three regions; A_t is the number of points in the upper region and R_t is the number in the lower region.

At each step t, the analyst can choose the next threshold $s_{t+1}(x)$ however she wishes, with only two constraints. First, $s_{t+1} \preceq s_t$ as stated before. Second, the large and small p-values (the p-values 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

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

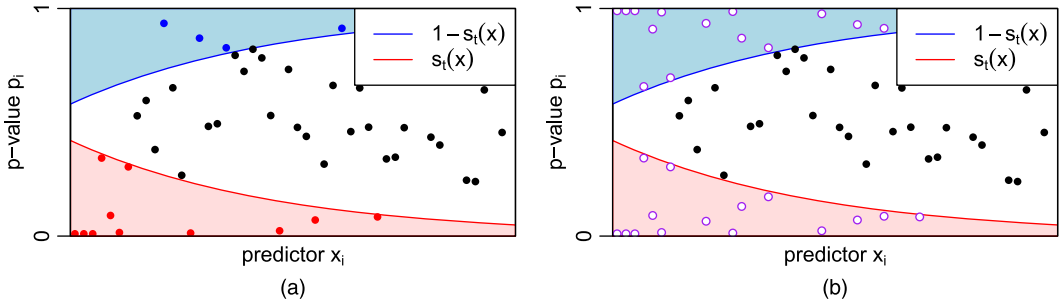

Fig. 1. Illustration of one step of AdaPT with a univariate predictor: (a) $A_t = 4$ and $R_t = 11$ are the numbers of points in the upper region and lower region respectively, leading to $\widehat{\text{FDP}} = (1+4)/11 \approx 0.45$ (if $\widehat{\text{FDP}} \leqslant \alpha$, we stop and reject the points in the lower region; 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 masked point is reflected across $p = 0.5$, leaving the analyst to impute which are the true p-values and which are the mirror images)

Thus, if $p_i = 0.01 \leqslant s_t(x_i)$ 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. Fig. 1(b) illustrates what the analyst can see: each masked point from Fig. 1(a) 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* FDR, fdr, as a function of x and p:

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

Formally, $\text{fdr}(p \mid x)$ 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 the 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 that 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 understand fully, her update rule for choosing $s_{t+1}(x)$. In this sense, we say that our method is fully *interactive*—the analyst's behaviour is arbitrary as long as she abides by a certain protocol for interacting with the algorithm.

Whereas 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 that the algorithm is close to stopping, the vast majority of p-values have already been revealed, and many of those 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.

AdaPT controls the 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 conditionally 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 in Table 1 summarizes AdaPT's procedure, using the generic subroutine 'update' to represent whatever process the analyst uses to select $s_{t+1}(x)$. Note that $s_{t+1}(x)$ is a random

Table 1. Algorithm 1: AdaPT

Input: predictors and p-values $(x_i, p_i)_{i\in[n]}$, initialization s_0 and target FDR level α
Procedure:
1, *for* $t = 0, 1, \ldots,$ *do*
2, $\quad \widehat{\mathrm{FDP}}_t \leftarrow (1 + A_t)/(R_t \vee 1)$
3, $\quad$ *if* $\widehat{\mathrm{FDP}}_t \leqslant \alpha$ *then*
4, $\quad\quad$ reject $\{H_i : p_i \leqslant s_t(x_i)\}$
5, $\quad\quad$ return s_t
6, $\quad$ *end if*
7, $\quad s_{t+1} \leftarrow \mathrm{update}\{(x_i, \tilde{p}_{t,i})_{i\in[n]}, A_t, R_t, s_t\}$
8, *end for*

function that is measurable to $\mathcal{F}_t$. Sections 3 and 4 discuss recommendations for a good update routine. It is worth mentioning that AdaPT reduces to the Barber–Candès method, inspired by Barber and Candès (2015) and proposed by Arias-Castro and Chen (2017), when $s_t(x)$ is a constant function for every t.

1.4. Related work

In recent work Ignatiadis *et al.* (2016) proposed the 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 AdaPT since it attempts to find optimal weights, but it is a little 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 guaranteed to control the FDR only asymptotically, as the number of bins stays fixed and the number of hypotheses in each bin grows to ∞. As a result, we must trust that n is sufficiently large 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 was proposed by Du and Zhang (2014) when the covariate is an auxiliary univariate p-value derived by prior information. However, similarly to Ignatiadis *et al.* (2016), it controls the FDR only 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 hypothesis.

Perhaps the procedure that is most closely related to ours is the *structure-adaptive BH algorithm* SABHA of Li and Barber (2016). 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, and 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 note 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 control the finite sample FDR provably by 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 shall 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 in multiple-testing problems; see for example Lewinger *et al.* (2007), Ferkingstad *et al.* (2008), Lawyer *et al.* (2009) and 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 model proposed deviates from the ground truth. By contrast, our method (and IHW and SABHA) rely only on validity of p-values (see the assumptions of theorem 1 in Section 2) 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 the 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 AdaPT's power on five real data sets and two simulated data sets, 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 at `https://github.com/lihualei71/adaptMT`, and the data and programs can be obtained also from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. AdaPT

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 shall 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 that we shall define shortly.

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

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

We similarly define an initial σ-field with all p-values masked: $\mathcal{F}_{-1} = \sigma[(x_i, \{p_i, 1 - p_i\})_{i=1}^n]$. 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 $\mathcal{F}_t$ grows only from t to $t+1$.

Lemma 1. For all $t \geqslant -1$, $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$.

Proof. We use induction on u to show that $\mathcal{F}_u \subseteq \mathcal{F}_t$ for any $u \leqslant t$. 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 p-values).

For $u \geqslant 0$, 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,

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

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 $\hat{t} \leqslant n$ 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 $\mathbb{P}_{H_i}(p_i \leqslant a) \leqslant a$ for all $0 \leqslant a \leqslant 1$—is *not* enough to guarantee FDR control. Instead, we say that a p-value p_i is *mirror conservative* if

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

where $[a_1, a_2]$ denotes the closed interval with end points a_1 and a_2. If p_i is discrete, condition (3) means that $p_i = 1 - a$ is at least as likely as $p_i = a$ for $a \leqslant 0.5$; if p_i has a continuous density, it means that 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 $B \sim \text{Bernoulli}(0.9)$); nor 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 the on-line appendix B.1 for proofs of these claims.

2.2. False discovery rate control

We are now ready to prove our main result: AdaPT controls the FDR in finite samples. The proof relies on a similar optional stopping argument to that 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* $p_i \leqslant s_t(x_i)$ and *null* $p_i \geqslant 1 - s_t(x_i)$ respectively. If the null p-values are uniform then, no matter how we choose $s_t(x)$ at each step, we shall always have $V_t \approx U_t$ and $\widehat{\text{FDP}}_t > U_t/(R_t \vee 1) \approx V_t/(R_t \vee 1)$.

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

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

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

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

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

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 AdaPT controls the FDR at level α, conditionally on $\mathcal{F}_{-1}$ and also marginally.

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

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

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

Let $m_i = \min\{p_i, 1-p_i\}$ and $b_i = \mathbf{1}\{p_i \geqslant 0.5\}$, 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,

$$U_t = \sum_{i \in \mathcal{C}_t} b_i$$

and

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

Further, define the σ-fields

$$\mathcal{G}_{-1} = \sigma\{(x_i, m_i)_{i=1}^n, (b_i)_{i \notin \mathcal{H}_0}\}$$

and

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

The assumptions of independence and mirror conservatism guarantee that $\mathbb{P}(b_i = 1 \mid \mathcal{G}_{-1}) \geqslant 0.5$ 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

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

and $R_t \in \mathcal{G}_t$ by a similar argument. It follows that $\hat{t} = \min\{t : \widehat{\mathrm{FDP}}_t \leqslant \alpha\}$ 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, conditionally on $\mathcal{G}_{-1}$, we can apply lemma 2 to obtain

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

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 and 6.3 further investigate the connection to knockoffs.

3. Guideline to choose thresholding rules

Although AdaPT controls the 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 updating, we consider a conditional two-groups model as the *working model*. As we shall see, under mild conditions, the Bayes optimal rejection thresholds are the level surfaces of the local FDR, fdr, defined as the probability that a hypothesis is null conditionally on its p-value. The local FDR was first discussed by Efron *et al.* (2001); see also Efron (2007). A similar result was 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 that

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

In addition, we assume that (x_i, H_i, p_i) are independent for $i \in [n]$. Unless otherwise stated we shall 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 FDR

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

We never observe H_i directly. Thus, although 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 $X \sim \nu$. Similarly to Sun *et al.* (2015), for any thresholding rule $s(x)$, we define the global FDR as

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

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

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

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

$$\max_s \mathrm{Pow}(s;\nu) \qquad \text{subject to } \mathrm{FDR}(s;\nu) \leqslant \alpha. \tag{4}$$

Next, define

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

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

$$\max_s \frac{Q_1(s)}{\mathbb{P}(H=1)} \qquad \text{subject to} \frac{Q_0(s)}{Q_0(s)+Q_1(s)} \leqslant \alpha \tag{5}$$

$$\Leftrightarrow \min_s -Q_1(s) \qquad \text{subject to} -\alpha Q_1(s) + (1-\alpha)Q_0(s) \leqslant 0 \tag{6}$$

$$\Leftrightarrow \min_s \int_{\mathcal{X}} -F_1\{s(x)\,|\,x\}\pi_1(x)\nu(\mathrm{d}x)$$
$$\text{subject to} \int_{\mathcal{X}} [-\alpha F_1\{s(x)\,|\,x\}\pi_1(x) + (1-\alpha)F_0\{s(x)\,|\,x\}\{1-\pi_1(x)\}]\nu(\mathrm{d}x) \leqslant 0. \tag{7}$$

The corresponding Lagrangian function can be written as

$$L(s;\lambda) = \int_{\mathcal{X}} [-(1+\lambda\alpha)F_1\{s(x)\,|\,x\}\pi_1(x) + \lambda(1-\alpha)F_0\{s(x)\,|\,x\}\{1-\pi_1(x)\}]\nu(\mathrm{d}x). \tag{8}$$

Let s^* be the optimum; then the Karush–Kuhn–Tucker condition (under regularity conditions) implies that

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

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) $f_1(p\,|\,x_i)$ is continuously non-increasing and $f_0(p\,|\,x_i)$ is continuously non-decreasing and uniformly bounded away from ∞ and
(b) ν is a discrete measure supported on $\{x_1,\ldots,x_n\}$ with $\nu(\{x_i : \mathrm{fdr}(0\,|\,x_i) < \alpha, f(0\,|\,x_i) > 0\}) > 0$.

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

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 the 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 that are associated with these values are of interest. We believe that it can be relaxed to more general measures and

we shall not discuss it because of 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 can control the FDR at α. As mentioned above, we can set s as the level surfaces of $\widehat{\text{fdr}}(p \mid x) = \hat{f}(1 \mid x)/\hat{f}(p \mid x)$ 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 $\text{fdr}(p \mid x)$, 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 favourably when no domain-specific expertise can be brought to bear.

More generally, we should model the data by 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 the local FDR, AdaPT nevertheless controls the global FDR. Thus, there is every reason to be relatively aggressive in choosing a modelling strategy.

4.1. Conditional density estimation via the expectation–maximization algorithm

Generically, we can model the conditional density by a parametric family where we assume that 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:

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

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

$$\begin{aligned} y_i &= g(p_i), \\ \mu_i &= B'(\eta_i). \end{aligned} \tag{11}$$

Using the standard argument, expression (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 $B''(\eta) = \text{var}_\eta(y_i) > 0$ and B' is bijective. Then there is a one-to-one mapping from μ_i to η_i, denoted by $\eta_i = \eta(\mu_i)$ by convention. In fact, $\eta(\cdot) = (B')^{-1}(\cdot)$. Then expression (10) can be reparameterized by using μ_i:

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

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

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

$$H_i \mid x_i \sim \text{Bernoulli}(\pi_{1i}), \qquad \log\left(\frac{\pi_{1i}}{1 - \pi_{1i}}\right) = \theta' \phi_\pi(x_i),$$

and

$$p_i|x_i, H_i \sim \begin{cases} h(p;\mu_i) & \text{if } H_i = 1, \\ 1 & \text{if } H_i = 0, \end{cases} \quad \text{with } \zeta(\mu_i) = \beta'\phi_\mu(x_i). \tag{14}$$

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

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

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

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

$$\begin{aligned} l(\theta, \beta; p, H, x) = &\sum_{i=1}^{n} (H_i \theta' \phi_\pi(x_i) - \log[1 + \exp\{-\theta'\phi_\pi(x_i)\}]) \\ &+ \sum_{i=1}^{n} H_i [y_i \eta \circ \zeta^{-1}\{\beta'\phi_\mu(x_i)\} - A \circ \zeta^{-1}\{\beta'\phi_\mu(x_i)\}]. \end{aligned} \tag{16}$$

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 shall proceed as though A_t and R_t are missing, so that the (y_i, H_i) pairs are mutually independent given the predictors, i.e. 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 $b_i = I(p_i \geqslant 0.5)$. 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 that t is fixed for the duration of the EM algorithm). For the E-step we compute the expectation of the log-likelihood,

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

which amounts to computing

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

and

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

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

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

The optimization above splits into two separate optimization problems: a logistic regression with predictors $\phi_\pi(x_i)$ and fractional responses $\hat{H}_i^{(r)}$ and a GLM with predictors $\phi_\mu(x_i)$, responses $\hat{y}_i^{(r,1)}$ and weights $\hat{H}_i^{(r)}$. Each of these GLM problems can be solved efficiently by using

Table 2. 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 and initialization $\hat{\theta}^{(0)}, \hat{\beta}^{(0)}$
  for $r = 1, 2, \ldots, m$ *do*
    (*E-step*)
      $\hat{H}_i^{(r)} \leftarrow \mathbb{E}_{\hat{\theta}^{(r-1)},\hat{\beta}^{(r-1)}}[H_i \mid D_t], \qquad i \in [n],$
      $\hat{y}_i^{(r,1)} \leftarrow \mathbb{E}_{\hat{\theta}^{(r-1)},\hat{\beta}^{(r-1)}}[y_i \mid D_t, H_i = 1]/\hat{H}_i^{(r)}$
    (*M-step*)
      $\hat{\theta}^{(r)} \leftarrow \mathrm{glm}(\hat{H}^{(r)} \sim \phi_\pi(x),\ \mathrm{family} = \mathrm{binomial})$
      $\hat{\beta}^{(r)} \leftarrow \mathrm{glm}(\hat{y}^{(r,1)} \sim \phi_\mu(x),\ \mathrm{family} = ...(\mathrm{link} = \zeta),\ \mathrm{weights} = \hat{H}^{(r)})$
  end for
Output: $\hat{\pi}_1(x) = [1 + \exp\{-\phi_\pi(x)'\hat{\theta}^{(m)}\}]^{-1}$, $\hat{\mu}(x) = \zeta^{-1}\{\phi_\mu(x)'\hat{\beta}^{(m)}\}$

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 by using R pseudocode. The family argument for estimating $\hat{\beta}^{(r)}$ depends on the form of exponential family (13). For example, equation (19) yields a gamma GLM in the beta mixture model (15).

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

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

one can even fit them directly by using any non-parametric method, such as random forests or neural networks, that targets estimating the conditional mean.

4.2. Selecting featurization

Suppose that 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 equispaced knots; for multivariate x, $\phi_{\pi,j}(x)$ and $\phi_{\mu,j}(x)$ could be subsets of covariates contained in x. At step t, we are permitted to fit a model for each featurization, using arbitrary methods (e.g. a GLM or a penalized GLM,), 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 that H_i is known, for the GLM (14) based on $(\phi_{\pi,j}(x), \phi_{\mu,j}(x))$ can be written as

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

Though l_j is not computable, we can replace it by

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

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

On the basis of $\{\tilde{l}_j\}_{j=1}^M$, we can use any information criterion for featurization selection. Our implementation uses the Bayes information criterion BIC as default, defined as

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

where $\mathrm{df}_{\pi,j}$ and $\mathrm{df}_{\mu,j}$ are respectively the degree of freedom of $\phi_{\pi,j}$ and $\phi_{\mu,j}$. 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 are divided into K folds. For the kth fold, the expected log-likelihood $\tilde{l}_{jk}$ is computed by taking the kth fold as the hold-out set and fitting the parameters on other folds. The selection is then based on $\tilde{l}_j = \Sigma_{k=1}^K \tilde{l}_{jk}$.

We emphasize that any of the 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 model (14), level surfaces of the local FDR are given by

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

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 $s_{t+1}(x) \leqslant s_t(x)$. 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 $s_{t+1}(x)$ deviates from $s_t(x)$ only slightly.

In this paper 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 equation (22). The choice of c can be computed in the following way.

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

$$\mathrm{fdr}_{t,i} = \frac{f(1 \mid x_i)}{f(p'_{t,i} \mid x_i)} = \frac{\hat{\pi}_{1i}\, h(1; \hat{\mu}_i) + 1 - \hat{\pi}_{1i}}{\hat{\pi}_{1i}\, h(p'_{t,i}; \hat{\mu}_i) + 1 - \hat{\pi}_{1i}}, \tag{23}$$

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 $\mathrm{fdr}_{t,i}$ among all *partially masked p-values*. (Strictly speaking, c should be slightly smaller than $\max_i \mathrm{fdr}_{t,i}$. 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

4.4.1. Initial thresholds

As shown in algorithm 1, AdaPT starts from some curve $s_0(x)$ and then slowly updates it. If the hypotheses are not ordered, then we can simply set $s_0(x) \equiv s_{0,1}$ with $s_{0,1} \leqslant 0.5$. 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.

4.4.2. 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.

4.4.3. 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 $\hat{t}_\alpha = \min\{t : \widehat{\mathrm{FDP}}_t \leqslant \alpha\}$ 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

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

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 AdaPT, we apply it to the GEOquery gene–dosage data (Davis and Meltzer, 2007), which have been analysed repeatedly as a benchmark for ordered testing procedures (Li and Barber, 2016, 2017; Lei and Fithian, 2016). We use algorithm 2 with a beta mixture model (15) for the E-step (see the on-line appendix A.1.1 for details) and a gamma GLM with canonical link function for the M-step. This data set consists of gene expression measurements for $n = 22\,283$ genes, in response to oestrogen treatments in breast cancer cells for five groups of patients, with different dosage levels and five trials in each. The task is to identify the genes responding to a low dosage. The p-values p_i for gene i are 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).

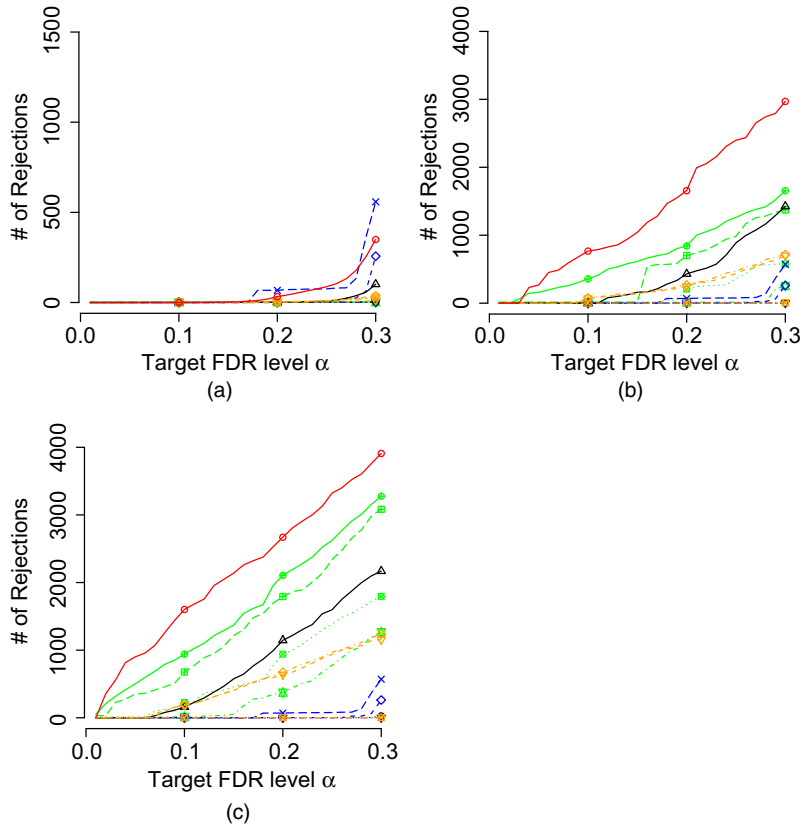

Fig. 2. Number of discoveries, in the gene–drug response data set, by each method at a range of target FDR levels α from 0.01 to 0.30 (each panel plots the results for an ordering, ranging from (a) random ordering through (b) moderately informative to (c) highly informative: ⊠, SeqStep; ⊞, HingeExp; ✡, ForwardStop; ⊕, Adaptive SeqStep; •–▽–•, BH; – –◇– –, Storey–BH; ×, Barber-Candès; △, SABHA (step); ···○···, SABHA (ordered); – –▽– –, IHW; •–◇–•, IHW (oracle); ☒, IF (oracle); —○—, AdaPT

We consider two orderings: first, a stronger (more informative) ordering based on a comparison with the highest dosage; second, a weaker (less informative) ordering based on a comparison with 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 (2016, 2017). 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 have been 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 12 other methods:

(a) SeqStep with parameter $C = 2$ (Barber and Candès, 2015);
(b) ForwardStop (Grazier G'Sell *et al.*, 2016);
(c) the accumulation test with the HingeExp function and parameter $C = 2$ (Li and Barber, 2017);
(d) Adaptive SeqStep with $s = q$ and $\lambda = 1 - q$ (Lei and Fithian, 2016);
(e) the BH procedure (Benjamini and Hochberg, 1995);
(f) Storey's BH procedure with threshold $\lambda = 0.5$ (Storey *et al.*, 2004);

(g) the Barber–Candès method (Barber and Candès, 2015; Arias-Castro and Chen, 2017);
(h) SABHA with $\tau = 0.5, \epsilon = 0.1$ and stepwise constant weights, monotone, taking values in $\{\epsilon, 1\}$ (see section 4.1 of Li and Barber (2016));
(i) SABHA with $\tau = 0.5, \epsilon = 0.1$ and monotone weights, taking values in $[\epsilon, 1]$ (see section 4.1 of Li and Barber (2016));
(j) IHW with the 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 cut-off 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 the FDR at level 1.134α (lemma 1 of Li and Barber (2016)) when the target level is α. Despite the potential anticonservativeness of these methods, to compare their best possible performance with AdaPT we do not make a correction. Fig. 2 shows the number of discoveries with different target FDR levels. We show the range of α only from 0.01 to 0.3 since it is rare to allow the 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 6–10 equiquantile knots via BIC at the initial step and kept the same afterwards; see Section 4.2.

Figs 2(b) and 2(c) 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 Fig. 2(a). In this case, the number of rejections drops dramatically and the power is almost the same as in the 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

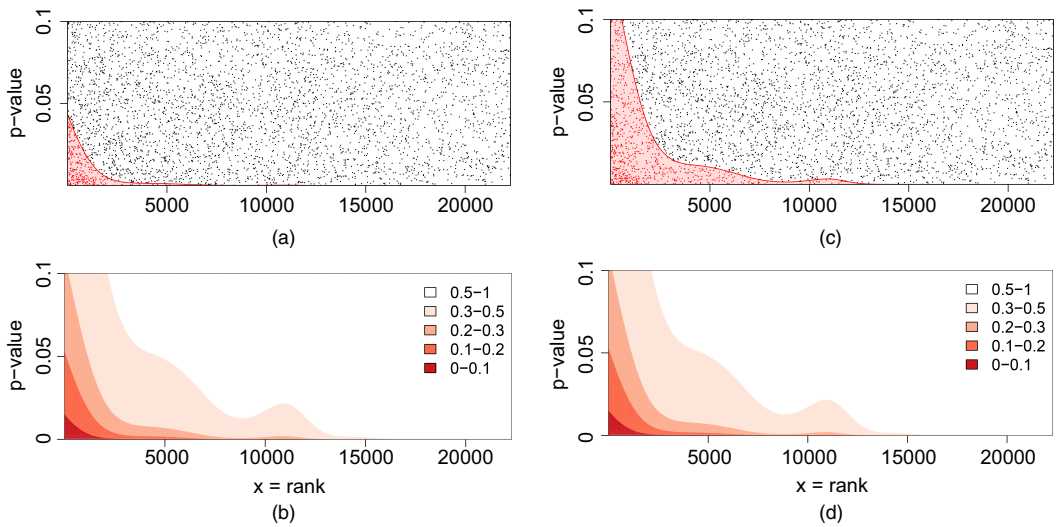

Fig. 3. Results for the gene–drug response data with moderately informative ordering of p-values, i.e. $\{p_i^{\mathrm{W}}\}$, with (a), (b) $\alpha = 0.05$ and (c), (d) $\alpha = 0.1$: (a), (c) ●, p-values and ●, rejected p-values; ———, thresholding rule $s(x)$; (b), (d) contour plots of the estimated local FDR

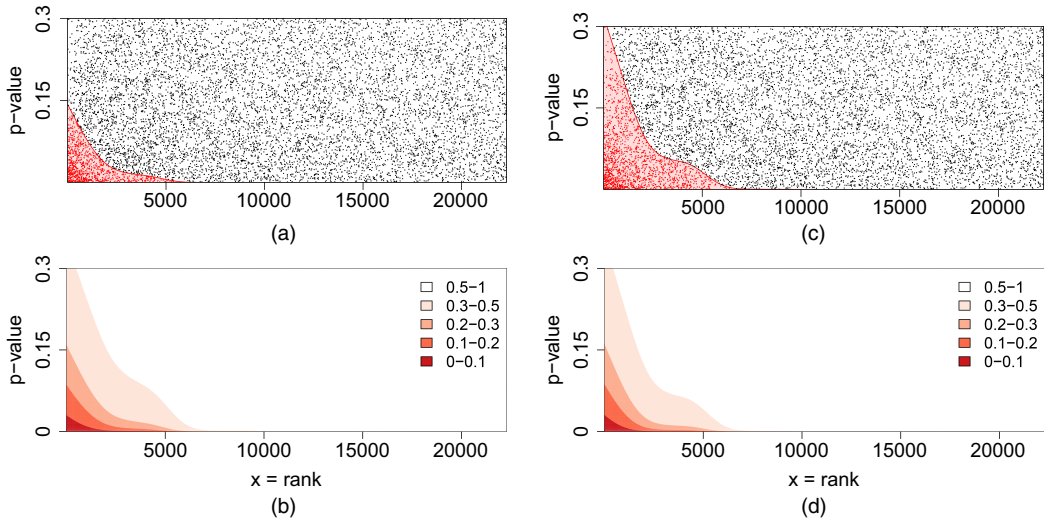

Fig. 4. Results for the gene–drug response data with highly informative ordering of p-values, i.e. $\{p_i^S\}$, with (a), (b) $\alpha = 0.05$ and (c), (d) $\alpha = 0.1$: (a), (c) p-values; ●, rejected p-values; ———, thresholding rule $s(x)$; (b), (d) contour plots of the estimated local FDR

p-values with highly informative ordering respectively in Fig. 3 and Fig. 4. It can be seen from Figs 3(b) and 3(d), and Figs 4(b) and 4(d) 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 whereas 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.

Finally, we measure the loss of information that is caused by partial masking: we first estimate the local FDR by using the set of (unmasked) p-values and the covariates, denoted by $\mathrm{fdr}^*(x)$. It can be regarded as the best possible estimate given the algorithm. Let $\mathrm{fdr}_t(x)$ denote the estimate of the local FDR at step t (based on partially masked p-values). Then we measure the loss of information by the correlation of $\{\mathrm{fdr}^*(x_i)\}_{i=1}^n$ and $\{\mathrm{fdr}_t(x_i)\}_{i=1}^n$. The results are shown in Fig. 5 where the x-axis corresponds to the target FDR, in a reverse order ranging from 0.5 to 0.01, and the 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 Section 1.3, the loss of information is quite small and even negligible after the target FDR drops to the 'practical' regime (e.g. below 0.2), where the correlation between $\{\mathrm{fdr}^*(x_i)\}_{i=1}^n$ and $\{\mathrm{fdr}_t(x_i)\}_{i=1}^n$ is almost 1. The pattern is even more significant in other data examples in the next subsection. This provides 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 the 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, 2016, 2017; Lei and Fithian, 2016).

5.2. Simulation studies

5.2.1. Example 1: a two-dimensional case

We generate the covariates x_is from an equispaced 50×50 grid in the area $[-100, 100] \times$

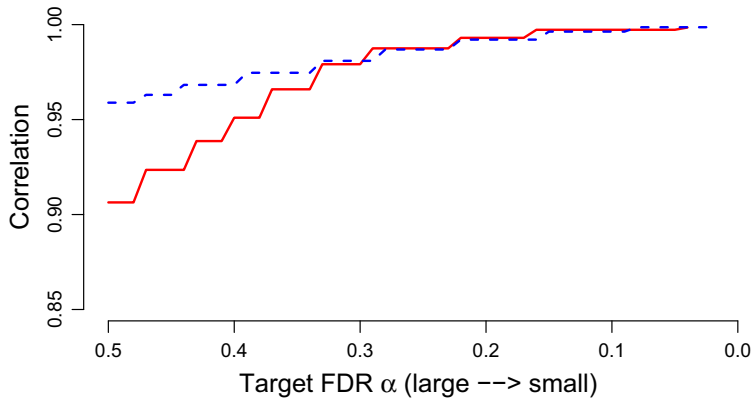

Fig. 5. Correlation of $\{\text{fdr}^*(x_i)\}_{i=1}^n$ and $\{\text{fdr}_t(x_i)\}_{i=1}^n$ for the gene–drug response dosage data set under the original, moderately informative (——) and highly informative orderings (– – –): the x-axis corresponds to the target FDR, in a reverse order ranging from 0.5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

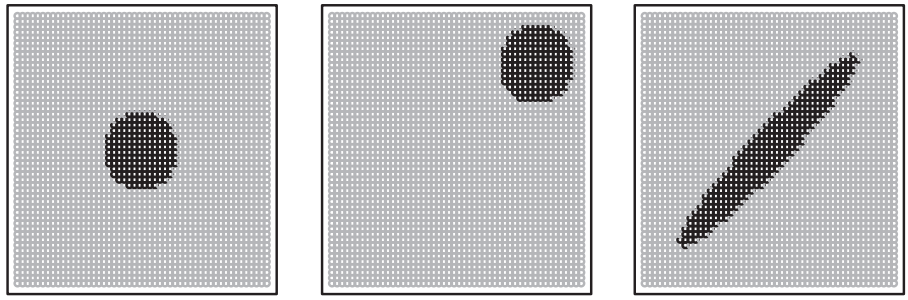

Fig. 6. Underlying ground truth for three cases in example 1 (each point represents a hypothesis (2500 in total) with nulls (●) and non-nulls (●)): (a) circle in the middle; (b) circle in a corner; (c) thin ellipse

$[-100, 100]$. We generate p-values independent and identically distributed from a one-sided normal test, i.e.

$$\begin{aligned} p_i &= 1 - \Phi(z_i), \\ z_i &\sim N(\mu, 1), \end{aligned} \tag{24}$$

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

In this case, it is not clear how to apply non-adaptive ordered testing procedures or the IF. Thus we compare AdaPT only with Storey's BH method, the Barber–Candès method, IHW using the default automatic parameter tuning procedure and SABHA using two-dimensional low total variation weights (see section 4.3 of Li and Barber (2016)). For AdaPT, we fit two-dimensional generalized additive models in the M-step, using R package `mgcv` with the knots selected automatically in every step by the generalized cross-validation 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 FDP and the power as

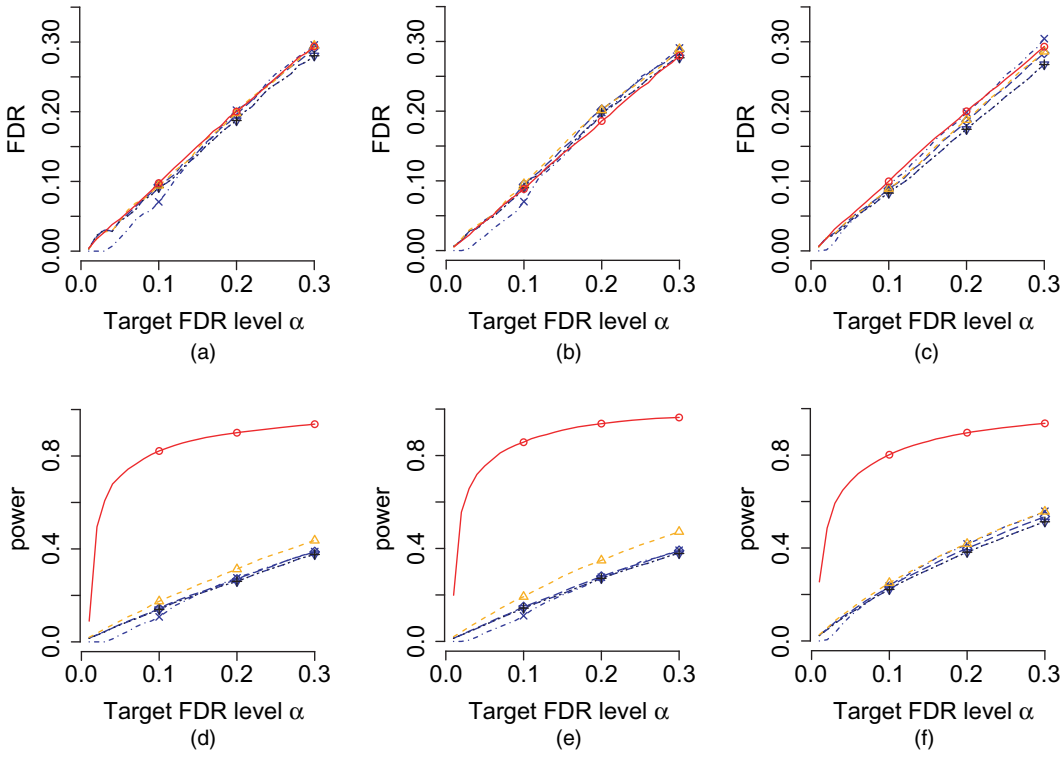

Fig. 7. (a)–(c) FDR and (d)–(f) power with $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$ in example 1 (▽, BH; ◇, Storey–BH; ×, Barber–Candès; +, SABHA; △, IHW; ○, AdaPT): (a), (d) circle in the middle; (b), (e) circle in the corner; (c), (f) thin ellipse

$$\begin{aligned} \mathrm{FDP}(\alpha) &= \frac{|\mathcal{R}_\alpha \cap \mathcal{H}_0|}{|\mathcal{R}_\alpha|}, \\ \mathrm{power}(\alpha) &= \frac{|\mathcal{R}_\alpha \cap \mathcal{H}_0^c|}{|\mathcal{H}_0^c|}. \end{aligned} \tag{25}$$

We repeat the above procedure on 100 fresh simulated data sets and calculate the average of $\mathrm{FDP}(\alpha)$ and $\mathrm{power}(\alpha)$ as the measure of FDR and power. The results are shown in Fig. 7. It is clearly seen that AdaPT controls the FDR like the other methods while achieving a significantly higher power.

To see why AdaPT gains power, we plot the estimated local FDR in Fig. 8 for the first case at the steps where $\widehat{\mathrm{FDP}}$ is first below 0.5, 0.3 and 0.1. As shown in the real examples, the fitted local FDR identifies the non-nulls quite accurately even at the early steps where most p-values are partially masked. The estimates become very stable and informative after reaching the practical regime of αs.

5.2.2. *Example 2: a 100-dimensional case*

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

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

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

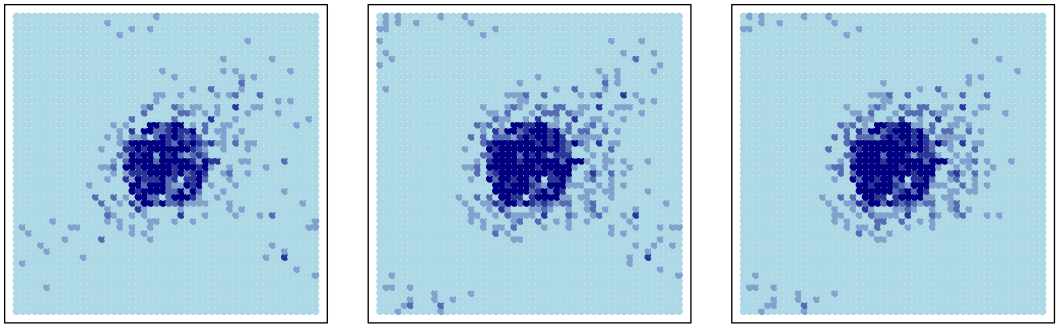

Fig. 8. Estimated local FDR in the first case of example 1 with (a) the target FDR level 0.5, with (b) the target FDR level 0.3 and with (c) the target FDR level 0.1: dark colour marks the hypotheses with low local FDR and light colour with high local FDR

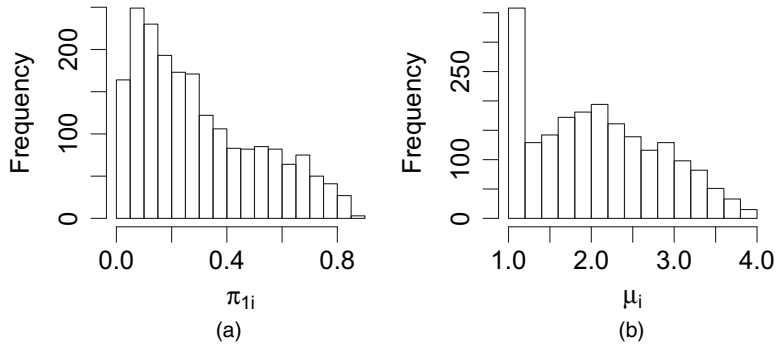

Fig. 9. Distributions of (a) π_{1i}s and (b) μ_is in example 2

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

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

$$\theta=(3,3,0,\ldots,0)^{\mathrm{T}},$$
$$\beta=(2,2,0,\ldots,0)^{\mathrm{T}}$$

and θ_0 is chosen so that $(1/n)\Sigma_{i=1}^n\pi_{1i}=0.3$. In this case, $\mathbb{E}[-\log(p_i)]=\mu_i$ under the alternative. Fig. 9 shows the histograms of the π_{1i}s and μ_is.

In this case, it is not clear how to apply non-adaptive ordered testing procedures, or IF or adaptive procedures like IHW and SABHA. Thus we compare AdaPT with only the BH method, Storey's BH method and the Barber–Candès method. For AdaPT, we fit L_1-regularized GLMs in the M-step (see the on-line 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 by using 100 replications. The results are plotted in Fig. 10. It is clearly seen that both AdaPTs control the FDR like the other methods while achieving a higher power. Not surprisingly, compared with AdaPT with L_1-regularized GLMs, AdaPT with oracle GLMs has a higher power. Nevertheless, this example shows the

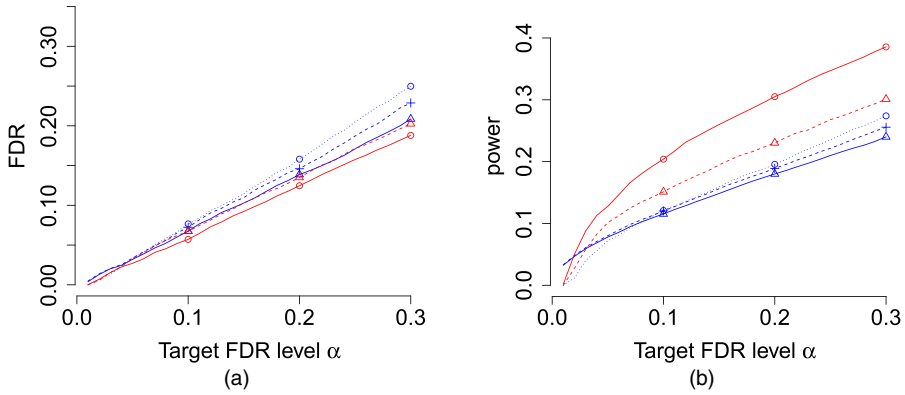

Fig. 10. (a) FDR and (b) power with $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$ in example 2: —△—, BH; ···+···, Storey–BH; ○, Barber–Candès; - -△- -, AdaPT; —○—, AdaPT (oracle)

unprecedented ability of AdaPT to improve power by squeezing information from a large set of noisy features.

5.3. *Other example applications*

In this section, we examine the performance of AdaPT on four more real data sets, which have been analysed in other references exploiting adaptive FDR control methods, e.g. Bourgon *et al.* (2010) and Ignatiadis *et al.* (2016). In all cases, we start with a brief introduction of the data set and show the plots on the number of rejections, like Fig. 2, the path of information loss, like Fig. 5, and the threshold curve and level curves of estimated local FDR with target FDR 0.1, like Fig. 3 and Fig. 4. We use the same settings for AdaPT as in the gene dosage data set: performing model selection at the initial step with candidate featurization being all combinations of spline basis with 6–10 equiquantile knots on $\pi(x)$ and $\mu(x)$, and fixing the selected model in subsequent updates.

5.3.1. *Bottomly data*

The Bottomly data set is an RNA sequencing data set targeting on detecting the differential expression in two mouse strains, C57BL/6J (B6) and DBA/2J (D2), collected by Bottomly *et al.* (2011), which is available from the ReCount repository (Frazee *et al.*, 2011), and analysed by Ignatiadis *et al.* (2016) by using IHW. It consists of gene expression measurements for $n = 13\,932$ genes. Following Ignatiadis *et al.* (2016), we analyse the data by using the `DEseq2` package (Love *et al.*, 2014) and use the logarithm of normalized count (averaged across all samples) as the univariate covariate for each gene. The results are plotted in Fig. 11. It is clearly seen that AdaPT produces significantly more discoveries than all the other methods and the loss of information 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.

5.3.2. *Airway data*

The airway data set is an RNA sequencing data set targeting identifying the differentially expressed genes in airway smooth muscle cell lines in response to dexamethasone, which was collected by Himes *et al.* (2014) and is available in R package `airway`. It is analysed in the

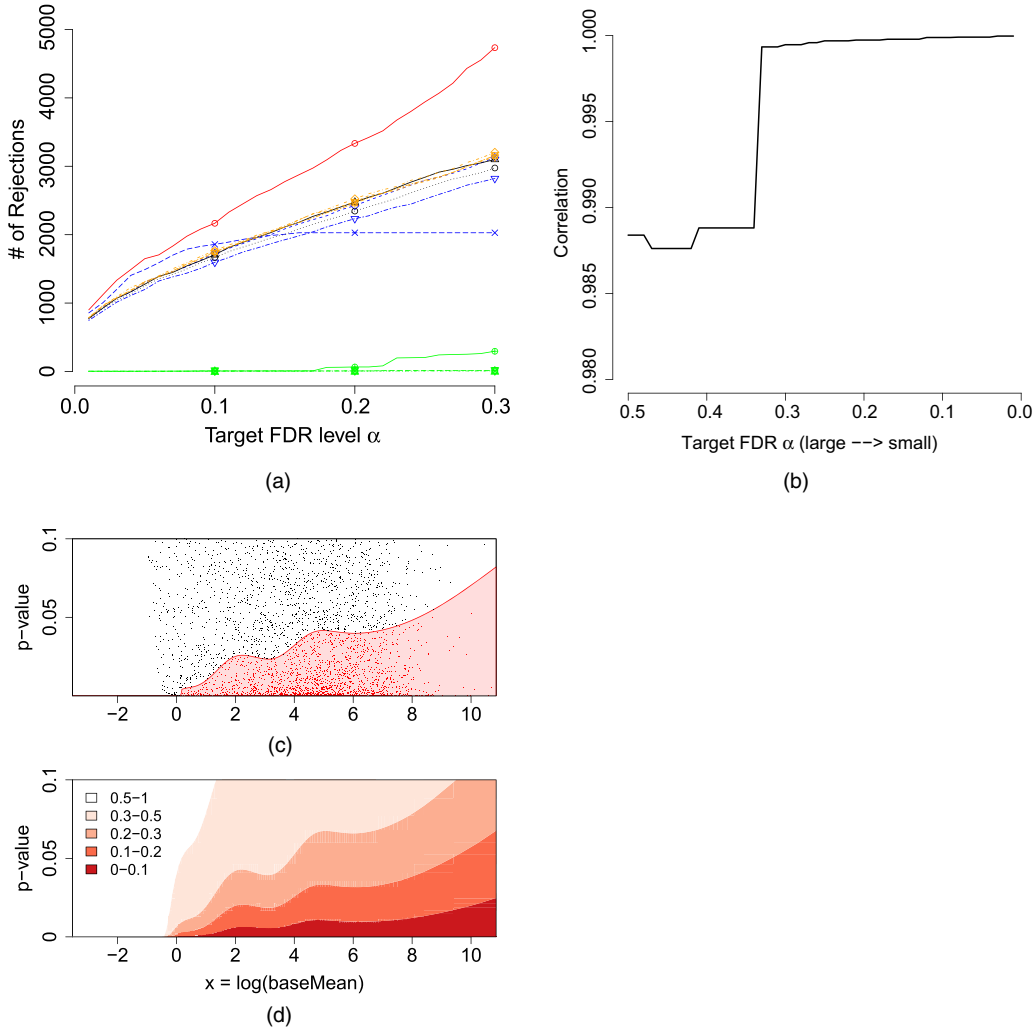

Fig. 11. Results for the Bottomly data set: (a) number of rejections (⊠, SeqStep; ⊞, HingeExp; ✡, ForwardStop; ⊕, Adaptive SeqStep; •−▽−•, BH; - -◇- -, Storey–BH; ×, Barber–Candès; △, SABHA (step); ···○···, SABHA (ordered); - -▽- -, IHW; •-◇-•, IHW (oracle); ☒, IF (oracle); —○—, AdaPT); (b) path of information loss; (c) threshold curve ($\alpha = 0.1$); (d) level curves of the estimated local FDR ($\alpha = 0.1$)

vignette of the `IHW` package by using the IHW method of Ignatiadis *et al.* (2016). As in the vignette and the previous example, we analyse the data by using the `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 Fig. 12. Again, AdaPT produces significantly more discoveries than all the other methods.

5.3.3. Pasilla data

The pasilla data set is also an RNA sequencing data set targeting 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

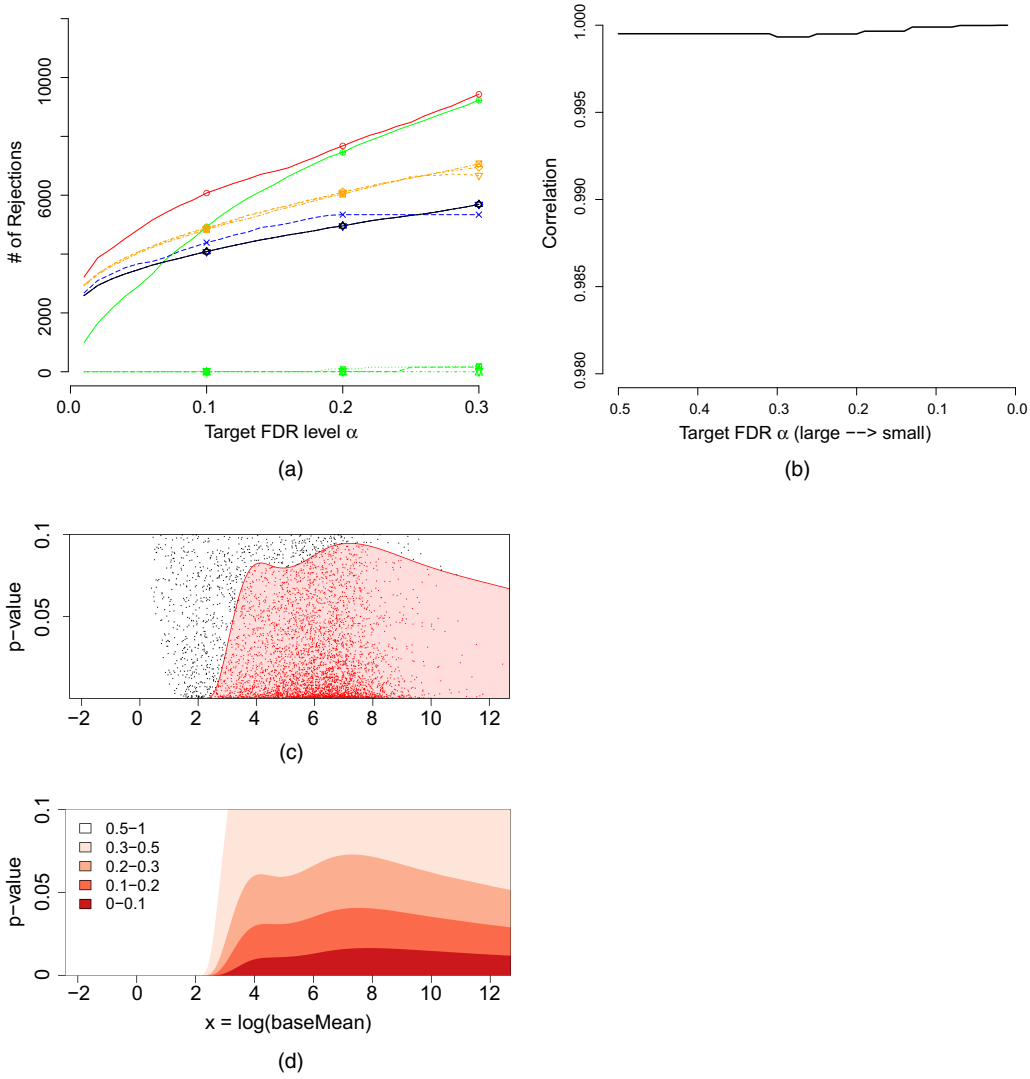

Fig. 12. Results for the airway data set: (a) number of rejections (⊠, SeqStep; ⊞, HingeExp; ✡, ForwardStop; ⊕, Adaptive SeqStep; •−▽−•, BH; - -◇- -, Storey–BH; ×, Barber–Candès; △, SABHA (step); ···○···, SABHA (ordered); - -▽- -, IHW; •−◇−•, IHW (oracle); ☒, IF (oracle); —○—, AdaPT); (b) path of information loss; (c) threshold curve ($\alpha = 0.1$); (d) level curves of the estimated local FDR ($\alpha = 0.1$)

analysed in the vignette of the `genefilter` package (Gentleman *et al.*, 2016) using the independent filtering method (Bourgon *et al.*, 2010). As in the vignette, we analyse the data by using the `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 Fig. 13. It is clear that we arrive at the same conclusion that AdaPT is more powerful than all the other methods.

5.3.4. Yeast proteins data

The yeast proteins data set SILAC is a proteomics data set, which was collected by Dephoure

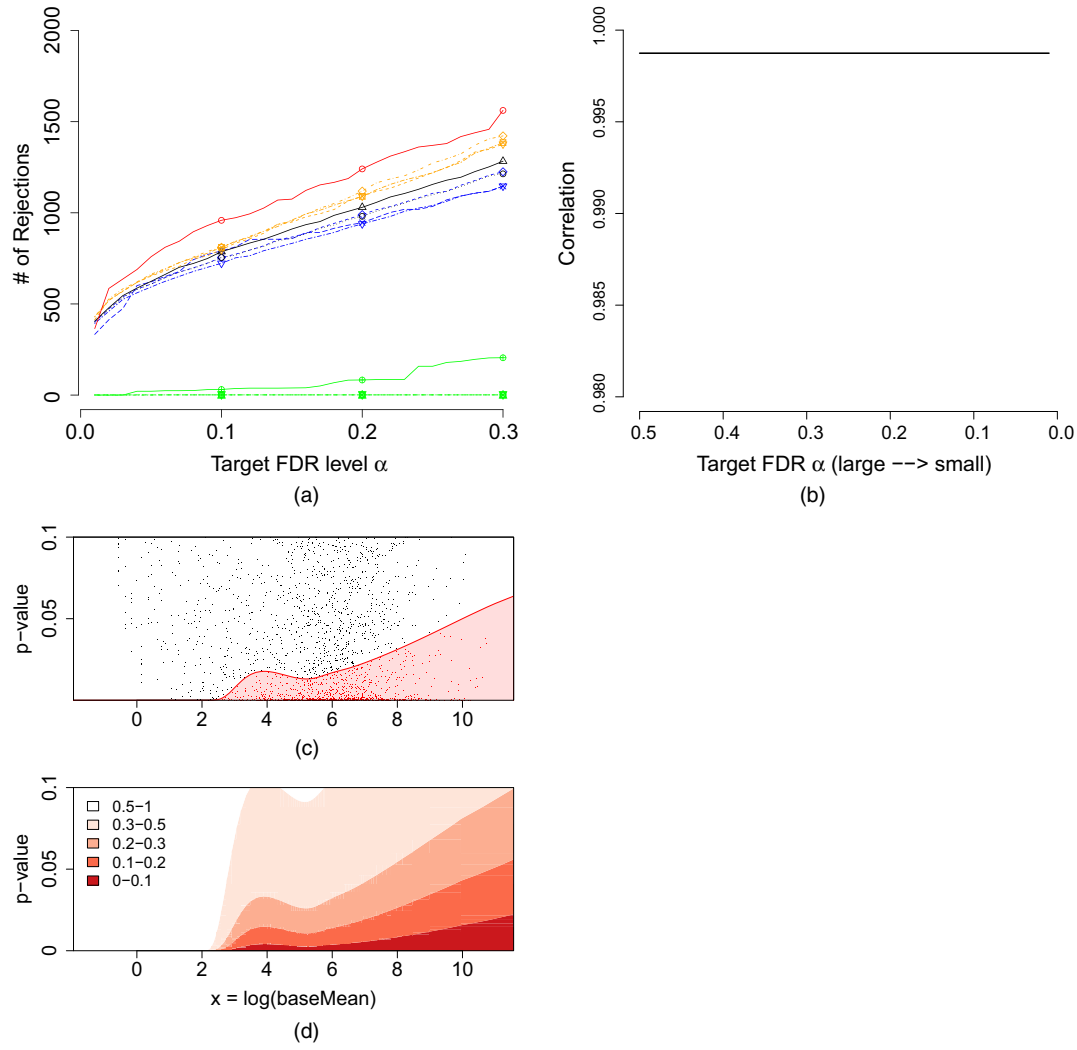

Fig. 13. Results for the pasilla data set: (a) number of rejections (⊠, SeqStep; ⊞, HingeExp; ✡, ForwardStop; ⊕, Adaptive SeqStep; •−▽−•, BH; - -◇- -, Storey–BH; ×, Barber–Candès; △, SABHA (step); · · ·○· · ·, SABHA (ordered); - -▽- -, IHW; •-◇-•, IHW (oracle); ☒, IF (oracle); —○—, AdaPT); (b) path of information loss; (c) threshold curve ($\alpha = 0.1$); (d) level curves of the estimated local FDR ($\alpha = 0.1$)

and Gygi (2012) and is 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 was analysed in Ignatiadis *et al.* (2016) by using the IHW method. As in Dephoure and Gygi (2012) and Ignatiadis *et al.* (2016), we calculate the p-values by using Welch's t-test and use as the univariate covariate the logarithm of the total number of peptides that were quantified across all samples for each gene. The results are plotted in Fig. 14. In this case, AdaPT has a similar performance to that of the Barber–Candès method and Storey's BH method. However, it still outperforms all the other methods. Furthermore, AdaPT learns the monotone pattern of the local FDR, which coincides with the heuristic.

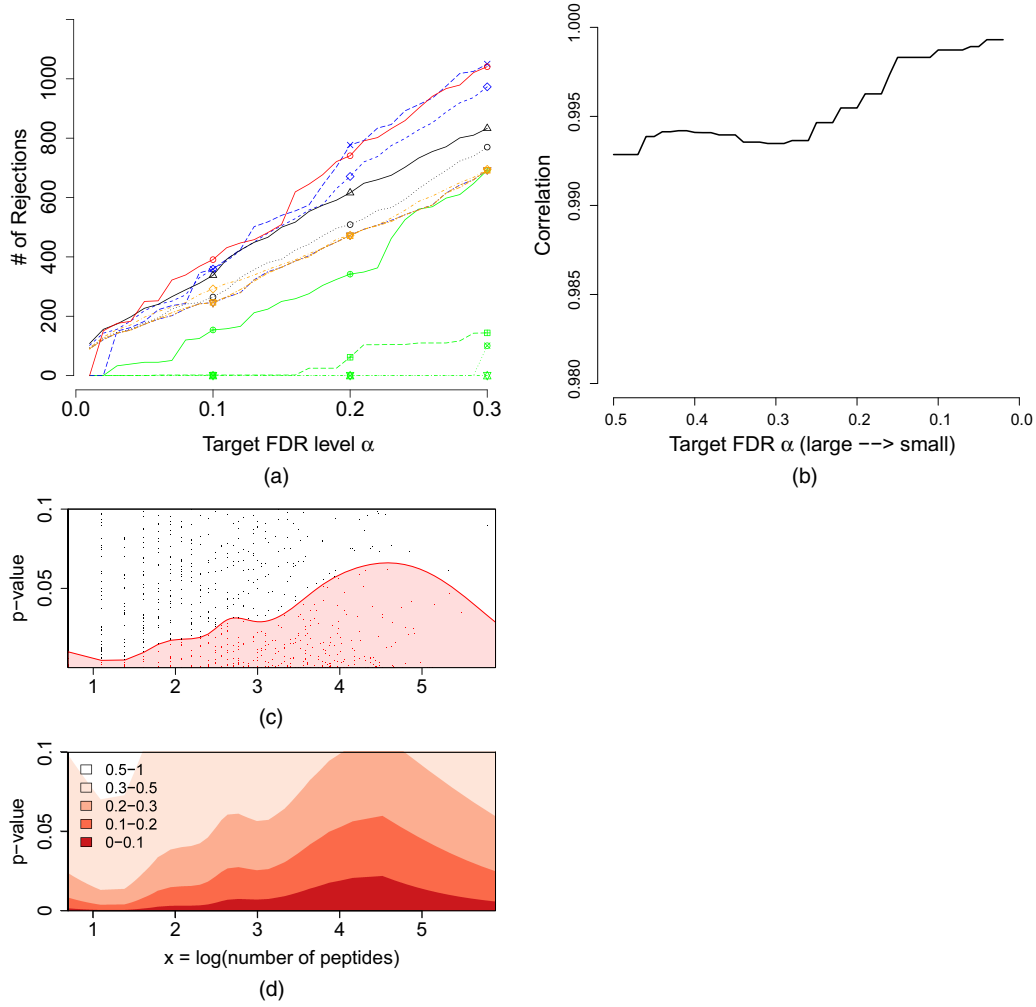

Fig. 14. Results for the yeast proteins data set: (a) number of rejections (⊠, SeqStep; ⊞, HingeExp; ✡, ForwardStop; ⊕, Adaptive SeqStep; •−▽−•, BH; - -◇- -, Storey–BH; ×, Barber–Candès; △, SABHA (step); ···○···, SABHA (ordered); - -▽- -, IHW; •−◇−•, IHW (oracle); ☒, IF (oracle); —○—, AdaPT); (b) path of information loss; (c) threshold curve ($\alpha = 0.1$); (d) level curves of the estimated local FDR ($\alpha = 0.1$)

6. Discussion

We have proposed the adaptive procedure AdaPT, which is 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 enables 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, the FDR is still controlled at the advertised level (though power could be adversely affected by

Table 3. Algorithm 3: AdaPT without thresholds

Input: predictors and p-values $(x_i, p_i)_{i\in[n]}$ and target FDR level α
Procedure:
1, initialize $\mathcal{R}_0 = [n]$
2, *for* $t = 0, 1, \ldots,$ *do*
3, $\quad R_t \leftarrow \#\{i \in \mathcal{R}_t : p_i \leqslant \frac{1}{2}\}$; $A_t \leftarrow \#\{i \in \mathcal{R}_t : p_i > \frac{1}{2}\}$
4, $\quad \widehat{\mathrm{FDP}}_t \leftarrow (1 + A_t)/(R_t \vee 1)$
5, $\quad$ *if* $\widehat{\mathrm{FDP}}_t \leqslant \alpha$ *then*
6, $\quad\quad$ reject $\mathcal{R}_t$;
7, $\quad$ *end if*
8, $\quad \mathcal{R}_{t+1} \leftarrow \mathrm{update}\{(x_i, \tilde{p}_{t,i})_{i\in[n]}, \mathcal{R}_t\}$
9, *end for*

overfitting).We show by 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 shown in Table 3 in the absence of $s(x)$.

Rephrasing algorithm 3, we start from partially masking all p-values, yielding a 'candidate rejection set' $\mathcal{R}_0 = [n]$, and then apply an arbitrary method to update $\mathcal{R}_t$ directly. The FDP estimator (step 4) is defined in an essentially identical way to that in 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 difficult to see that our implementation in Section 4 can be reformulated in a more simple and straightforward way: in each step we estimate the local FDR for each partially masked p-value and peel off a 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 favourable scores and to proceed with refitted scores by exploiting the p-values revealed. Heuristically, the most statistical meaningful score is the 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 developed a general framework for controlling the FDR under structural constraints. We refer readers to Lei *et al*. (2017) for more thoughts in this vein.

6.2. Extension to dependent data by 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, AdaPT does not control the FDR in finite samples for dependent p-values. In particular, there is a danger of overfitting to local random effects that are shared by nearby hypotheses: to AdaPT, such random effects are treated as signals to discover.

It could be interesting to pursue a hybrid method using ideas from AdaPT and the Knockoff+ procedure 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 $j = 1, \ldots, d$. The key

step in Barber and Candès (2015) is to compute another matrix $\tilde{X} \in \mathbb{R}^{n \times d}$ with $\tilde{X}'\tilde{X} = X'X$ and $\tilde{X}'X = X'X - D$, for some diagonal $D \in \mathbb{R}^{d \times d}$ with positive entries; this can be done provided that $n \geqslant 2d$ and X has full column rank.

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

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

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, conditionally on $\mathcal{F}_{-1}$, $b_j = 1 - \operatorname{sgn}(w_j)$ is a mirror conservative 'binary p-value', i.e. $(b_j)_{j \in \mathcal{H}_0}$ are IID Bern$(\frac{1}{2})$ 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, AdaPT 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 $s_t(|w_j|) = 0.5\mathbf{1}\{|w_j| \geqslant q_t\}$. More generally, we can use AdaPT and interactively change the threshold that we use.

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

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

In that case $(p_j)_{j \in \mathcal{H}_0}$ are IID 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 by using $v + \tilde{v}$ as a predictor. Although 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 $d \geqslant 2n$. In that case Barber and Candès (2015) suggested using the knockoff matrix $\tilde{X}$ of d more orthonormal columns which are also orthogonal to the columns of X. Then $X'y \sim \mathcal{N}_d(\beta, \sigma^2 I_d)$ whereas $\tilde{X}'y \sim \mathcal{N}_d(0, \sigma^2 I_d)$, independently.

In this case, using the lasso, forward stepwise regression or virtually any other model selection path procedure on the design matrix $[X\tilde{X}]$ is identical to selecting variables in decreasing order of absolute value of $|X'_j y|$ and $|\tilde{X}'_j y|$, or, equivalently, in increasing order of the two-tailed p-values $p_j = 2 - 2\Phi(|X'_j y|/\sigma)$ and $p_j^* = 2 - 2\Phi(|\tilde{X}'_j y|/\sigma)$ (this is true whether or not σ^2 is known). As a result, if we operationalize the Knockoff+ procedure by using for example the 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 IID pair (p_j, p_j^*) in knockoffs.

The two most salient differences between AdaPT and Knockoff+ in this case are as follows.

(a) AdaPT enables 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 subroutine update, the AdaPT result is a deterministic function of the original data.

6.4. Extension: estimating local false discovery rate

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 that we have used to produce these estimates, they may provide the best summary of evidence against the individual hypothesis H_i. By contrast, the level of significance for the 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 the FDR at some level $\alpha \ll 1$. Even apart from such perversities, it will typically be that $\widehat{\mathrm{fdr}}(p_i \mid x_i) > \alpha$ for many of the rejected hypotheses.

Despite their more favourable interpretation, however, the local FDR estimates that are 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:

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

or

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

If we take an infinitely large window, we obtain $\widehat{\mathrm{fdp}}^{+}_{t,\infty} = \widehat{\mathrm{FDP}}_t$; thus, these estimators adaptively estimate the FDP for p-values revealed in the next w steps of the algorithm, in much the same way that $\widehat{\mathrm{FDP}}_t$ estimates the FDP 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.

Acknowledgements

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

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A. and Weindruch, R. (2002) A

mixture model approach for the analysis of microarray gene expression data. *Computnl Statist. Data Anal.*, **39**, 1–20.
Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, no. 10, article R106.
Arias-Castro, E. and Chen, S. (2017) Distribution-free multiple testing. *Electron. J. Statist.*, **11**, 1983–2001.
Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.
Barber, R. F. and Candès, E. J. (2016) A knockoff filter for high-dimensional selective inference. *Preprint arXiv:1602.03574*. Department of Statistics, University of Chicago, Chicago.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Benjamini, Y. and Hochberg, Y. (1997) Multiple hypotheses testing with weights. *Scand. J. Statist.*, **24**, 407–418.
Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *Ann. Statist.*, **41**, 802–837.
Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K. and Hitzemann, R. (2011) Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using RNA-seq and microarrays. *PLOS One*, **6**, no. 3, article e17820.
Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natn. Acad. Sci. USA*, **107**, 9546–9551.
Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E. and Graveley, B. R. (2011) Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.*, **21**, 193–202.
Davis, S. and Meltzer, P. S. (2007) GEOquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, **23**, 1846–1849.
Dephoure, N. and Gygi, S. P. (2012) Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci. Signlng*, **5**, article rs2.
Dobriban, E. (2016) A general convex framework for multiple testing with prior information. *Preprint arXiv:1603.05334*. Stanford University, Stanford.
Dobriban, E., Fortney, K., Kim, S. K. and Owen, A. B. (2015) Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika*, **102**, 753–766.
Dobson, A. J. and Barnett, A. (2008) *An Introduction to Generalized Linear Models*. Boca Raton: CRC Press.
Du, L. and Zhang, C. (2014) Single-index modulated multiple testing. *Ann. Statist.*, **42**, 1262–1311.
Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. L. (2015) Preserving statistical validity in adaptive data analysis. In *Proc. 47th A. Symp. Theory of Computing*, pp. 117–126. New York: Association for Computing Machinery.
Efron, B. (2007) Size, power and false discovery rates. *Ann. Statist.*, **35**, 1351–1377.
Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.
Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. and Kong, A. (2008) Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Statist.*, **2**, 714–735.
Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. Department of Statistics, University of California, Berkeley.
Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C. and Owen, A. B. (2015) Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLOS Genet.*, **11**, no. 12, article e1005728.
Frazee, A. C., Langmead, B. and Leek, J. T. (2011) Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinform.*, **12**, article 449.
Genovese, C. R., Roeder, K. and Wasserman, L. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.
Gentleman, R., Carey, V., Huber, W. and Hahne, F. (2016) genefilter: genefilter: methods for filtering genes from high-throughput experiments. *R Package Version 1.54.2*. European Bioinformatics Institute, Cambridge.
Grazier G'Sell, M. G., Wager, S., Chouldechova, A. and Tibshirani, R. (2016) Sequential selection procedures and false discovery rate control. *J. R. Statist. Soc.* B, **78**, 423–444.
Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderman, B., Whitaker, R. M., Duan, Q., Lasky-Su, J. and Nikolos, C. (2014) RNA-seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLOS One*, **9**, no. 6, article e99625.
Hu, J. X., Zhao, H. and Zhou, H. H. (2010) False discovery rate control with groups. *J. Am. Statist. Ass.*, **105**, 1215–1227.
Huber, W. and Reyes, A. (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*. European Molecular Biology Laboratory, Heidelberg.
Ignatiadis, N. and Huber, W. (2017) Covariate-powered weighted multiple testing with false discovery rate control. *Preprint arXiv:1701.05179*. Department of Statistics, Stanford University, Stanford.

Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Meth.*, **13**, 577–580.
Lawyer, G., Ferkingstad, E., Nesvåg, R., Varnäs, K. and Agartz, I. (2009) Local and covariate-modulated false discovery rates applied in neuroimaging. *NeuroImage*, **47**, 213–219.
Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.
Lei, L. and Fithian, W. (2016) Power of ordered hypothesis testing. In *Proc. Int. Conf. Machine Learning*, pp. 2924–2932.
Lei, L., Ramdas, A. and Fithian, W. (2017) STAR: a general interactive framework for fdr control under structural constraints. *Preprint arXiv:1710.02776*. Department of Statistics, University of California, Berkeley.
Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J. and Thomas, D. C. (2007) Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidem.*, **31**, 871–882.
Li, A. and Barber, R. F. (2016) Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *Preprint arXiv:1606.07926*. Department of Statistics, University of Chicago, Chicago.
Li, A. and Barber, R. F. (2017) Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Statist. Ass.*, **112**, 837–849.
Love, M. I., Anders, S. and Huber, W. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.*, **15**, no. 12, article 550.
Markitsis, A. and Lai, Y. (2010) A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, **26**, 640–646.
Parker, R. and Rothenberg, R. (1988) Identifying important results from multiple statistical tests. *Statist. Med.*, **7**, 1031–1043.
Pounds, S. and Morris, S. W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.
Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.
Storey, J. D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc.* B, **69**, 347–368.
Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA*, **100**, 9440–9445.
Sun, W., Reich, B. J., Cai, T. T., Guindani, M. and Schwartzman, A. (2015) False discovery control in large-scale spatial multiple testing. *J. R. Statist. Soc.* B, **77**, 59–83.
Tian, X. and Taylor, J. E. (2018) Selective inference with a randomized response. *Ann. Statist.*, **46**, 679–710.
Tukey, J. W. (1994) *The Collected Works of John W. Tukey*, vol. 8, *Multiple Comparisons, 1948-1983*. New York: Chapman and Hall.
Yekutieli, D. (2012) Adjusted Bayesian inference for selected parameters. *J. R. Statist. Soc.* B, **74**, 515–541.
Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. and Thompson, W. K. (2014) Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, **30**, 2098–2104.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material of "AdaPT: an interactive procedure for multiple testing with side information"'.