Biometrika (2021), **108**, 2, *pp.* 253–267
Printed in Great Britain

doi: 10.1093/biomet/asaa064
Advance Access publication 22 October 2020

A general interactive framework for false discovery rate control under structural constraints

By LIHUA LEI

Department of Statistics, Stanford University, 202 Sequoia Hall,
390 Serra Mall, Stanford, California 94305, U.S.A.
lihualei@stanford.edu

AADITYA RAMDAS

Department of Statistics and Data Science, Carnegie Mellon University,
132H Baker Hall, Pittsburgh, Pennsylvania 15213, U.S.A.
aramdas@cmu.edu

and WILLIAM FITHIAN

Department of Statistics, University of California, Berkeley,
301 Evans Hall, Berkeley, California 94720, U.S.A.
wfithian@berkeley.edu

Summary

We propose a general framework based on selectively traversed accumulation rules for interactive multiple testing with generic structural constraints on the rejection set. It combines accumulation tests from ordered multiple testing with data-carving ideas from post-selection inference, allowing highly flexible adaptation to generic structural information. Our procedure defines an interactive protocol for gradually pruning a candidate rejection set, beginning with the set of all hypotheses and shrinking the set with each step. By restricting the information at each step via a technique we call masking, our protocol enables interaction while controlling the false discovery rate in finite samples for any data-adaptive update rule that the analyst may choose. We suggest update rules for a variety of applications with complex structural constraints, demonstrate that selectively traversed accumulation rules perform well in problems ranging from convex region detection to false discovery rate control on directed acyclic graphs, and show how to extend the framework to regression problems where knockoff statistics are available in lieu of p-values.

Some key words: Accumulation test; Data carving; False discovery rate; Interactive multiple testing; Knockoff; Masking.

1. Introduction

From a classical statistical perspective, data analysis can be divided into two distinct types: exploratory analysis is a flexible and iterative process of searching the data for interesting patterns while pursuing only a loose error guarantee, or none at all, while confirmatory analysis involves performing targeted inferences on questions that were preselected for focused study. Selective inference blends exploratory and confirmatory analysis by allowing for inference on questions that may be selected in a data-adaptive way, but most selective inference methods still require the analyst to pre-commit to selection rules before observing the data, falling short of the freewheeling nature of true exploratory analysis. By contrast, interactive methods constitute a subset of selective inference methods that allow the analyst to react to data, consult their own internal judgement, and revise their models and research plans to adapt to patterns they may not have expected to find, while still achieving valid inferences.

We consider the problem of multiple hypothesis testing with p-values $p_1, \ldots, p_n$, where each p_i corresponds to a different null hypothesis H_i. A multiple testing method examines the p-values, possibly along with additional data, and decides which null hypotheses to reject. Let $\mathcal{H}_0 = \{i : H_i \text{ is true}\}$ denote the set of truly null hypotheses, and let $\mathcal{R} = \{i : H_i \text{ is rejected}\}$ denote the rejection set. Then $R = |\mathcal{R}|$ is the number of rejections and $V = |\mathcal{R} \cap \mathcal{H}_0|$ is the number of erroneous rejections. Benjamini & Hochberg (1995) defined the false discovery proportion as $V/\max\{1, R\}$ and famously proposed controlling its expectation, the false discovery rate, at some prespecified level α.

This article proposes a new framework for interactive multiple testing in structured settings with false discovery rate control, called selectively traversed accumulation rules. Our approach is especially well suited to settings where one wishes to impose structural constraints on the set of rejected hypotheses, for example to enforce a hierarchy principle for interactions in a regression or analysis of variance problem, or to detect a convex spatial region where the signal exceeds a certain level in a signal processing application. In certain cases, enforcing a structural constraint is important for logical coherence or interpretability. For instance, in hierarchical testing (e.g., Yekutieli, 2008) where hypotheses are represented as nodes on a tree, the parent of a rejected hypothesis must be rejected as well. In other settings, the structural constraint serves as a type of side information which reflects the scientific domain knowledge. By using this information judiciously, the statistical power can be boosted if the true signals satisfy the constraint, exactly or at least approximately.

More formally, our procedure controls the false discovery rate in finite samples while guaranteeing that the rejection set $\mathcal{R}$ belongs to $\mathcal{K}$, a collection of subsets satisfying the constraint. For instance, in the case of hierarchical testing, $\mathcal{K}$ includes all subsets that correspond to rooted subtrees. The notion of structure is interpreted in a rather general way: $\mathcal{K}$ could also depend on auxiliary covariate information x_i about the hypothesis H_i. For example, in the spatial testing case, x_i gives the geographic location in $\mathbb{R}^2$, and $\mathcal{K}$ may include all subsets that can be expressed as the intersection of a convex set on $\mathbb{R}^2$ with $(x_i)_{i=1}^n$.

Selectively traversed accumulation rules generalize the notion of masking used by the adaptive p-value thresholding method of Lei & Fithian (2018), which achieves its error control guarantee by judiciously limiting the analyst's knowledge of the data. As such, it is natural to view selectively traversed accumulation rules as iterative interactions between two agents: the analyst, who drives the search for discoveries based on partial observation of the data, and a hypothetical oracle, who observes the full dataset and gradually reveals information to the analyst based on the latter's actions. Typically the oracle is a computer program, and the analyst is either a human or an automated adaptive search algorithm based on predefined modelling assumptions. In § 3 we discuss generic strategies for defining good automated rules.

A key difference between selectively traversed accumulation rules and most earlier methods is that they give the analyst power to enforce structural constraints on the final rejection set. Previous approaches such as the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995), independent hypothesis weighting (Ignatiadis et al., 2016), the structure-adaptive Benjamini–Hochberg algorithm (Li & Barber, 2019) and adaptive p-value thresholding (Lei & Fithian, 2018) all produce potentially heterogeneous thresholds for p-values, and then reject all hypotheses whose p-values are below their corresponding thresholds. Because any given p-value could be above the chosen threshold, none of these methods can enforce structural constraints. On the other hand, various algorithms exist that are tailored to particular structural constraints, such as the methods of Yekutieli (2008) and Lynch & Guo (2016) for hierarchical testing on trees, and those of Lynch & Guo (2016) and Ramdas et al. (2019a) for testing on directed acyclic graphs. However, these methods are nonadaptive and noninteractive in the sense that they can neither learn the structural information from data nor incorporate extra side information, and they apply to very specific types of constraints. To the best of our knowledge, the method proposed here is the first data-adaptive and interactive multiple testing framework that can accommodate generic structural constraints on the rejection set.

2. Selectively traversed accumulation rules

2.1. *The framework*

We assume that for each hypothesis H_i $(i = 1, \ldots, n)$, the oracle observes a generic covariate $x_i \in \mathcal{X}$ and a p-value $p_i \in [0, 1]$; in § 6 we discuss a generalization to the setting where knockoff statistics instead of p-values are available. In addition, the analyst may impose a generic structural constraint $\mathcal{K} \subseteq 2^{[n]}$ which represents the allowable rejection sets, where $\subseteq$ denotes a subset. We require that $\emptyset \in \mathcal{K}$.

The selectively traversed accumulation rules proceed by adaptively generating a sequence of candidate rejection sets $[n] = \mathcal{R}_0 \supsetneq \mathcal{R}_1 \supsetneq \cdots$, where $\supsetneq$ denotes a strict superset. At step t, the oracle estimates the false discovery proportion of the current rejection set as

$$\widehat{\mathrm{FDP}}_t = \frac{h(1) + \sum_{i\in\mathcal{R}_t} h(p_i)}{1 + |\mathcal{R}_t|}, \tag{1}$$

where $h : [0, 1] \to [0, \infty)$ is nondecreasing and bounded, with $\int_0^1 h(p)\,\mathrm{d}p = 1$, for example $h(p) = 2 \times \mathbb{1}\{p \geqslant 0.5\}$. The function h is called an accumulation function, and the estimator (1) is based on the accumulation test of Li & Barber (2016), itself a generalization of procedures proposed by Barber & Candès (2015) and G'Sell et al. (2016). Informally, $\sum_{i\in\mathcal{R}_t} h(p_i)$ plays the role of estimating $V_t = |\mathcal{R}_t \cap \mathcal{H}_0|$.

To allow for the analyst to make data-dependent choices without inflating Type I error, we generalize a technique due to Lei & Fithian (2018) called masking. Specifically, for any choice of accumulation function h as described above, we show how to derive a masking function g, constructed so that $h(p)$ is mean-independent of $g(p)$ when $p \sim \mathrm{Un}[0, 1]$:

$$E_{u\sim U[0,1]}\{h(u) \mid g(u)\} = E_{u\sim \mathrm{Un}[0,1]}\{h(u)\} = 1 \tag{2}$$

almost surely. For example, if we choose $h(p) = 2 \times \mathbb{1}\{p \geqslant 0.5\}$, then we may choose $g(p) = \min\{p, 1 - p\}$ by observing that, for a null uniform p-value p, we have $E\{h(p) \mid g(p)\} = E\{h(p)\} = 1$. In § 2.3 we describe a general recipe for constructing such a function g. Even though the masking function g is constructed using condition (2), we will show next that when null p-values are not exactly uniform, the same (g, h) pairs satisfy a more general property that will be crucial in the proof of false discovery rate control. The proof is presented in the Supplementary Material.

Proposition 1. *If the density of a null p-value p is nondecreasing and the functions h and g are chosen such that condition* (2) *holds, then*

$$E\{h(p) \mid g(p)\} \geqslant 1$$

almost surely.

The selectively traversed accumulation rules protocol works by first revealing all masked p-values $g(p_i)$ to the analyst. Then, as the analyst shrinks the rejection set, p-values that can no longer be rejected get unmasked, meaning that the analyst observes all of the p-values p_i for $i \notin \mathcal{R}_t$. Revealing $g(p)$ to the analyst is an example of data carving (Fithian et al., 2017), where part of a random variable, $g(p_i)$, is used for selection while the remainder, $h(p_i)$, is used for inference. Because these two views of the data are designed to be orthogonal to each other under the null, the masked p-values and covariates, along with prior information, together provide guidance to the analyst on how to adaptively shrink the rejection set. Unlike other approaches (e.g., Dwork et al., 2015), masking does not introduce extra randomness. This is desirable in scientific research as it prevents cheating by specifying a favourable random seed.

At each step t, the oracle reports $\widehat{\text{FDP}}_t$ to the analyst. If $\widehat{\text{FDP}}_t \leqslant \alpha$, then the entire procedure halts and $\mathcal{R}_t$ is rejected. Otherwise, the analyst is responsible for selecting a smaller rejection set $\mathcal{R}_{t+1} \subsetneq \mathcal{R}_t$ using covariates, intuition and any desired statistical model or procedure, along with the oracle's revealed information, subject to the constraint that $\mathcal{R}_{t+1} \in \mathcal{K}$. After the analyst chooses $\mathcal{R}_{t+1}$, the oracle re-estimates the false discovery proportion and reveals p_i for $i \in \mathcal{R}_t \setminus \mathcal{R}_{t+1}$. The analyst may then update their model, prior, constraints or intuition, and the process repeats until $\widehat{\text{FDP}}_t \leqslant \alpha$.

The update rule from $\mathcal{R}_t$ to $\mathcal{R}_{t+1}$ is a user-specified subroutine, and should exploit only the information contained in the σ-field representing all information the analyst is allowed to observe by time t:

$$\mathcal{F}_t = \sigma\left[\{x_i, g(p_i)\}_{i=1}^{n},\ (p_i)_{i\notin\mathcal{R}_t},\ \sum_{i\in\mathcal{R}_t} h(p_i)\right]. \tag{3}$$

For instance, at step $t = 0$, the analyst is free to use $\{x_i, g(p_i)\}_{i=1}^{n}$ to train an arbitrary model to estimate which hypotheses are most likely to be true, and choose $\mathcal{R}_1$ by eliminating the most likely hypothesis subject to the constraint that $\mathcal{R}_1 \in \mathcal{K}$. We will discuss specific rules tailored to different problems in § 4 and § 5. For now, the update rule is any subroutine that produces $\mathcal{R}_{t+1} \subsetneq \mathcal{R}_t$ such that $\mathcal{R}_{t+1} \in \mathcal{K}$ and with its outcome $\mathcal{F}_t$-measurable. Because $\mathcal{R}_t$ shrinks with each step, revealing more information to the analyst, the σ-fields form a filtration with $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. This can easily be proved by induction; see the Supplementary Material for details. As a consequence, the information available to the analyst accrues over time. Algorithm 1 summarizes the procedure.

Algorithm 1. Selectively traversed accumulation rules.

```
Input: Predictors and $p$-values $(x_i, p_i)_{i=1}^{n}$, constraint set $\mathcal{K}$, target false discovery rate level $\alpha$
$\mathcal{R}_0 = [n]$
While $\mathcal{R}_t \neq \emptyset$
    $\widehat{\text{FDP}}_t \leftarrow \frac{1}{1+|\mathcal{R}_t|}\left\{h(1) + \sum_{i\in\mathcal{R}_t} h(p_i)\right\}$
    If $\widehat{\text{FDP}}_t \leqslant \alpha$ or $\mathcal{R}_t = \emptyset$
```

```
        Stop and return $\mathcal{R}_t$, and reject $\{H_i : i \in \mathcal{R}_t\}$
    Select $\mathcal{F}_t$-measurable $\mathcal{R}_{t+1} \subsetneq \mathcal{R}_t$ with $\mathcal{R}_{t+1} \in \mathcal{K}$
Output: $\mathcal{R}_t$ as the rejection set
```

Remark 1. In some applications, the analyst may want to choose the structural constraint $\mathcal{K}$ after looking at the data. Indeed, selectively traversed accumulation rules allow $\mathcal{K}$ to be chosen based on $\mathcal{F}_0$ and even to vary with t. Likewise, there is no requirement that $\mathcal{R}_t \in \mathcal{K}$ for every t, provided we impose $\mathcal{R}_t \in \mathcal{K}$ as an additional condition for stopping the algorithm. We discuss these details in the Supplementary Material.

The mechanism used by selectively traversed accumulation rules to enforce structural constraints is different from that for learning structural information. The former is based on the fact that $\mathcal{R}_{t+1}$ can be updated without relying on the size of the p-values, in contrast to Benjamini–Hochberg-type algorithms, whereas the latter is enabled by masking functions, which provide a partial view of the data without sacrificing validity.

Selectively traversed accumulation rules can be viewed as a generalization of the ordered multiple testing framework of Li & Barber (2016), in which a full preordering of hypotheses based on outside data or prior knowledge must be supplied as input to the analysis; with a low-quality preordering the method may be powerless, as shown in Li & Barber (2016) and Lei & Fithian (2016). By contrast, for our method, a preordering is just one potential source of side information that may or may not be available in any given case; other possible kinds of side information include spatial structure, covariate information, and a partial ordering from a directed acyclic graph. Our method then determines a data-adaptive ordering based on interactive guidance from the scientist or an algorithm acting on their behalf. This interactive ordering respects any constraints the analyst has imposed, and uses both the side information and the masked p-values $g(p_i)$. A counterpart of accumulation tests' preordering, the interactive ordering is simply the order in which the scientist or algorithm decides to peel off unpromising hypotheses from the candidate rejection set, based on masked p-values and prior information. If the ordering is prespecified before seeing the data and masked p-values are ignored entirely, and if there is no interaction, then our method reduces to an accumulation test with H_n peeled off first, then H_{n-1}, and so on.

The flexibility of our method comes from carving the p-value into two parts, $g(p)$ and $h(p)$, with the first part used to adaptively determine the ordering and the second part used to control the false discovery rate. Our use of the masked p-values is a selective-inference free lunch: compared with accumulation tests, we are using the same $h(p_i)$ values in the same way to estimate the false discovery proportion; we have brought more information $g(p_i)$ to bear on guiding our method without inflating the false discovery rate or requiring any further correction. In other words, accumulation tests can be thought of as also calculating $g(p_i)$ values and then simply discarding them. On the other hand, the masked p-values can be highly informative about which hypotheses are nonnull. For example, observing that $g(p_i) = \min\{p_i, 1-p_i\} = 10^{-8}$ is extremely suggestive of the following: (i) H_i is most likely false; (ii) $h(p_i) = 2 \times \mathbb{1}(p_i > 0.5)$ is most likely 0; and (iii) possibly other hypotheses H_j with spatial locations or covariate values close to those of H_i are more likely to be false as well and to have $h(p_j) \approx 0$. More generally, by examining in aggregate all of the masked p-values, as well as most of the unmasked ones in later stages of the procedure, we can learn areas of the covariate space with many nonnull hypotheses, and focus the power on those by placing them nearer the front of the list, that is, by peeling them off last. When the true signals do not satisfy the constraint, the enforced constraint narrows down potential rejection sets, and may affect power negatively compared with algorithms which do not enforce the constraint; however, the data-adaptive exploration made possible by masking allows users to learn the structure that could compensate for the power loss at no cost of false discoveries. In fact, the theory and algorithm would be exactly identical if the prespecified $\mathcal{K}$ is replaced by a data-dependent constraint $\mathcal{K}_t$, as long as $\mathcal{K}_t$ is predictable, i.e., $\mathcal{F}_{t-1}$-measurable.

In the Supplementary Material we conduct an asymptotic analysis under the slightly more general framework of Li & Barber (2016) to quantify the benefit of using masking functions. In a nutshell, in the absence of an informative preordering, the accumulation test is powerless while the masking functions have some power in most practical cases. Moreover, even when an informative preordering is available, one can still improve the power further by combining the preordering with the masked p-values to obtain an even better ordering.

2.2. *False discovery rate control*

Intuitively, the quantity $\sum_{i\in\mathcal{R}_t} h(p_i)$ is a conservative estimator of $V_t = |\mathcal{H}_0 \cap \mathcal{R}_t|$, the number of false rejections we would incur if we rejected the set $\mathcal{R}_t$: each null hypothesis in $\mathcal{R}_t$ contributes at least 1 in expectation, and the nonnull hypotheses contribute a nonnegative amount. Hence $|\mathcal{R}_t|^{-1}\sum_{i\in\mathcal{R}_t} h(p_i)$ can be interpreted as an upwardly biased estimator of the false discovery proportion of rejection set $\mathcal{R}_t$. With the correction term in (1), it can be proved that Algorithm 1 controls the false discovery rate in finite samples; however, the analyst may update the rejection set $\mathcal{R}_t$ under certain regularity conditions on the joint distribution of p-values.

Theorem 1. *Assume that the following hold:*

(i) *the null p-values $(p_i)_{i\in\mathcal{H}_0}$ are mutually independent, and are independent of the nonnulls $(p_i)_{i\notin\mathcal{H}_0}$, conditional on the covariates $(x_i)_{i=1}^n$;*

(ii) *each null p-value has a nondecreasing density, which may differ across p-values.*

If the accumulation function h is nondecreasing and the masking function g satisfies condition (2), then, conditional on $\{x_i, g(p_i)\}_{i=1}^n$, the selectively traversed accumulation rules control the false discovery rate at level α. Hence, unconditionally, the set $\mathcal{R}_\tau$ chosen using Algorithm 1 satisfies $E\{\text{FDP}(\mathcal{R}_\tau)\} \leqslant \alpha$.

The assumption (i) is common in the multiple testing literature; (ii) strengthens the usual assumption that a null p-value is stochastically larger than uniform, but it is significantly weaker than assuming exact uniformity. It also strengthens the mirror-conservatism proposed in Lei & Fithian (2018). This theorem is proved using a martingale argument, which is elaborated in the Supplementary Material.

Remark 2. Following Li & Barber (2016), we could allow h to be unbounded, replace $h(1)$ in (1) with a constant $C > 0$, and halt when $\widehat{\text{FDP}}_t$ is below a corrected level:

$$\frac{C + \sum_{i\in\mathcal{R}_t} h(p_i)}{1 + |\mathcal{R}_t|} \leqslant \alpha \int_0^1 \{h(p) \wedge C\}\, \mathrm{d}p.$$

However, one can show that replacing such an unbounded accumulation function by its bounded counterpart $h^C(p) = \{h(p) \wedge C\}/\int_0^1 \{h(p) \wedge C\}\, \mathrm{d}p$ results in a strictly more powerful procedure. Hence, we assume without loss of generality that h is bounded.

2.3. *Masking functions*

We now give a recipe for constructing a masking function g for a generic accumulation rule h.

THEOREM 2. *Let $h : [0,1] \to [0,\infty)$ be any nonnegative nondecreasing function with $\int_0^1 h(p)\,\mathrm{d}p = 1$, and let*

$$H(p) = \int_0^p \{h(x) - 1\}\,\mathrm{d}x.$$

Then we have the following two conclusions.

(i) *There exists a continuous and strictly decreasing function $s(p)$ which is also differentiable, except on a set of measure zero, i.e., a Lebesgue null set, such that*

$$H\{s(p)\} = H(p), \quad s(0) = 1,\ s(1) = 0.$$

(ii) *The masking function $g(p) = \min\{p, s(p)\}$ satisfies condition* (2).

This theorem is proved in the Supplementary Material. The aforementioned function $s(p)$ can be obtained numerically by a quick binary search whenever $H(\cdot)$ can be computed efficiently. All accumulation functions in this paper satisfy the conditions of the above theorem, and so their associated $s(p)$ functions are almost surely differentiable. Letting p_* denote the unique solution to $s(p_*) = p_*$, for any $q < p_*$, the set $\{p : g(p) = q\}$ contains exactly two points, and hence such a $g(p)$ is very informative because it only masks one bit of information. For example, if $h(p) = 2 \times \mathbb{1}(p \geqslant 0.5)$, it is easy to show from Theorem 2 that $g(p) = \min\{p, 1-p\}$. Then $g(p) = 0.01$ implies that $p = 0.01$ or $p = 0.99$, and the analyst just needs to make a guess from two candidate values.

For brevity, throughout the main article we focus on the simple accumulation function $h(p) = 2 \times \mathbb{1}(p \geqslant 0.5)$ with masking function $g(p) = \min\{p, 1-p\}$; we have found that this simple choice works reasonably well in all the settings we tried. We provide several other examples and examine their performance in the Supplementary Material.

3. IMPLEMENTATION

3.1. *Guidance on updating rejection sets*

Theorem 1 showed that the false discovery rate is controlled no matter how the analyst chooses $\mathcal{R}_{t+1}$ based on $\mathcal{F}_t$; however, having a good update rule will be vital to implementing selectively traversed accumulation rules in any given context. Still, we recommend that any human-in-the-loop interactions be grounded in principled data analysis. For example, our method is to some degree susceptible to the same free-rider dynamic as other data-adaptive multiple testing procedures such as AdaPT and knockoffs; namely, if we find 100 strong signals in one part of the dataset, we might be able to add two or three more favourite hypotheses from elsewhere without threatening false discovery rate control. Well-chosen and principled constraints on the rejection set can serve as a safeguard against this behaviour, which is epistemically problematic even if not formally disallowed. In addition, too-frequent looks at the data may tempt users to go back and modify their previous rejection sets; such actions are prohibited by Theorem 1 and would break the theoretical guarantee.

In general, we can describe an update rule in three steps: (i) find all candidate sets of hypotheses that we can peel off from $\mathcal{R}_t$ without leaving the constraint set $\mathcal{K}$; (ii) compute a score using all the information in $\mathcal{F}_t$ that measures the likelihood of each candidate set having nonnulls; and (iii) delete the candidate set with the worst score. The Supplementary Material provides a flowchart to summarize the pipeline schematically. The inclusion of candidate sets marks the fundamental difference between selectively traversed accumulation rules and adaptive p-value thresholding, because the essential candidate sets of the latter are simply all remaining hypotheses, while the former operates in a more greedy way.

As a concrete example, suppose that the hypotheses correspond to vertices of a tree and one aims to detect a rooted subtree of signals. Then the deletion candidates are all hypotheses on leaf nodes of the subtree given by $\mathcal{R}_t$, because deletion of any leaf node does not change the rooted subtree structure of the rejection set.

The next step is to compute a score for each candidate. Heuristically, the score should be highly correlated with the p-values. As discussed in § 2, the most straightforward score for the ith hypothesis is $g(p_i)$; we refer to it as the canonical score. For a candidate set that contains multiple hypotheses, we define the canonical score as the average of the $g(p_i)$ values. A larger canonical score provides stronger evidence that the candidate set is mostly null. Although the canonical score is straightforward to use, the user is allowed to fit any model using the covariates and the partially masked p-values. Thus, one can estimate the signal strength, or a posterior probability of being null, as the score; we call this a model-assisted score.

Finally, given the score, it is natural to remove the least favourable candidate. For instance, when using the canonical score, the candidate with the largest score will be removed. On the other hand, if the score measures the signal strength or the likelihood of being nonnull, the candidate with the smallest score will be removed.

Our principle here can be summarized as follows: given a working model or belief about the data-generating process, we have a generic expectation-maximization-based pipeline that incorporates the model into our method to produce scores yielding the adaptive ordering. The researcher can always incorporate their favourite data-generating model into our method, improving the power if their model is good, but never violating the false discovery rate control if their model is inaccurate.

3.2. *Conditional one-group model as the working model*

Although the canonical scores are effective in many situations, as will be shown in § 5, they do not fully exploit covariate information, apart from enforcing the constraint. For instance, when the hypotheses are arranged spatially, one might expect the nonnulls to concentrate on a few clusters, or that the signal strength will be smooth on the underlying space. Such prior knowledge may be neither reflected directly from the p-values nor used explicitly to strengthen the p-values; instead, one can use a working model to assist in calculating the scores. We emphasize that no matter how misspecified the working model is, the false discovery rate will still be controlled.

Lei & Fithian (2018) proposed a conditional two-group model and used the estimated local false discovery rates as model-assisted scores. They proved that the local false discovery rate gives the optimal score for ordering hypotheses. The model can be fitted by an expectation-maximization algorithm (Lei & Fithian, 2018, Appendix A).

Despite the approximate optimality of the above approach, the expectation-maximization algorithm for conditional two-group models is computationally intensive, because it fits two separate models for the proportion and the signal strength of nonnulls at each iteration. Additionally, Lei & Fithian (2018, Appendix A.3) pointed out some instability issues of the algorithm. For these reasons, we propose a conditional one-group model as an alternative:

$$p_i \sim f(p; \mu_i), \quad \eta(\mu_i) = \beta'\phi(x_i), \tag{4}$$

where f is the density function, η is the link function and ϕ is an arbitrary featurization. As will be detailed in the Supplementary Material, μ_i can be estimated via an expectation-maximization algorithm. The model-assisted scores are then given by the estimates $\hat{\mu}_i$. Although the $\hat{\mu}_i$ lose the optimality guarantee, they provide a good proxy for how promising each hypothesis is. On the other hand, (4) is easier to fit as it involves only one set of parameters. At step t, the algorithm only needs to impute the masked p-values. Thus it is computationally more efficient and partially solves the issue raised in Lei & Fithian (2018).

Finally, as discussed in Lei & Fithian (2018), p-values may not be the objects most amenable to modelling, in which case one can either model transformed p-values or directly model the data used to produce the p-values. For example, in many applications z-values are available and one-sided p-values are obtained via the transformation $p_i = 1 - \Phi(z_i)$, where Φ is the distribution function of a standard Gaussian. In this case, we can directly model z_i as $z_i \sim N(\mu_i, 1)$.

4. Example 1: convex region detection

4.1. *Problem set-up*

In some applications, the p-values may be associated with features $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ that encode some contextual information, such as predictor variables or a spatial location, and which may be associated with the underlying signal. We may wish to use this feature information to discover regions of the feature space where the signal is relatively strong; for example, Drevelegas (2010) seeks a convex region to locate the boundary of tumours.

As a concrete mathematical example, suppose that for each point x_i on a regular spatial grid we observe an independent observation $z_i \sim N\{f(x_i), 1\}$ for some nonnegative function f, and we aim to discover the region $\mathcal{C} = \{x : f(x) > 0\}$, where we have some prior belief that $\mathcal{C}$ is convex; for example, if f is known to be a concave function, then $\mathcal{C}$ is a super-level set and is hence convex. Since we cannot expect to find $\mathcal{C}$ perfectly, we may hope to discover a smaller region $\hat{\mathcal{C}}$ that is mostly contained within $\mathcal{C}$, i.e.,

$$\frac{\mathrm{Vol}(\hat{\mathcal{C}} \cap \mathcal{C}^{\mathrm{c}})}{\mathrm{Vol}(\hat{\mathcal{C}})} \leqslant \alpha. \tag{5}$$

We can frame the above as a multiple testing problem by computing a one-sided p-value $p_i = 1 - \Phi(z_i)$ for each H_i, and constraining the rejection set to be of the form $\mathcal{R} = \{i : x_i \in \hat{\mathcal{C}}\}$ for some convex set $\hat{\mathcal{C}}$, leading to a constraint $\mathcal{K}$ on the allowable rejection sets. If the grid is relatively fine, then the false discovery rate is a natural error criterion to control since the false discovery proportion of $\mathcal{R}$ approximates criterion (5).

As another example, from supervised learning, suppose that for $i = 1, \ldots, n$ we observe a pair consisting of a feature x_i and a response y_i, and we want to find a subregion of the feature space $\mathcal{X}$ where the y values tend to be relatively large. In bump-hunting (Friedman & Fisher, 1999) one seeks to discover a rectangle in predictor space; see the Supplementary Material for further discussion.

More generally, we may want to discover a set of hypotheses $\mathcal{R} = \{i : i \in \hat{\mathcal{C}}\}$, where $\hat{\mathcal{C}}$ is convex, or is a rectangle, or satisfies some other geometric property. To the best of our knowledge, no existing procedure can solve such problems while guaranteeing false discovery rate control. To achieve these goals using the selectively traversed accumulation rules framework, we need only to define the procedure-specific functions elaborated in § 3. To illustrate, we first consider an example where $x_i \in \mathbb{R}^2$.

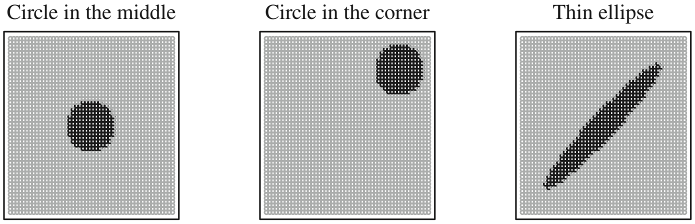

Fig. 1. True underlying signal for our convex region detection simulation. Each of 2500 points represents a hypothesis; black indicates nonnulls with $\mu = 2$.

4.2. *Procedure*

To preserve convexity, we consider an automated procedure that gradually peels off the boundaries of the point cloud $\{x_i\}$. At each iteration, we choose a direction $\theta \in [0, 2\pi)$ and a small constant δ, and peel off a proportion δ of points that are farthest in this direction.

Specifically, for each angle θ we define a candidate set $C(\theta; \delta)$ to be observed as the set of indices corresponding to the δ-proportion of points that are farthest in the direction θ.

If the goal is to detect an axis-parallel box, θ can be restricted to take only values in $\left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$; otherwise we set θ to an equispaced grid on $[0, 2\pi]$ of length 100. Given a score S_i for each hypothesis, which we will soon define, we may evaluate the signal strength of each candidate set as the average of the S_i. Then we update $\mathcal{R}_t$ as

$$\mathcal{R}_t := \mathcal{R}_{t-1} \setminus C(\hat{\theta}; \delta), \quad \hat{\theta} := \arg\min_{\theta} \operatorname{Avg}\{S_i : i \in C(\theta; \delta)\}.$$

Finally, to define the scores S_i, we can directly use the canonical score $g(p_i)$. However, in most problems of this type, where x_i represents the location in some continuous space, it is reasonable to assume that the distributions of p-values are smoothly varying. In particular, we use the conditional one-group model on p-values with the beta distribution, $p_i \sim \mathrm{Be}(1/\mu_i, 1)$, as the working model and fit a generalized additive model (Hastie & Tibshirani, 1990) using the smooth spline basis of x_i as the featurization ϕ defined in (4). This working model is motivated by the conditional two-group gamma generalized linear model in Lei & Fithian (2018).

4.3. *Simulation results*

We consider an artificial dataset, where the predictors form an equispaced 50×50 grid in the region $[-100, 100] \times [-100, 100]$. Let $\mathcal{C}_0$ be a convex set on $\mathbb{R}^2$ and let $\mathcal{H}_0^c = \{x_i : x_i \in \mathcal{C}_0\}$. We generate independent and identically distributed p-values from a one-sided normal test, i.e.,

$$p_i = 1 - \Phi(z_i), \quad z_i \sim N(\mu, 1), \tag{6}$$

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. 1 displays three types of $\mathcal{C}_0$ that we conduct tests on.

Although our procedure is the only one that can enforce the convexity, it is still illuminating to compare it with other procedures and assess the power. In particular, we consider the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) and adaptive p-value thresholding (Lei & Fithian, 2018). We implement selectively traversed accumulation rules with the model-assisted score based on the generalized additive model.

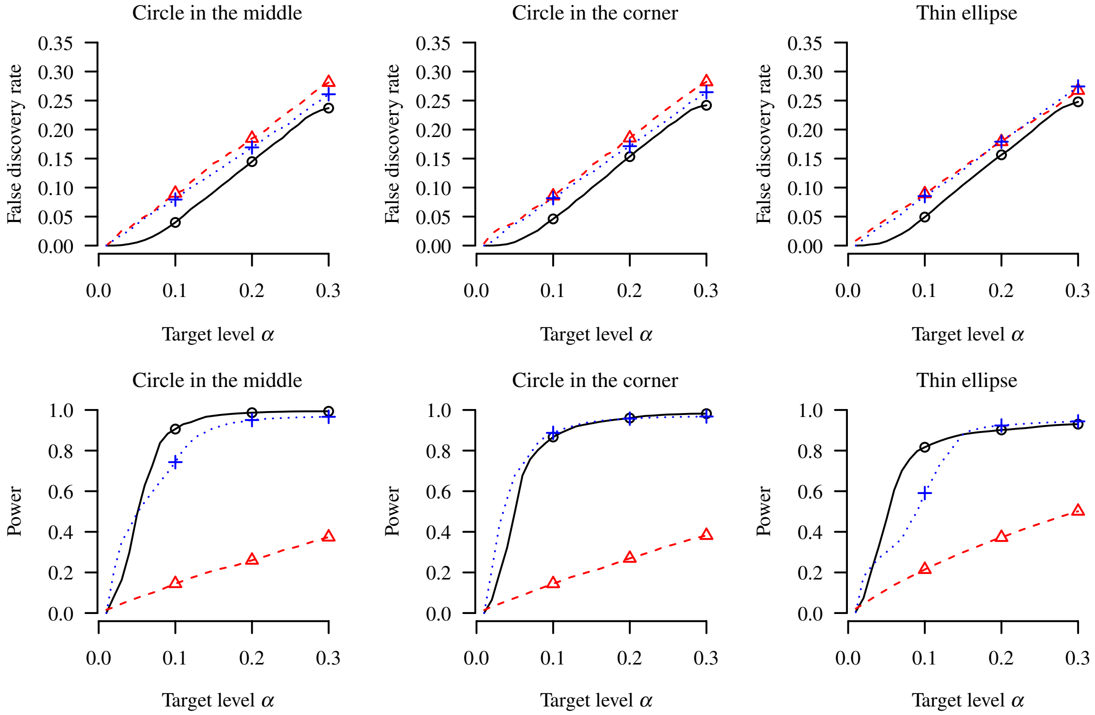

Fig. 2. Comparison of our method (black solid) with the Benjamini–Hochberg procedure (red dashed) and adaptive p-value thresholding (blue dotted) for convex region detection.

For each procedure and level α, we calculate the false discovery proportion and the power,

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

where $\mathcal{R}(\alpha)$ is the rejection set at level α. We then estimate the false discovery rate and the power by averaging FDP(α) and power(α) over 100 sets of independently generated p-values. The results are plotted in Fig. 2 for a range of α values from 0.01 to 0.3. We see that our method is comparable to adaptive p-value thresholding and more powerful than the Benjamini–Hochberg procedure, neither of which enforces the convexity constraint. Furthermore, it is worth mentioning that the model-assisted selectively traversed accumulation rules achieve high power despite using a generic and misspecified generalized additive beta model for the p-values.

The power gain over the Benjamini–Hochberg procedure is due in part to the fact that the convexity constraint reflects the truth and our method implicitly builds it into the selection. It also comes from our method's effective learning of the underlying spatial structure. To examine this, we plot $\hat{\mu}(x)$ in Fig. 3. The top panels show the initial score, which uses only the partially masked p-values $g(p_i)$ and x_i. The bottom panels show the oracle result when fitting the model to fully observed p-values. Surprisingly, even the initial estimate is good enough to clearly show the contour of the nonnulls, and it is nearly as good as the final estimate when using fully observed p-values; in fact, the correlation between the two estimates is above 0.98 in all three cases. This explains why our method can accurately pinpoint the nonnulls and hence increase the power.

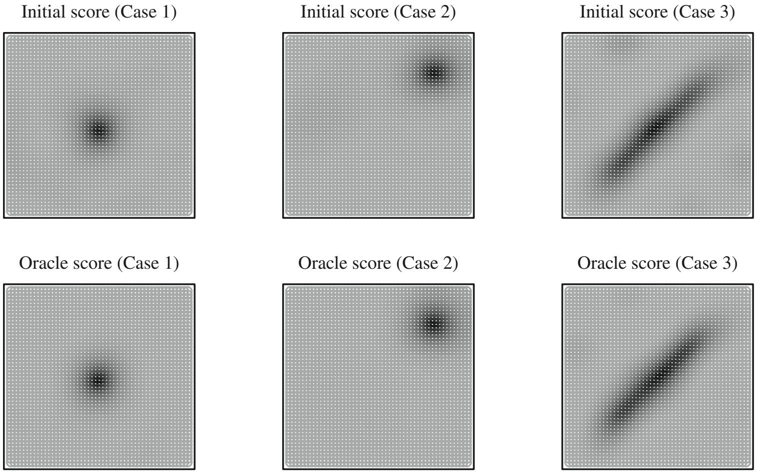

Fig. 3. Model-assisted scores of selectively traversed accumulation rules for circle in the middle (Case 1), circle in the corner (Case 2) and thin ellipse (Case 3); darker pixels represent higher scores.

5. Example 2: testing on directed acyclic graphs

5.1. *Set-up and procedure*

In another scenario, hypotheses are arranged on a directed acyclic graph, and the nonnulls are known a priori to satisfy the heredity principle. The strong or weak heredity principle, also known as the effect hierarchical principle (Wu & Hamada, 2000), states that an effect is significant only if, respectively, all of its parent effects are significant or one of its parent effects is significant. An application to variable selection in factorial experiments under the heredity principle is discussed in the Supplementary Material.

Multiple testing on directed acyclic graphs has extensive applications in genomics (e.g., (Goeman & Mansmann, 2008; Saunders et al., 2014; Meijer & Goeman, 2015)) and clinical trials (e.g., Dmitrienko & Tamhane, 2013). However, most prior work deals with familywise error rate control, and the setting of false discovery rate control is relatively under-studied. To the best of our knowledge, the only existing false discovery rate control procedures for directed acyclic graphs are those proposed in G. Lynch's 2014 PhD thesis from the New Jersey Institute of Technology, and in the sequential setting by Ramdas et al. (2019a) and the multi-layer setting by Ramdas et al. (2019b). All the aforementioned works concern the strong heredity principle.

It is straightforward to use our method to guarantee false discovery rate control under both the strong and the weak heredity principles. For the strong heredity principle, we select the candidates to be all the leaf nodes. For the weak heredity principle, we select the candidates to be all nodes such that after removing them the remaining graph satisfies the principle. We also present an application to a factorial experiment in the Supplementary Material.

5.2. *Simulation results*

We focus on the strong heredity principle to compare selectively traversed accumulation rules with existing methods. To account for the structure, we consider three types of directed acyclic graphs: a shallow regular graph, with four layers and 250 nodes in each layer; a deep regular graph, with 10 layers and 100 nodes in each layer; and a triangular graph, with five layers having 50, 100, 200, 300 and 350 nodes each. For each graph, we take 50 nodes satisfying the strong heredity principle to be nonnull and generate p-values using (6) with $\mu = 2$ for nonnulls. The settings are illustrated in Fig. 4.

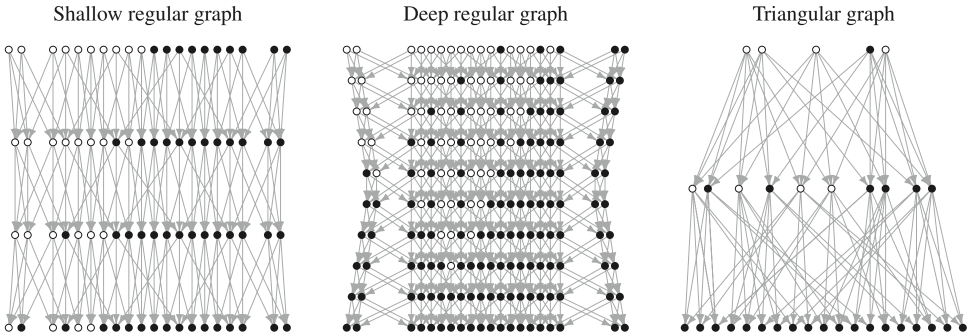

Fig. 4. Three types of graphs used in our simulation studies, with fewer nodes shown for the sake of clarity. Each of 1000 nodes represents a hypothesis, with black ones being the nulls.

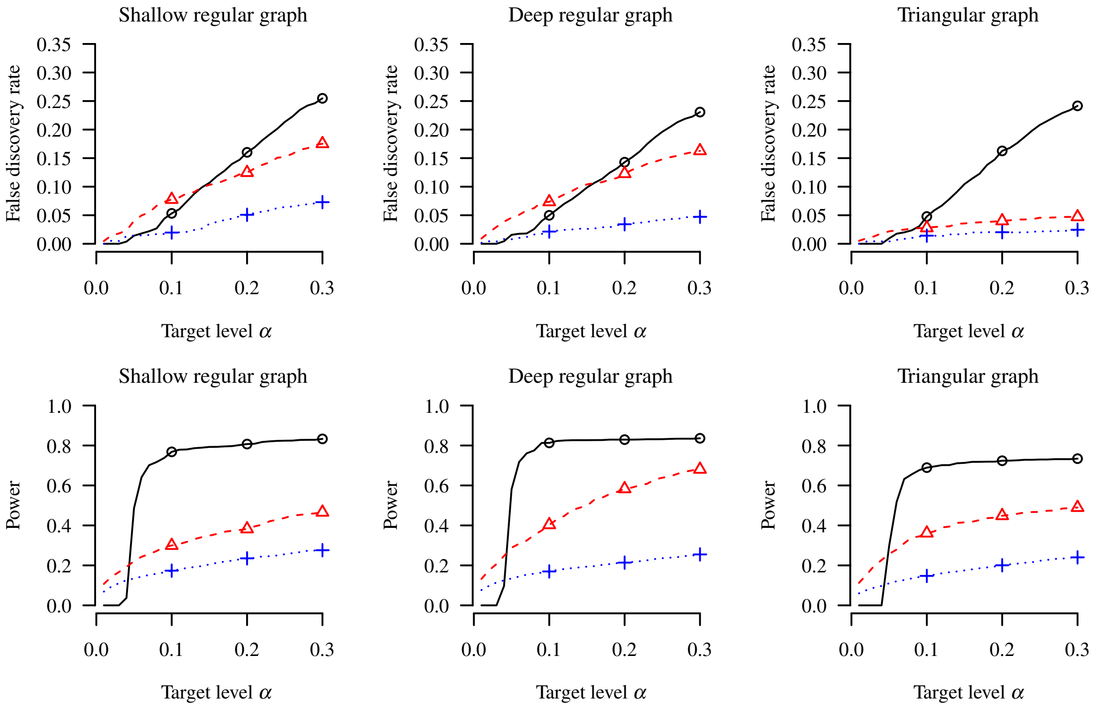

Fig. 5. Comparison of our method (black solid) with the methods of Lynch & Guo (2016) (red dashed) and Ramdas et al. (2019a) (blue dotted) for testing on directed acyclic graphs.

We compare our method using canonical scores, as described in § 3.1, with the method of G. Lynch, referred to as the self-consistent rejection procedure, and a recently developed sequential method, referred to as greedily evolving rejections, which is a generalization of Lynch's hierarchical test proposed by Ramdas et al. (2019a). The results are plotted in Fig. 5. It is clear that in all cases our method is more powerful than the others when α is not too small. When α is small, our method has no power because of the finite-sample correction constant $h(1)$ in (1), which requires at least $h(1)/\alpha - 1$ rejections to get a nonempty rejection set at level α. In our case $h(1) = 2$, but $h(1)$ can be reduced by choosing other accumulation functions. A few examples that resolve this issue are provided in the Supplementary Material.

6. Discussion

This article has focused on the case where the test statistics are independent. In more general settings, our framework dovetails naturally with the knockoff framework proposed by Barber & Candès (2015) and extended by Candès et al. (2018), converting complex regression problems into independent one-bit p-values for each variable. When running sparse regression algorithms, the underlying variables often have some structure, such as a tree structure with wavelet coefficients in compressed sensing, and one may want to use the knockoff procedure to select a structured subset of variables.

With this motivation in mind, let W_i denote the knockoff statistic for hypothesis i and define $p_i = (1+\mathbb{1}\{W_i > 0\})/2$. Then knockoff constructions guarantee that the $(p_i)_{i\in\mathcal{H}_0}$ are independent p-values, conditional on $(|W_i|)_{i=1}^n$ and $(p_i)_{i\notin\mathcal{H}_0}$. The absolute value $|W_i|$ could be viewed as the free side information $g(p_i)$ in Algorithm 1, while the location of the variable on the tree would be the structural information x_i. Although the constructed p_i does not technically have a decreasing density, the accumulation function $h(p_i) = 2 \times \mathbb{1}\{p_i > 1/2\} = \mathrm{sign}(W_i) + 1$ nevertheless has expectation 1 under the null. The analyst can then use selectively traversed accumulation rules to interactively pick a structured subset of variables.

Like adaptive p-value thresholding, knockoffs as defined in Barber & Candès (2015) do not enforce constraints on the rejection set; nor do they allow interactive adaptation as more p-values are unmasked. Combining our method with knockoff statistics enables control of the false discovery rate in many interesting regression problems with constraints on the rejection set, such as hierarchy constraints for interactions in a regression.

When the underlying problem is a sparse regression problem, the aforementioned knockoff procedure can be used to construct independent p-values even when the covariates are not orthogonal. In most settings it is not yet known how to construct knockoffs, but often one can still construct dependent p-values, and an important open problem is to provide guarantees in such settings. As shown in the Supplementary Material, the experiments under dependence yield encouraging results, especially under negative dependence, but we do not currently know how to prove any results on robustness with respect to deviations from independence. Such results would immediately be applicable to several other settings, such as ordered testing (Li & Barber, 2016) and knockoffs (Barber & Candès, 2015).

Supplementary material

Supplementary material available at *Biometrika* online includes all proofs, algorithmic details, other applications, comparison of different masking functions, a sensitivity analysis under dependent p-values, and an asymptotic power analysis. The `R` code used to produce all results in the paper is available at `https://github.com/lihualei71/STAR`.

References

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Candès, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: ‘Model-X’ knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B **80**, 551–77.

Dmitrienko, A. & Tamhane, A. C. (2013). General theory of mixture procedures for gatekeeping. *Biomet. J.* **55**, 402–19.

Drevelegas, A. (2010). *Imaging of Brain Tumors with Histological Correlations*. New York: Springer.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science* **349**, 636–8.

Fithian, W., Sun, D. & Taylor, J. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.

Friedman, J. H. & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statist. Comp.* **9**, 123–43.

Goeman, J. J. & Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24**, 537–44.

G'Sell, M. G., Wager, S., Chouldechova, A. & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Statist. Soc.* B **78**, 423–44.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Boca Raton, Florida: Chapman & Hall/CRC.

Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Meth.* **13**, 577–80.

Lei, L. & Fithian, W. (2016). Power of ordered hypothesis testing. In *Proc. 33rd Int. Conf. Machine Learning (ICML)*. JMLR W&CP **48**, pp. 2924–32.

Lei, L. & Fithian, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Statist. Soc.* B **80**, 649–79.

Li, A. & Barber, R. F. (2016). Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Statist. Assoc.* **112**, 1–38.

Li, A. & Barber, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Statist. Soc.* B **81**, 45–74.

Lynch, G. & Guo, W. (2016). On procedures controlling the FDR for testing hierarchically ordered hypotheses. *arXiv:* 1612.04467.

Meijer, R. J. & Goeman, J. J. (2015). A multiple testing method for hypotheses structured in a directed acyclic graph. *Biomet. J.* **57**, 123–43.

Ramdas, A., Chen, J., Wainwright, M. J. & Jordan, M. I. (2019a). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika* **106**, 69–86.

Ramdas, A. K., Barber, R. F., Wainwright, M. J. & Jordan, M. I. (2019b). A unified treatment of multiple testing with prior knowledge using the p-filter. *Ann. Statist.* **47**, 2790–821.

Saunders, G., Stevens, J. R. & Isom, S. C. (2014). A shortcut for multiple testing on the directed acyclic graph of gene ontology. *BMC Bioinformatics* **15**, 349.

Wu, J. & Hamada, M. (2000). *Experiments: Planning, Analysis, and Optimization*. Hoboken, New Jersey: Wiley.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *J. Am. Statist. Assoc.* **103**, 309–16.

[*Received on* 8 *June* 2019. *Editorial decision on* 15 *May* 2020]