ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 4, *pp.* 1247–1268

The *p*-filter: multilayer false discovery rate control for grouped hypotheses

Rina Foygel Barber

University of Chicago, USA

and Aaditya Ramdas

University of California at Berkeley, USA

[Received January 2016. Revised August 2016]

Summary. In many practical applications of multiple testing, there are natural ways to partition the hypotheses into groups by using the structural, spatial or temporal relatedness of the hypotheses, and this prior knowledge is not used in the classical Benjamini–Hochberg procedure for controlling the false discovery rate (FDR). When one can define (possibly several) such partitions, it may be desirable to control the *group FDR* simultaneously for all partitions (as special cases, the 'finest' partition divides the n hypotheses into n groups of one hypothesis each, and this corresponds to controlling the usual notion of FDR, whereas the 'coarsest' partition puts all n hypotheses into a single group, and this corresponds to testing the global null hypothesis). We introduce the *p-filter*, which takes as input a list of n p-values and $M \geqslant 1$ partitions of hypotheses, and produces as output a list of n or fewer discoveries such that the group FDR is provably *simultaneously* controlled for all partitions. Importantly, since the partitions are arbitrary, our procedure can also handle multiple partitions which are non-hierarchical. The p-filter generalizes two classical procedures—when $M = 1$, choosing the finest partition into n singletons, we exactly recover the Benjamini–Hochberg procedure, whereas, choosing instead the coarsest partition with a single group of size n, we exactly recover the Simes test for the global null hypothesis. We verify our findings with simulations that show how this technique can not only lead to the aforementioned multilayer FDR control but also lead to improved *precision* of rejected hypotheses. We present some illustrative results from an application to a neuroscience problem with functional magnetic resonance imaging data, where hypotheses are explicitly grouped according to predefined regions of interest in the brain, thus allowing the scientist to employ field-specific prior knowledge explicitly and flexibly.

Keywords: False discovery rate; Grouped hypotheses; Multilayer; Multilevel; Multiple testing; Multiresolution; *p*-filter

1. Introduction

One of the biggest concerns in the reproducibility crisis that is faced by modern data analysis is the practice of testing hundreds or thousands of hypotheses often arising from a single experiment. One of the earliest methods to gain some control on the number of false discoveries (null hypotheses that were incorrectly rejected by the scientist) is the Bonferroni correction, which controls the *familywise error rate*, which requires that the probability of making any false discoveries must be bounded by α. This procedure, which compares each p-value against the 'corrected' threshold α/n (where n is the number of hypotheses), is known to lead to extremely low power. Since then, a wide range of methods have been proposed as alternatives, such as

Address for correspondence: Rina Foygel Barber, Department of Statistics, University of Chicago, Room 108, 5734 South University Avenue, Chicago, IL 60637, USA.
E-mail: rina@uchicago.edu

 1369–7412/17/791247

the test by Simes (1986) for the 'global null' hypothesis (testing whether all hypotheses are null). Closely related to Simes's test, the most practically popular method is the procedure by Benjamini and Hochberg (BH) (1995) for controlling the *false discovery rate* (FDR).

We refer to 'true signals' to mean those tests for which the null hypothesis is actually false (and should be rejected), and 'nulls' or 'true nulls' to mean those tests with no real signal, where the null hypothesis is true (and should not be rejected). Our 'discoveries' are those tests which our method identifies as probably true signals (i.e. our algorithm's rejected null hypotheses). A false discovery is, of course, a false rejection: a null hypothesis that was rejected by our algorithm (proclaimed as a discovery) but is in fact a true null hypothesis.

We propose an algorithm called the *p*-filter, which is an elegant conceptual unification and generalization of the BH procedure and Simes's test for the global null, which is useful in practical scenarios when the scientist can naturally partition the hypotheses being tested into groups and desires to control both the *overall FDR* (controlling the number of falsely discovered hypotheses) and the *group FDR* (controlling the number of falsely discovered groups). We say that a group is falsely discovered if there is at least one hypothesis that is rejected within that group, but in reality the group consists entirely of nulls. Our procedure can also handle multiple partitions, which are referred to as 'layers', which *are not necessarily required to be hierarchical*; the *p*-filter provides FDR control simultaneously at the level of each specified layer. Practitioners may use prior knowledge to group hypotheses that they expect to be either simultaneously false or simultaneously true, or to organize the hypotheses according to some discipline-specific natural partitioning. At a high level, the *p*-filter works by filtering the groups in each partition, or layer, searching for groups that pass some threshold of evidence for a true signal; in the end, a hypothesis is rejected if and only if it passes through every layer of the filter.

Consider an example from neuroscience where controlling the FDR is both crucial and already popular since the early adoption popularized by Genovese *et al*. (2002). Consider showing a patient some stimulus and recording some physiological correlate of her brain activity (using, say, functional magnetic resonance imaging (MRI)). Suppose that we consider brain locations (voxels) $z_1, \ldots, z_V$, at times $t_1, \ldots, t_S$ after presentation of the stimulus, and formulate the following VS many null hypotheses:

$$H^0_{(v,s)}: \text{ the stimulus is independent of activity at } v, \text{ at delay } s \text{ after presentation.}$$

In addition to controlling the usual FDR by using the trivial partition (treating each (v, s) as its own group), we may want to ensure that the group FDR is *also* simultaneously small, where one may partition the hypotheses into voxels (grouping (v, s) for fixed v, across all delays s) and/or into time points (grouping (v, s) for fixed s, across all voxels v in some functional region). Note that, in this example, the three layers are not hierarchical—when we partition by space and by time, neither partition can be nested inside the other.

Another area where such groupings may be natural is bioinformatics or statistical genetics—when looking for associations between genes and proteins, it may make sense to group proteins with similar amino acid structure, and/or to group genes with similar nucleic acid sequences, perhaps employing prior knowledge from existing gene ontologies. We also expect our work to find favour in other spatiotemporal applications of the FDR, whenever rejected hypotheses are expected to be contiguous in space and/or time.

1.1. Related work

The nearest comparison with our method in the literature is the work of Benjamini and Bogomolov (BB) (2014) who proposed a hierarchical FDR control procedure, which was developed

further by Peterson *et al*. (2016). We return to this work in Section 3, where we discuss groupwise FDR control, and again in Section 6, where we compare our method with theirs conceptually and empirically. Yekutieli (2008) also considered the problem of testing hypotheses which are arranged in a hierarchy; this is of course related to simultaneously controlling groupwise and overall FDR. Our procedure is more general than both of these methods since, for the p-filter, the various layers or partitions are not required to form a hierarchy.

Other recent references have examined related questions. Here we briefly describe several such works, but many more exist in the literature. In the variable selection problem for a regression framework, Meinshausen (2008) considered hierarchical tests for handling clusters of highly correlated variables. A different setting also involving grouped hypotheses arose in Hu *et al*. (2010), where the goal was to control the overall FDR only, but different groups of hypotheses have different proportions of true signals *versus* nulls; by estimating these proportions for each group separately, their method increases the power to detect true signals in the high signal groups. Hypotheses may be grouped in a data-dependent or adaptive way in some applications; for example in spatial data where locally contiguous regions can form a 'cluster' of discoveries; the problem of controlling false discoveries at the cluster level was studied by Chouldechova (2014) and Sun *et al*. (2015).

1.2. Outline

In Section 2, we recall various standard definitions and the standard FDR procedure of Benjamini and Hochberg (1995). Next, we present our method; for clarity, we split our exposition into two parts. In Section 3, we show how to control the FDR simultaneously for individual hypotheses and at the group level, if our set of hypotheses is partitioned into groups. This leads into the more general setting of Section 4, where we develop the p-filter for controlling the FDR across an arbitrary number of (possibly non-nested) partitions, or layers. An algorithm for running the p-filter efficiently is given in Section 5. We then examine the empirical performance of our method on simulated data in Section 6 and on functional MRI data in Section 7. We give some concluding remarks in Section 8. Proofs for our theoretical results are deferred to Appendix A.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from either `http://www.stat.uchicago.edu/~rina/pfilter.html` or

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

2. Background

We assume that the reader is familiar with the classical set-up of frequentist hypothesis testing. In this paper, we assume that we are given a set of p-values, denoted by the vector $P \in [0,1]^n$, each corresponding to a different question (a different null hypothesis), and we wish to select some subset of these tests as our 'discoveries' (i.e. to reject some subset of the corresponding null hypotheses) while retaining some form of control over the number of false discoveries. For the remainder of the paper, let $\mathcal{H}^0 \subseteq [n]$ be the set of tests (hypotheses) designated as 'true nulls', and let $\hat{S}$ be the set that is selected as our discoveries on the basis of the observed p-values.

2.1. Benjamini–Hochberg procedure for false discovery rate control

For an algorithm that chooses a set of hypotheses to reject (denoted here by $\hat{S}$), the seminal paper by Benjamini and Hochberg (1995) proposed to measure its performance via the *FDR*, defined as

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \hat{S}|}{1 \vee |\hat{S}|}\right]$$

where $|\mathcal{H}^0 \cap \hat{S}|$ is the number of false discoveries (null hypotheses that are true and are incorrectly rejected) and $|\hat{S}|$ is the total number of discoveries (all hypotheses that are rejected). The notation $1 \vee |\hat{S}|$ in the denominator is defined as $\max\{1, |\hat{S}|\}$ and ensures that, if no rejections are made, then the false discovery proportion (FDP) is defined as 0.

Given the vector of p-values $P = (P_1, \ldots, P_n)$, the BH procedure with target FDR level α is defined by calculating

$$\hat{k}_\alpha(P) = \max\left\{k \in \{1, \ldots, n\} : \left|\left\{i : P_i \leqslant \frac{\alpha k}{n}\right\}\right| \geqslant k\right\},$$

with the convention that we set $\hat{k}_\alpha(P) = 0$ if this set is empty. Equivalently, if $P_{(i)}$ is the ith smallest p-value, then

$$\hat{k}_\alpha(P) = \max\left\{k \in \{1, \ldots, n\} : P_{(k)} \leqslant \frac{\alpha k}{n}\right\}.$$

The method then rejects the $\hat{k}_\alpha(P)$ smallest p-values or, equivalently, rejects all p-values that are less than or equal to $\alpha \hat{k}_\alpha(P)/n$. Benjamini and Hochberg (1995) then showed that this procedure provably controls the FDR at level α if the p-values are independent. Subsequent work by Benjamini and Yekutieli (2001) proved that this result holds under a relaxed condition, the positive regression dependence on a subset (PRDS) assumption, where the p-values are allowed to have positive dependence (see equation (5) in Section 3.1 for details).

2.2. Simes test for the global null

The BH procedure is closely related to earlier work by Simes (1986), which, for a vector of p-values $P = (P_1, \ldots, P_n)$, tests the *global null* hypothesis (which is also called the intersection hypothesis), i.e. tests whether *all* of these n p-values are null (there are no true signals). To perform this test, first calculate the Simes p-value

$$\mathrm{Simes}(P) = \min_{1 \leqslant k \leqslant n} \frac{P_{(k)} n}{k},$$

where, as before, $P_{(k)}$ is the kth smallest p-value in the list $P_1, \ldots, P_n$. The global null hypothesis is then rejected if $\mathrm{Simes}(P) \leqslant \alpha$, where α is the prespecified level of the test (the desired type I error rate).

To see the connection to the BH procedure, for any $n \geqslant 1$ and $\alpha \in [0, 1]$, write

$$P \in \mathrm{BH}(\alpha)$$

whenever $\hat{k}_\alpha(P) \geqslant 1$, i.e. this is equivalent to the statement that the set of p-values P leads to at least one rejection, when applying the BH procedure with target FDR level α. We then say that P *passes* the BH procedure at level α. Examining the definition of the BH procedure, we see that

$$\mathrm{Simes}(P) = \min\{\alpha \in [0, 1] : P \in \mathrm{BH}(\alpha)\},$$

i.e. the Simes p-value is the minimum threshold α for which P passes the BH procedure. In other words,

$$\mathrm{Simes}(P) \leqslant t \Leftrightarrow P \in \mathrm{BH}(t) \Leftrightarrow \hat{k}_t(P) \geqslant 1 \tag{1}$$

for any $t \in [0, 1]$. We should note that the Simes p-value really is a p-value in the true sense of the word—if the p-values are independent and uniform, then the Simes p-value is uniformly distributed under the global null (i.e. if $P_1, \ldots, P_n$ are independent and uniformly distributed). This is because

$$\Pr\{\text{Simes}(P) \leqslant t\} = \Pr\{P \in \text{BH}(t)\} = t$$

where the latter equality is a property of the BH procedure under the global null (Benjamini and Hochberg, 1995). Under positive dependence (i.e. PRDS), the Simes p-value becomes conservative, with $\Pr\{\text{Simes}(P) \leqslant t\} \leqslant t$ by properties of the BH procedure under positive dependence (Benjamini and Yekutieli, 2001).

2.3. *False discovery rate control only at the group level: interpolating between Simes and Benjamini–Hochberg*

Suppose for a moment that all the p-values are independent and uniformly distributed, and that we have partitioned our hypotheses into G groups of size $n_1, n_2, \ldots, n_G$, with $n = n_1 + \ldots + n_G$:

$$\underbrace{P_1, \ldots, P_{n_1}}_{\text{group } 1}, \underbrace{P_{n_1+1}, \ldots, P_{n_1+n_2}}_{\text{group } 2}, \ldots, \underbrace{P_{n_1+\ldots+n_{G-1}+1}, \ldots, P_n}_{\text{group } G},$$

and we wish to select a subset of these groups, $\hat{S}_{\text{grp}} \subseteq [G]$, so that the proportion of null groups is not too high. (In this setting, a 'null group' is a group consisting entirely of null hypotheses.)

We can consider the following simple procedure for this problem. Using the Simes p-value, we could reduce this to a standard multiple-testing problem: specifically, we compute the Simes p-values for each of the G groups,

$$\text{Simes}(P_{A_1}), \ldots, \text{Simes}(P_{A_G}),$$

where $A_g = \{n_1 + \ldots + n_{g-1} + 1, \ldots, n_1 + \ldots + n_g\}$ is the set of indices belonging to group g, and P_{A_g} is the vector of p-values belonging to this group. Then, apply the BH procedure with threshold α to this new list of p-values to produce a set $\hat{S}_{\text{grp}}$ of (group) discoveries. If the n p-values are independent, then, since we are simply applying the BH procedure to a set of p-values (which are independent and, for each null group, are uniformly distributed), we can then expect this procedure to control the group level FDR, and indeed it immediately follows that

$$\mathbb{E}\left[\frac{|\mathcal{H}^0_{\text{grp}} \cap \hat{S}_{\text{grp}}|}{1 \vee |\hat{S}_{\text{grp}}|}\right] \leqslant \alpha.$$

(Of course, we would have no corresponding guarantee for the overall FDR when the hypotheses are considered individually rather than in groups; our multilayer method, which is introduced shortly, gives this type of simultaneous guarantee.)

In fact, we can view this type of group FDR procedure as an interpolation between the Simes test of the global null and the Benjamini–Hochberg procedure, i.e. both the Simes test and the BH procedure are actually special cases of the group FDR control method that is described in this section, obtained by considering two extremes: one group of size n (corresponding to the Simes test of the global null), or n groups of size 1 (corresponding to the BH procedure). It is intuitively pleasing that our multilayer method, which is to be introduced later, also specializes to the Simes test and the BH procedure in the case of only one layer, exactly in the fashion that was mentioned above.

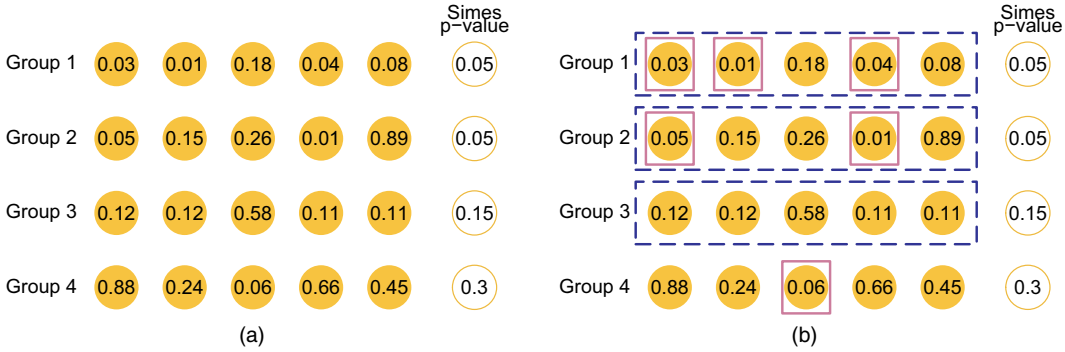

Fig. 1. (a) 20 p-values and their groupings (into rows) (●) and the Simes p-values for the groups (○) and (b) the discoveries made by running the BH procedure on the 20 p-values, with $\alpha = 0.2$ (□), and the discoveries made by running the group FDR controlling procedure independently (BH applied to the Simes p-values for each group), with $\alpha = 0.2$ (⊏⊐)

2.4. Independent group and individual level discoveries may conflict

Unfortunately, for two (or more) layers, controlling the FDR both at the group level (by using the Simes plus BH procedure in the previous subsection) and independently at the individual level (using the BH procedure) may cause conflicts in rejected groups and individual hypotheses. The example in Fig. 1 is meant to demonstrate exactly this issue, highlighting the complications that may arise in the multilayer setting, even for just two layers. Here, we divide 20 p-values into four groups of five p-values each, and we choose to control the FDR at the individual and group levels, both at $\alpha = 0.2$.

The row and individual level rejections in Fig. 1 are in conflict, because the third row is discovered at the group level but does not contain any hypotheses that are discovered at the individual level; conversely, the fourth row was not discovered at the group level but has a p-value discovered at the individual level. Thus although these outcomes guarantee FDR control at the group and individual levels, the output of this procedure is not *internally consistent*. If we throw away all rejections that are in conflict, by rejecting the individual hypotheses only from the first two rows (i.e. taking the intersection of rejections at different layers) and discarding the group rejection of the third row and the individual rejection in the fourth row, we now have a result that is internally consistent, but unfortunately we have lost the guarantees of FDR control—indeed, it is a well-known property of the BH procedure that rejecting fewer hypotheses than recommended by the BH procedure may sometimes increase the FDR.

There are special cases where the group and individual level rejections may not be in conflict, as discussed in Peterson *et al.* (2016). However, this certainly does not generalize to arbitrarily many layers of arbitrary groups being tested at arbitrary levels. This motivates the further study of procedures that can provide *simultaneous* FDR guarantees for multiple possibly non-hierarchical layers. Indeed, our general and efficient p-filter algorithm provably circumvents the obstacles mentioned above in quite some generality.

3. Controlling false discovery rate for individual hypotheses and for groups

Assume again that we partition our set of n hypotheses (and their corresponding p-values) into G groups of size $n_1, n_2, \ldots, n_G$:

$$\underbrace{P_1, \ldots, P_{n_1}}_{\text{group } 1}, \underbrace{P_{n_1+1}, \ldots, P_{n_1+n_2}}_{\text{group } 2}, \ldots, \underbrace{P_{n_1+\ldots+n_{G-1}+1}, \ldots, P_n}_{\text{group } G},$$

with $n = n_1 + \ldots + n_G$. Let $\mathcal{H}^0 \subseteq [n]$ index the unknown set of null hypotheses, and define the set of null groups (groups that contain only null hypotheses) as

$$\mathcal{H}^0_{\text{grp}} = \{g : A_g \subseteq \mathcal{H}^0\}$$

where A_g is the set of indices belonging to group g as before. We now consider the problem of controlling the FDR at the individual and the group level simultaneously, possibly for different target FDR levels α_{ov} and α_{grp}.

3.1. Overall false discovery rate and group level false discovery rate

We now present our proposed method, the *p*-filter, for controlling the FDR at both granularities, i.e. the standard overall FDR and the group level FDR.

First, consider a pair of thresholds $(t_{\text{ov}}, t_{\text{grp}}) \in [0,1] \times [0,1]$ (we show below how the *p*-filter chooses these thresholds adaptively; for the purposes of definitions assume that they are given). At this pair of thresholds, we define the set of all discoveries (rejections) that are made by the algorithm as

$$\hat{S} = \hat{S}(t_{\text{ov}}, t_{\text{grp}}) = \left\{i : P_i \leqslant t_{\text{ov}} \text{ and } \text{Simes}(P_{A_{g(i)}}) \leqslant t_{\text{grp}}\right\}, \tag{2}$$

where $g(i)$ is the group to which P_i belongs. In other words, a hypothesis is rejected if and only if its *p*-value P_i is below the overall threshold t_{ov} *and* the Simes *p*-value for its group, $P_{A_{g(i)}}$, is below the group threshold t_{grp}. Next, define the set of group discoveries as $\hat{S}_{\text{grp}} = \hat{S}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}}) = \{g : \hat{S}(t_{\text{ov}}, t_{\text{grp}}) \cap A_g \neq \emptyset\}$, i.e. any group with at least one discovery is considered to be a selected group.

Ideally, for any choice $(t_{\text{ov}}, t_{\text{grp}})$, we would like to be able to measure the overall FDP at these thresholds,

$$\text{FDP}_{\text{ov}} = \text{FDP}_{\text{ov}}(t_{\text{ov}}, t_{\text{grp}}) = \frac{|\mathcal{H}^0 \cap \hat{S}(t_{\text{ov}}, t_{\text{grp}})|}{1 \vee |\hat{S}(t_{\text{ov}}, t_{\text{grp}})|}, \tag{3}$$

and the group FDP,

$$\text{FDP}_{\text{grp}} = \text{FDP}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}}) = \frac{|\mathcal{H}^0_{\text{grp}} \cap \hat{S}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}})|}{1 \vee |\hat{S}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}})|}. \tag{4}$$

To estimate these quantities, we define the estimated overall FDP as

$$\widehat{\text{FDP}}_{\text{ov}} = \widehat{\text{FDP}}_{\text{ov}}(t_{\text{ov}}, t_{\text{grp}}) = \frac{n t_{\text{ov}}}{1 \vee |\hat{S}(t_{\text{ov}}, t_{\text{grp}})|},$$

and the estimated group FDP as

$$\widehat{\text{FDP}}_{\text{grp}} = \widehat{\text{FDP}}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}}) = \frac{G t_{\text{grp}}}{1 \vee |\hat{S}_{\text{grp}}(t_{\text{ov}}, t_{\text{grp}})|}.$$

We use the 'hats' in our estimated FDP notation, to remind the reader that these quantities are empirical; we can explicitly calculate them from the data P since they do not depend on knowing the underlying true set of nulls $\mathcal{H}^0$.

To understand these definitions, note that, if there are $|\mathcal{H}^0|$ many null *p*-values which are uniformly distributed, then we expect roughly $|\mathcal{H}^0| t_{\text{ov}} \leqslant n t_{\text{ov}}$ many of them to lie below the threshold t_{ov}, and similarly for the $|\mathcal{H}^0_{\text{grp}}| \leqslant G$ many null groups. Therefore the numerators in $\widehat{\text{FDP}}_{\text{ov}}(t_{\text{ov}}, t_{\text{grp}})$ and $\widehat{\text{FDP}}(t_{\text{ov}}, t_{\text{grp}})$ give intuitive (over)estimates of the numerators in the true

FDPs, $\mathrm{FDP}_{\mathrm{ov}}(t_{\mathrm{ov}}, t_{\mathrm{grp}})$ and $\mathrm{FDP}_{\mathrm{grp}}(t_{\mathrm{ov}}, t_{\mathrm{grp}})$ respectively. (This is the motivation underlying the BH procedure, extended also to the group setting.)

For any target FDR bounds $(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})$, define the set of admissible thresholds

$$\hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}}) = \{(t_{\mathrm{ov}}, t_{\mathrm{grp}}) \in [0,1] \times [0,1] : \widehat{\mathrm{FDP}}_{\mathrm{ov}} \leqslant \alpha_{\mathrm{ov}} \text{ and } \widehat{\mathrm{FDP}}_{\mathrm{grp}} \leqslant \alpha_{\mathrm{grp}}\}.$$

Our first result shows that the set $\hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}}) \subseteq [0,1] \times [0,1]$ has a well-defined maximum.

Theorem 1. Fix any $\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}} \in [0,1]$ and any vector of p-values $P \in [0,1]^n$. Define

$$\hat{t}_{\mathrm{ov}} = \max\{t_{\mathrm{ov}} \in [0,1] : \exists\, t_{\mathrm{grp}} \in [0,1] \text{ such that } (t_{\mathrm{ov}}, t_{\mathrm{grp}}) \in \hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})\}$$

and

$$\hat{t}_{\mathrm{grp}} = \max\{t_{\mathrm{grp}} \in [0,1] : \exists\, t_{\mathrm{ov}} \in [0,1] \text{ such that } (t_{\mathrm{ov}}, t_{\mathrm{grp}}) \in \hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})\}.$$

Then $(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}}) \in \hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})$.

Intuitively, this result implies that $\hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})$ is a region in $[0,1] \times [0,1]$ that has a maximum 'corner': a point $(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$ such that $(t_{\mathrm{ov}}, t_{\mathrm{grp}}) \leqslant (\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$ for all points $(t_{\mathrm{ov}}, t_{\mathrm{grp}}) \in \hat{\mathcal{T}}(\alpha_{\mathrm{ov}}, \alpha_{\mathrm{grp}})$. We remark that $\hat{t}_{\mathrm{ov}}$ and $\hat{t}_{\mathrm{grp}}$ always take values in a discrete grid:

$$\hat{t}_{\mathrm{ov}} \in \left\{\alpha_{\mathrm{ov}} \frac{k}{n} : k = 0, \ldots, n\right\}$$

and

$$\hat{t}_{\mathrm{grp}} \in \left\{\alpha_{\mathrm{grp}} \frac{k}{G} : k = 0, \ldots, G\right\}.$$

The construction given in theorem 1 defines our procedure: the p-filter procedure, applied to the given p-values P and given partition into groups, returns the set of rejections or discoveries given by $\hat{S}(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$. Recall that the set of discoveries $\hat{S}(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$ consists of all hypotheses whose individual p-value P_i and group p-value $\mathrm{Simes}(P_{A_{g(i)}})$ *both* lie below their respective adaptive thresholds; the name 'p-filter' refers to this process, where the rejected p-values are those that pass through both an individual level filter and a group level filter.

With our method now defined, we turn to a theorem on FDR control at both the individual and the group level. First, we introduce the PRDS assumption, which was originally formulated by Benjamini and Yekutieli (2001):

$$\begin{gathered}\text{for any non-decreasing set } D \subseteq [0,1]^n \text{ and any } i \in \mathcal{H}^0, \\ t \mapsto \Pr\{P \in D | P_i \leqslant t\} \text{ is a non-decreasing function over } t \in (0,1].\end{gathered} \tag{5}$$

(In Benjamini and Yekutieli's (2001) work, the assumption involves the probability $\Pr(P \in D | P_i = t)$, rather than conditioning on the event $P_i \leqslant t$ as in assumption (5) which we prefer for later convenience; however, as discussed in their work, the two assumptions are equivalent. Also, recall that a set $D \in \mathbb{R}^n$ is called non-decreasing if $x \in D$ implies that $y \in D$ for any $y \geqslant x$ in the orthant ordering (i.e. $y \geqslant x$ if $y_i \geqslant x_i$ for all i).) We also assume that each true null p-value is uniformly distributed—in fact, our assumption is more flexible:

$$\text{for any } i \in \mathcal{H}^0, \Pr(P_i \leqslant t) \leqslant t \quad \text{for all } t \in [0,1]. \tag{6}$$

This assumption holds trivially if $P_i \sim \mathrm{uniform}[0,1]$ but also allows for a misspecified null distribution in some settings. We are now ready to state our result.

Theorem 2. Let the p-values P satisfy assumptions (5) and (6) above, let $(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$ be defined as in theorem 1 and let $\hat{S}(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})$ be the set of discoveries returned by the p-filter, as defined in equation (2). Then the p-filter controls both the overall and the group FDR, i.e.

$$\mathbb{E}[\mathrm{FDP}_{\mathrm{ov}}(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})] \leqslant \alpha_{\mathrm{ov}} \frac{|\mathcal{H}^0|}{n}$$

and

$$\mathbb{E}[\mathrm{FDP}_{\mathrm{grp}}(\hat{t}_{\mathrm{ov}}, \hat{t}_{\mathrm{grp}})] \leqslant \alpha_{\mathrm{grp}} \frac{|\mathcal{H}^0_{\mathrm{grp}}|}{G}.$$

In fact, the set-up that is described here is a special case of a multilayer FDR framework that we describe below, where we seek to control the FDR simultaneously across multiple partitions or partitions of the hypotheses. First, however, we describe an existing approach to the grouped FDR problem to compare it with our method for this setting.

3.2. *Existing work: within-group false discovery rate and group level false discovery rate*

In recent work, Benjamini and Bogomolov (2014) proposed a related method for the multiple-hypothesis testing problem with grouped structure. In their method, the first step is a screening step to select a set of groups of interest, $\hat{S}_{\mathrm{grp}}$; the mechanism for this screening step is determined by the user subject to some mild conditions. The second step is then to test the p-values within each selected group: for each $g \in \hat{S}_{\mathrm{grp}}$, run a selection procedure that controls the FDR at the level $\alpha_{\mathrm{ov}}|\hat{S}_{\mathrm{grp}}|/G$. Peterson *et al.* (2016) developed this method further by examining a specific choice for the screening step.

(a) First, apply the BH procedure with threshold α_{grp} to the Simes p-values of the G groups,

$$\mathrm{Simes}(P_{A_1}), \ldots, \mathrm{Simes}(P_{A_G}),$$

to select a set of groups $\hat{S}_{\mathrm{grp}}$. The group level FDP is now given by $|\mathcal{H}^0_{\mathrm{grp}} \cap \hat{S}_{\mathrm{grp}}|/(1 \vee |\hat{S}_{\mathrm{grp}}|)$.

(b) Next, for each selected group $g \in \hat{S}_{\mathrm{grp}}$, run the BH procedure with threshold $\alpha_{\mathrm{ov}}|\hat{S}_{\mathrm{grp}}|/G$ on the p-values within the group, P_{A_g}. Let $\hat{S}_g$ be the selected set within group g. The FDP within group g is now given by $|\mathcal{H}^0 \cap \hat{S}_g|/(1 \vee |\hat{S}_g|)$.

Peterson *et al.* (2016) showed that the first step ensures that the group level FDR is controlled at level α_{grp}:

$$\mathbb{E}\left[\frac{|\mathcal{H}^0_{\mathrm{grp}} \cap \hat{S}_{\mathrm{grp}}|}{1 \vee |\hat{S}_{\mathrm{grp}}|}\right] \leqslant \alpha_{\mathrm{grp}}.$$

(This follows from the properties of Simes's test and the BH procedure.) Furthermore, Benjamini and Bogomolov's (2014) results guarantee that the resulting average FDP across all selected groups is controlled as

$$\mathbb{E}\left[\frac{\sum_{g \in \hat{S}_{\mathrm{grp}}} \text{FDP in group } g}{1 \vee |\hat{S}_{\mathrm{grp}}|}\right] = \mathbb{E}\left[\frac{\sum_{g \in \hat{S}_{\mathrm{grp}}} |\mathcal{H}^0 \cap \hat{S}_g|/(1 \vee |\hat{S}_g|)}{1 \vee |\hat{S}_{\mathrm{grp}}|}\right] \leqslant \alpha_{\mathrm{ov}}, \tag{7}$$

under the assumption that p-values in one group are independent of the other groups (with positive dependence allowed within each group).

Our p-filter method clearly has much in common with this procedure, but the two offer different types of guarantee. The p-filter does not offer control of the averaged within-group FDR; our guarantee is different, giving overall FDR control across all hypotheses selected. Depending on the setting, one or the other measure of false discovery control may be more desirable. We also note that the p-filter extends to a more general setting, which is discussed next, and is unique in allowing us to move to multiple partitions which are not necessarily arranged hierarchically, and allows dependence between p-values across groups.

4. Multilayer false discovery rate control

We now turn to the more general problem of *multilayer* FDR control, where we seek to control the FDR across a range of arbitrary partitions of the hypotheses.

Suppose that we are given n p-values, $P_1, \dots, P_n \in [0,1]$, with an unknown set of nulls $\mathcal{H}^0 \subseteq [n]$. Furthermore, suppose that we have M partitions (layers) of interest, with the mth partition having G_m groups:

$$A_1^m, \dots, A_{G_m}^m \subseteq [n]$$

for $m = 1, \dots, M$. To return to the example that was mentioned in Section 1, in a functional MRI study with V voxels and S time points, we might consider three layers:

(a) layer $m = 1$ considers every voxel and time point separately (VS groups);
(b) layer $m = 2$ considers each voxel across all time points (V groups);
(c) layer $m = 3$ considers each time point across all voxels within each of R regions of interest (ROIs) (SR groups).

Define the null set for the mth partition as

$$\mathcal{H}_m^0 = \{g \in [G_m] : A_g^m \subseteq \mathcal{H}^0\},$$

and, given a set $\hat{S} \subseteq [n]$ of rejections, we define the mth rejection set as

$$\hat{S}_m = \{g \in [G_m] : \hat{S} \cap A_g^m \neq \emptyset\}.$$

In our running functional MRI example, for instance, $\mathcal{H}_2^0$ is the set of voxels v such that (v, s) is a null across *all* time points s, whereas $\hat{S}_2$ is the set of voxels v for which (v, s) is a discovery for *any* time point s.

Given a selected set $\hat{S}$, define the FDP for the mth partition as

$$\mathrm{FDP}_m(\hat{S}) = \frac{|\hat{S}_m \cap \mathcal{H}_m^0|}{1 \vee |\hat{S}_m|}.$$

Now we describe the p-filter procedure for this more general setting. Consider any thresholds $(t_1, \dots, t_M) \in [0,1]^M$. We let

$$\begin{aligned}\hat{S}(t_1, \dots, t_M) &= \bigcap_{m=1}^{M} \left(\bigcup_{g=1,\dots,G_m : \mathrm{Simes}(P_{A_g^m}) \leqslant t_m} A_g^m \right) \\ &= \{i : \text{for all } m, \mathrm{Simes}(P_{A_{g(m,i)}^m}) \leqslant t_m\}, \end{aligned} \tag{8}$$

where $g(m, i)$ indexes the group that P_i belongs to in the mth partition, i.e. for each partition m we take the union of all groups whose Simes p-value is less than or equal to t_m; by taking the intersection across all layers, we see that a p-value P_i is selected if, at every layer m, its group

$A^m_{g(m,i)}$ passes this test. Correspondingly, we have

$$\hat{S}_m(t_1,\dots,t_M) = \{g \in [G_m] : \hat{S}(t_1,\dots,t_M) \cap A^m_g \neq \emptyset\}. \tag{9}$$

We then let

$$\mathrm{FDP}_m(t_1,\dots,t_M) = \frac{|\hat{S}_m(t_1,\dots,t_M) \cap \mathcal{H}^0_m|}{1 \vee |\hat{S}_m(t_1,\dots,t_M)|},$$

and define the estimated FDP as

$$\widehat{\mathrm{FDP}}_m(t_1,\dots,t_M) = \frac{G_m t_m}{1 \vee |\hat{S}_m(t_1,\dots,t_M)|}.$$

Now define

$$\hat{\mathcal{T}}(\alpha_1,\dots,\alpha_M) = \{(t_1,\dots,t_M) \in [0,1]^M : \widehat{\mathrm{FDP}}_m(t_1,\dots,t_M) \leqslant \alpha_m \text{ for all } m\}.$$

The next result proves that theorem 1 extends to this more general setting, meaning that the set $\hat{\mathcal{T}}(\alpha_1,\dots,\alpha_m)$ does indeed have a well-defined maximum point, thus defining our method.

Theorem 3. Fix any $\alpha_1,\dots,\alpha_M \in [0,1]$ and any vector of p-values $P \in [0,1]^n$. Define

$$\hat{t}_m = \max\{t_m : \exists t_1,\dots,t_{m-1},t_{m+1},\dots,t_M \text{ such that } (t_1,\dots,t_M) \in \hat{\mathcal{T}}(\alpha_1,\dots,\alpha_M)\}$$

for each $m = 1,\dots,M$. Then

$$(\hat{t}_1,\dots,\hat{t}_M) \in \hat{\mathcal{T}}(\alpha_1,\dots,\alpha_M).$$

The p-filter then selects the set

$$\hat{S}(\hat{t}_1,\dots,\hat{t}_M).$$

As before, we remark that these adaptive thresholds take values on a discrete grid, with

$$\hat{t}_m \in \left\{\alpha_m \frac{k}{G_m} : k = 0,\dots,G_m\right\} \tag{10}$$

for each m, but it is possible to find $(\hat{t}_1,\dots,\hat{t}_M)$ efficiently and without exhaustive search over this grid; see our algorithm given in Section 5.

Next, our main theorem shows that the FDR is controlled simultaneously for each partition, by our p-filter.

Theorem 4. Let the p-values $P \in [0,1]^n$ satisfy assumptions (5) and (6) above, and let $(\hat{t}_1,\dots,\hat{t}_M)$ be defined as in theorem 3. Then, for each $m = 1,\dots,M$, the method controls the FDR for the mth partition,

$$\mathbb{E}[\mathrm{FDP}_m(\hat{t}_1,\dots,\hat{t}_M)] \leqslant \alpha_m \frac{|\mathcal{H}^0_m|}{G_m}.$$

Clearly, this is a generalization of the setting that was considered previously, where the overall FDR and the group FDR can be controlled by defining two partitions: one that splits $[n]$ into n many singleton sets, and one that is defined by the group structure. Note that the theoretical results for the initial setting, theorems 1 and 2, are simply special cases of the more general results, theorems 3 and 4 respectively. Unlike the overall FDR or group FDR setting, however, in general the M partitions do not need to be nested; they are not constrained to form a hierarchy of partitions.

The proofs of theorems 1–4 are deferred to Appendix A, but one of the main ingredients is a technical lemma that could be of broader interest, and hence we state it below. First, for convenience, we define some notation: since the ratio '0/0' often arises in FDR control results, we let

$$\frac{a}{b} = \begin{cases} a/b, & \text{if } b \neq 0, \\ 0, & \text{if } a = b = 0, \\ \text{undefined}, & \text{otherwise.} \end{cases} \tag{11}$$

We shall use this to define conditional probability when the conditioned event has probability 0, i.e. for two events A and B,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \begin{cases} (\text{the usual definition}), & \Pr(B) > 0, \\ 0, & \Pr(B) = 0. \end{cases}$$

Let X be an arbitrary random variable and write $F(y) = \Pr(X \leqslant y)$ for the cumulative density function of X. Hence, it trivially follows that

$$\mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant y\}}{F(y)}\right] \leqslant 1 \qquad \text{for any fixed constant } y.$$

Our main lemma below states that the above result also holds for certain random Y.

Lemma 1. Let $X, Y \in \mathbb{R}$ be random variables satisfying the assumption that

$$\text{for any } y, \text{ the function } x \mapsto \Pr(Y < y | X < x) \text{ is non-decreasing in } x. \tag{12}$$

Then, we have

$$\mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant Y\}}{F(Y)}\right] \leqslant 1.$$

The proof of lemma 1 is given in Appendix A. This lemma gives an immediate corollary allowing us to understand the interaction between a null p-value P_i and any function of the vector of p-values. First note that, for any null p-value P_i, our superuniformity assumption (6) can be restated as

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \leqslant t\}}{t}\right] \leqslant 1 \qquad \text{for any fixed threshold } t.$$

The following corollary states that the above result continues to remain true for certain random thresholds.

Corollary 1. Let P_i be null, satisfying superuniformity assumption (6), and assume that $P \in [0,1]^n$ is PRDS with respect to P_i as in expression (5). Then, for any function $f : [0,1]^n \to [0, \infty)$ that is non-increasing (with respect to the orthant ordering), we have

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \leqslant f(P)\}}{f(P)}\right] \leqslant 1.$$

Proof. We apply lemma 1 by setting $X = P_i$ and $Y = f(P)$. Fix any $y \in \mathbb{R}$, and define $D = \{p \in \mathbb{R}^n : f(p) < y\}$. Since f is a non-increasing function, this means that D is a non-decreasing set. Therefore,

$$\Pr(Y < y | X < x) = \Pr(P \in D | P_i < x)$$

is a non-decreasing function of x, by the PRDS assumption (5). (Although the PRDS assumption is stated using $\Pr(P \in D | P_i \leqslant x)$, we can replace this with conditioning on *strict* inequality by taking limits.) Writing F to be the cumulative distribution function of $X = P_i$ and applying lemma 1,

$$1 \geqslant \mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant Y\}}{F(Y)}\right] = \mathbb{E}\left[\frac{\mathbf{1}\{P_i \leqslant f(P)\}}{F\{f(P)\}}\right] \geqslant \mathbb{E}\left[\frac{\mathbf{1}\{P_i \leqslant f(P)\}}{f(P)}\right],$$

where the last step holds because $F\{f(P)\} \leqslant f(P)$ always by assumption (6).

We note that this result is an extension of the work of Benjamini and Yekutieli (2001) for analysing FDR control of the BH procedure under the PRDS assumption; as part of their work, they proved an analogous result for the specific function $f(P) = \alpha \hat{k}_\alpha(P)/n$ under the assumption $P_i \sim \text{uniform}[0, 1]$.

4.1. Comments on power and precision of the p-filter

The following points are worthy of note. As mentioned earlier, running the p-filter with one partition, which is the trivial finest partition, is exactly equivalent to running the classical BH procedure. Similarly, running the p-filter with two partitions, the finest with threshold α_1, and any other partition with threshold α_2, is exactly equivalent to running the BH procedure if we set $\alpha_2 = \infty$. This observation can be further generalized to the case of M partitions: running the p-filter with the first partition being the trivial partition and with $\alpha_2 = \alpha_3 = \ldots = \alpha_M = \infty$ is exactly equivalent to running the BH procedure with $\alpha = \alpha_1$.

Since the set of discoveries is non-decreasing as a function of the thresholds $\alpha_1, \ldots, \alpha_M$, running the p-filter with non-trivial (i.e. finite) $\alpha_1, \ldots, \alpha_M$ leads to a set of discoveries that is no larger than the set that is produced by the BH procedure with threshold $\alpha = \alpha_1$; often the set is strictly smaller, and so the p-filter's power is strictly lower. At the same time, we may often have lower *achieved* FDR as well, even at the individual level (the overall FDR), since the added layers of the p-filter can increase the precision of our discoveries.

As a simple example, consider a two-layer partition with n groups of size 1 at level α_1, and with one group of size n at level α_2. We compare with the BH procedure with $\alpha = \alpha_1$ (which is equivalent to setting $\alpha_2 = \infty$). Then under the global null, if all p-values are independent and uniform, the probability of at least one rejection is equal to α_1 for the BH procedure and is equal to $\min\{\alpha_1, \alpha_2\}$ for the p-filter; under the global null this probability is equal to the FDR, so we see a lower *achieved* FDR for the p-filter.

5. Algorithm

In this section we present an efficient algorithm for implementing our method, given in algorithm 1 (Table 1). This algorithm yields the correct solution according to the following result.

Table 1. Algorithm 1: the p-filter for multilayer FDR control

Input: a vector of p-values $P \in [0, 1]^n$; target FDR levels $\alpha_1, \ldots, \alpha_M$; partition m given by $A_1^m, \ldots, A_{G_m}^m \subseteq [n]$ for $m = 1, \ldots, M$
Initialize: thresholds $t_1 = \alpha_1, \ldots, t_M = \alpha_M$;
repeat
  for $m = 1, \ldots, M$ *do*
    update the mth threshold; defining $\hat{S}_m(\cdot)$ as in equation (9), let

$$t_m \leftarrow \max\left\{T \in [0, t_m] : \frac{G_m T}{1 \vee |\hat{S}_m(t_1, \ldots, t_{m-1}, T, t_{m+1}, \ldots, t_M)|} \leqslant \alpha_m\right\}$$

  end for
until the thresholds $t_1, \ldots, t_M$ are all unchanged in the last round
Output: adaptive thresholds $\hat{t}_1 = t_1, \ldots, \hat{t}_M = t_M$

Theorem 5. The output of algorithm 1 is the vector of thresholds $(\hat{t}_1, \ldots, \hat{t}_m)$ defined in theorem 3.

Next we assess the run time of this algorithm. First, by definition of the algorithm, the t_ms cannot increase; therefore the sets $\hat{S}_m(t_1, \ldots, t_M)$ cannot increase over the iterations of the algorithm, and so the denominator in the update step in algorithm 1 is non-increasing. Therefore, for each run of the outer loop (the 'repeat ... until' loop), either t_m decreases strictly for some m, or all t_ms stay the same and so the algorithm terminates. Furthermore, observe that the maximizer t_m to the update step must lie in the set $\{\alpha_m k/G_m : k = 0, \ldots, G_m\}$. This means that there can be at most $G_1 + \ldots + G_M$ distinct instances where one of the t_ms decreases, and so the algorithm terminates after at most $G_1 + \ldots + G_M + 1$ passes through the outer loop.

6. Experiments with simulated data

In this section we examine two designs: one setting where we seek to control individual level and group level FDR control as discussed in Section 3, and a second more complex setting where we consider three different partitions of the hypotheses simultaneously. We compare the p-filter with the BH method and with Benjamini and Bogomolov's (2014) method. For both experiments, all p-values in the simulations are independent and are generated as follows:

$$X \sim \mu + \mathcal{N}(0, 1), \qquad p\text{-value} = 1 - \Phi(X) \tag{13}$$

where Φ is the standard Gaussian cumulative distribution function, with $\mu = 0$ for nulls and $\mu > 0$ for true signals. Larger values of μ correspond to stronger true signals that are easier to detect. All simulations were run in R (R Core Team, 2015). (R code implementing our method through algorithm 1, along with scripts reproducing all the simulated experiments that are presented in this paper, can be found at `http://www.stat.uchicago.edu/~rina/pfilter.html`.)

6.1. Grouped setting

In our first simulation, we consider a simple grouped scenario: we have $n = 1000$ hypotheses, partitioned into 100 groups of size 10. There are 55 true signals: one in group 1, two in group 2, ..., and 10 in group 10. The target FDR levels were set at $\alpha = 0.2$ for all methods (for both the overall and the group level FDR).

Fig. 2 shows the outcome of one trial run of the simulation (with $\mu = 3$); for convenience, we display our tests in a 10×100 array where each column corresponds to a group and the first 10 columns contain all the true signals. We see that the p-filter and the BB method both select very few null columns (groups), which is desirable; in fact, the results from these two methods are nearly identical. The BH method, which does not use the partition of the hypotheses, selects many null groups (columns) but is also slightly better able to find the true signals.

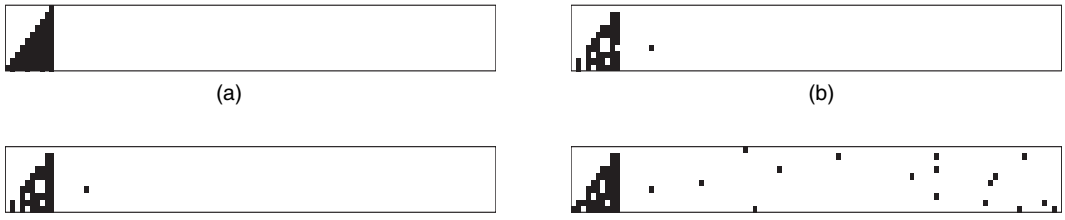

Fig. 2. Demonstration of one trial run of the groupwise sparsity simulation (Section 6.1): (a) true signals; (b) p-filter; (c) BB; (d) BH

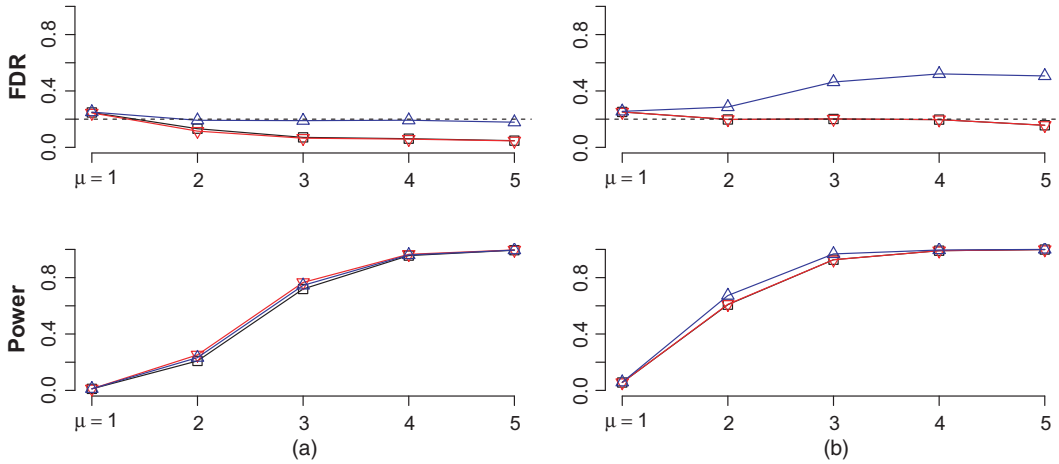

Fig. 3. Results for the groupwise sparsity simulation (Section 6.1), averaged over 100 trials (· · · · · ·, target FDR level for each of the partitions; □, *p*-filter; ▽, BB; △, BH): (a) by entry; (b) by column

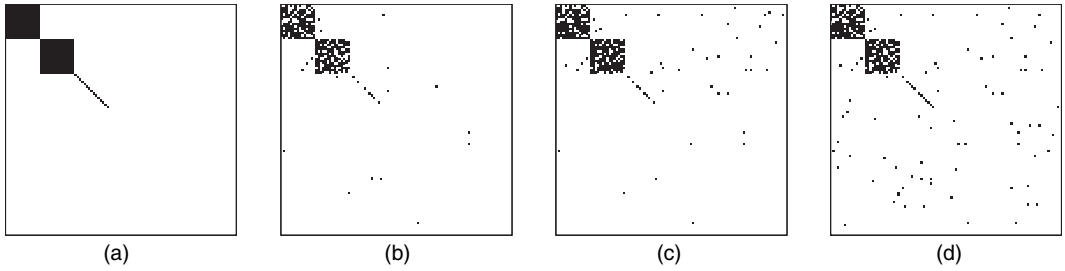

Fig. 4. Demonstration of one trial of the rowwise and columnwise sparsity simulation (Section 6.2): (a) true signals; (b) *p*-filter; (c) BB; (d) BH

Results across a range of μ-values are shown in Fig. 3, plotting the FDR and power for this array of hypotheses at the individual (entrywise) and group (columnwise) levels. We see that the three methods have very similar power (with slightly higher power for the BH method), but different FDR control properties: although all three methods control the entrywise FDR, as expected we see that the BH method does not control the columnwise FDR. The *p*-filter and BB methods show nearly identical results, with very slightly higher entrywise power for the BB method.

6.2. Multilayer setting

We now consider a setting where the structure of the true signals is best captured by using multiple partitions of the data. In this setting, the $n = 10\,000$ hypotheses are arranged into a 100×100 grid. The true signals lie in two 15×15 blocks, plus 15 additional signals that lie along a diagonal, and are therefore alone in their respective rows and columns (Fig. 4(a)). Therefore, they are sparse at the individual (entrywise) level but also are rowwise sparse and columnwise sparse. The 15 signals along the diagonal make this simulation more challenging for the *p*-filter and BB methods, which are best able to find signals that are grouped together. We again compare three methods: the *p*-filter (with three layers: entries, rows and columns), the BB procedure (where the groups are defined by the rows) and the BH procedure. The target FDR levels were set at $\alpha = 0.2$ for all methods (for the overall FDR and the rowwise and columnwise FDR).

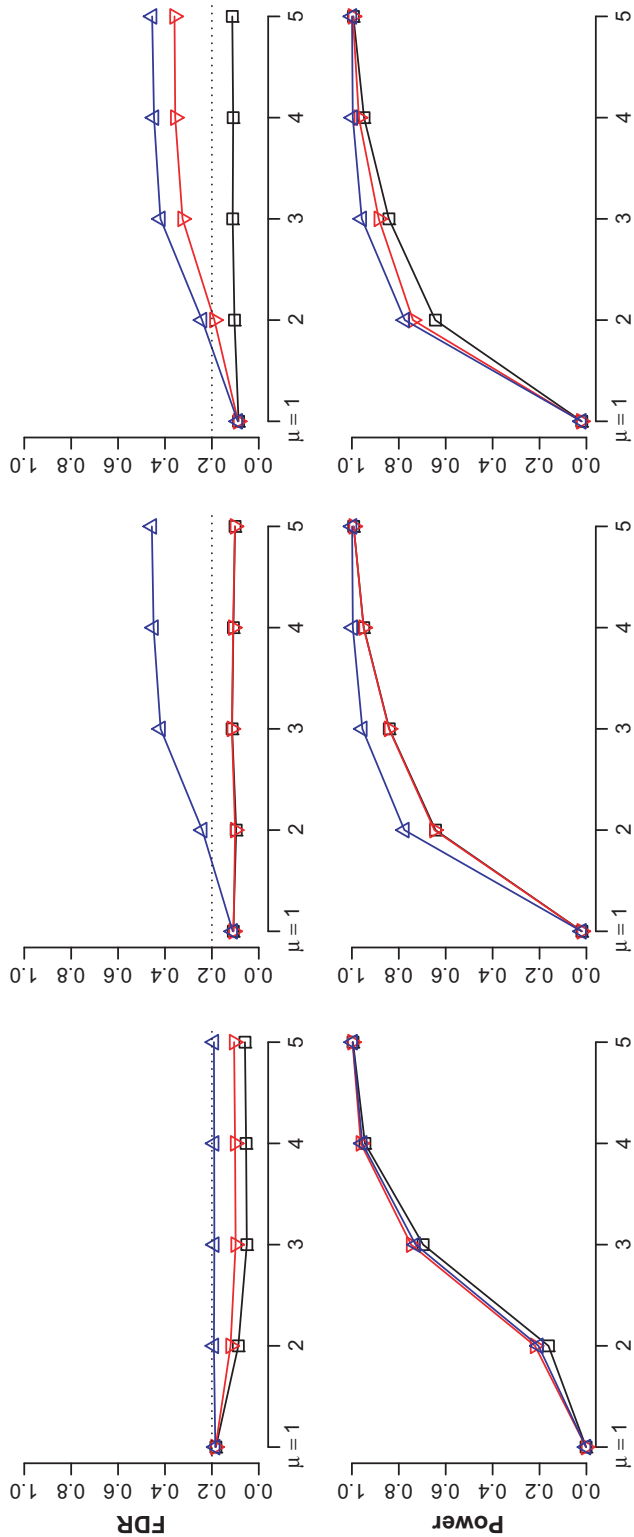

Fig. 5. Results for the rowwise and columnwise sparsity simulation (Section 6.2), averaged over 100 trials (· · · · · ·, target FDR level for each of the partitions; □, *p*-filter; ▽, BB; △, BH): (a) by entry; (b) by row; (c) by column

Fig. 4 shows the outcome of one trial run of the simulation (with $\mu = 3$). We see that the p-filter selects few null rows and columns, which is desirable. The BB procedure, with the groups defined as rows, selects few null rows but many null columns. The BH procedure, which does not use row or column information, selects many null rows and columns. In contrast, the BH procedure is much better able to find the sparse signals along the diagonal, as expected.

Results across a range of μ-values are shown in Fig. 5, plotting the FDR and power at the entrywise, rowwise and columnwise levels for this two-dimensional array of hypotheses. At the entrywise level, the three methods have similar power and all control the FDR. For rows, the BH procedure does not control the rowwise FDR as expected but can achieve higher power (due to the sparse signals along the diagonal). For columns, the BH and BB procedures both lose FDR control as expected, with a corresponding slight increase in power. The p-filter controls all three forms of FDR, as guaranteed by our theoretical results, and achieves good power across the three layers.

7. Experiment with functional magnetic resonance imaging data

We now demonstrate one way to use spatial and temporal prior information to aid inference in neuroscientific applications. (R code reproducing this experiment can be found at `http://www.stat.uchicago.edu/~rina/pfilter.html`.) We use freely available functional MRI data from Wehbe *et al.* (2014). Eight subjects read a chapter of *Harry Potter and the Sorcerer's Stone* while words were presented one at a time. The total presentation time is 2710 s and the available data consist of 1355 volumes of functional MRI activity (one scan every 2 s) for each of the eight subjects, each scanned with the same timeline of stimulus presentation. Each subject's brain is represented in 3 mm $\times$ 3 mm $\times$ 3 mm voxels, which are all normalized to the same co-ordinate space, with 41 073 voxels common to all eight subjects. The text was annotated with multiple types of intermediate features: in the analysis that follows, we use the semantic annotations that are available in the paper's accompanying on-line supporting information.

7.1. Temporal prior information

The functional MRI machine measures the haemodynamic response, which is a delayed response that is the neural correlate of brain activity corresponding to changes in the magnetic field due to blood flowing into the brain as a result of brain activity. Every functional MRI sample could therefore be approximated as the superposition of events happening in the 8–10 preceding seconds. It is hence appropriate to ask whether the features that are presented at time t can predict the brain activity at time $t + s$, for $s = 4, 6, 8$ s, which based on prior knowledge corresponds to the peak of the haemodynamic response.

7.2. Computing p-values for feature–activity correlations

For each delay s, we use the same predictive encoding model as proposed in Wehbe *et al.* (2014), where one fits a linear regression model from the text's semantic features to each voxel's recorded functional MRI brain activity delayed by s s. Data from all eight subjects are used for this to boost the signal-to-noise ratio. It was determined in Wehbe *et al.* (2015) that the results obtained and conclusions on this data set are quite stable to various modelling and algorithmic choices like regularization and smoothing. Hence the exact methods that are used are not very relevant for our present purposes and the reader is directed to Wehbe *et al.* (2014, 2015) for more details. This finally yields a p-value $P_{v,s}$ for each voxel v and delay s. Each p-value $P_{v,s}$ represents the question 'Is voxel v correlated with the semantic features of the text presented s s earlier?'. Fig. 6 displays these p-values on a brain (for $s = 6$ s, on a negative logarithm scale), using the

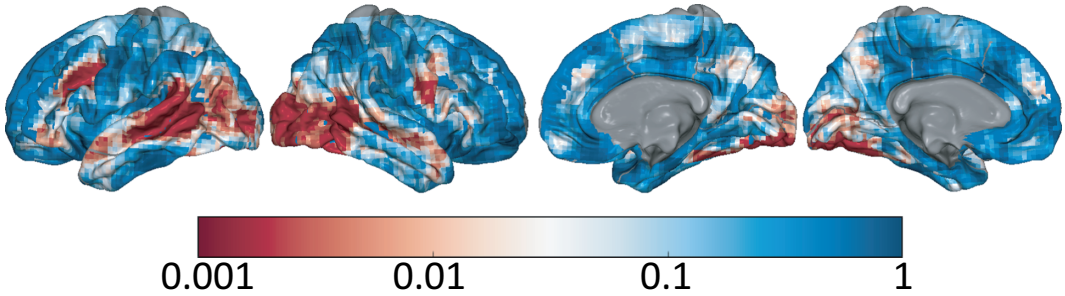

Fig. 6. For time delay $s = 6$, original p-values (between 1 and 10^{-3}) are plotted on a negative log-scale, one for each voxel in the brain, on the outside (lateral, left) and inside (medial, right) of the brain: ■, high correlation between semantic features and brain activity $s = 6$ s after stimulus presentation; ■, very low correlation; ■, no readings

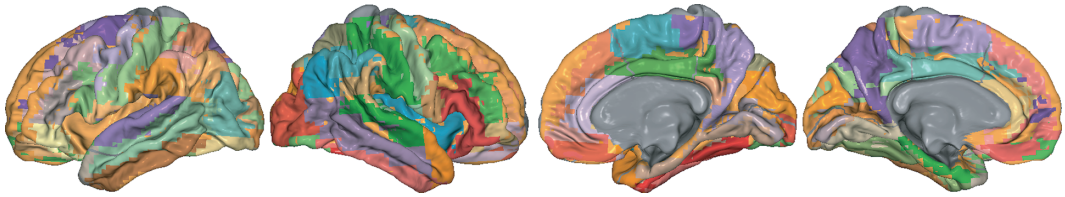

Fig. 7. The 90 ROIs of the brain as used by our experiments, each in a different colour (the colours have no meaning and are purely for easy visualization)

Pycortex software by Gao *et al.* (2015). The natural spatiotemporal correlation in the brain data, along with the 'searchlight' procedure that was used in Wehbe *et al.* (2014), results in a slightly smoothed set of positively correlated p-values, which we assume satisfy PRDS.

7.3. Spatial prior information

Neuroscientists often divide the voxels into ROIs, which are intended to be functionally distinct areas of the brain, like the visual cortex, the hippocampus and the auditory cortex. Fig. 7 shows the 90 ROIs that we use in this paper, marked in different colours for easy visualization. Although the exact number of ROIs and their precise boundaries are still debated, these still provide reasonable prior information for contiguous regions of space where the activity may be correlated with the input stimulus.

7.4. Applying the p-filter

We provide three different non-hierarchically arranged partitions. The first is the trivial finest partition with $41\,073 \times 3$ individual p-values, denoted $P_{v,s}$ as before, for $v = 1, \ldots, 41\,073$ and $s = 4, 6, 8$. The second partition uses temporal information to group $P_{v,4}$, $P_{v,6}$ and $P_{v,8}$ for each v ($41\,073$ many groups). The third partition uses spatial information to group $P_{v,s}$ for all voxels v in the same ROI, for each s (90×3 many groups). We set $\alpha_1 = 0.05$, $\alpha_2 = 0.05$ and $\alpha_3 = 0.1$. For $s = 6$, Fig. 8 displays the rejected p-values in red, and the non-rejected p-values in grey.

The ground truth is, of course, unknown, and this example serves as one possible way to construct layers and to analyse the given brain data. It is now a fairly standard procedure in neuroscience to use the BH procedure (in this case, directly on the input $41\,073 \times 3$ p-values)—recall that this just corresponds to a special case of our p-filter procedure: one that does not explicitly take temporal or spatial structure into account. As mentioned earlier, when used

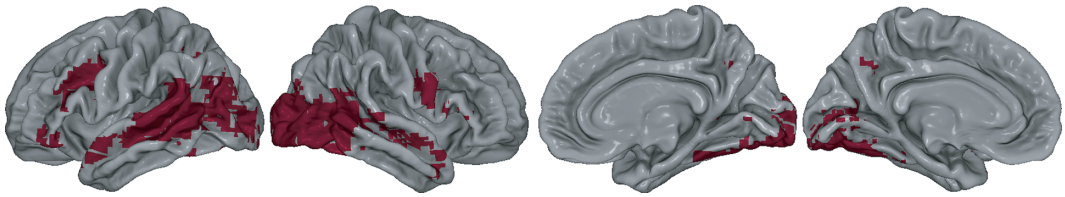

Fig. 8. For time delay $s = 6$ s, we display the final results obtained by the p-filter method: ■, discoveries; ■, non-discoveries

to control the FDR at both the individual voxel and the group levels, our procedure may have lower power than the usual BH procedure, since p-values must pass individual *and* group level constraints; however, as demonstrated in the earlier simulations, these constraints often help to achieve *nearly* the same power but with a sizable reduction in the achieved number of false discoveries, resulting in an improved *precision*. This may allow the scientist possibly to employ higher FDR thresholds, as has been recognized as important for functional MRI data by Lieberman and Cunningham (2009).

8. Conclusion

We introduced an extremely flexible method, the p-filter, that simultaneously controls the FDR across multiple layers (partitions of p-values), which is a guarantee that is significantly more general than existing work. We gave an efficient algorithm for computing the set of discoveries (i.e. rejected p-values), given all the p-values, their various partitions and a target FDR for each partition. We demonstrated its usefulness in simulations—when the pattern of true signals was naturally grouped across rows and columns, we applied the p-filter for entrywise, rowwise and columnwise FDR control, and achieved higher precision, i.e. nearly the same power at lower FDR. We conjecture that this approach may find widespread usage in spatiotemporal or other multimodal applications where p-values can naturally be grouped in many ways across modalities.

Acknowledgements

The authors thank the American Institute of Mathematics's Workshop on 'Inference in high-dimensional regression', where this collaboration started. The authors are also very grateful to Leila Wehbe, who generously shared her time, plotting tools and functional MRI data.

Appendix A: Proofs

A.1. Proof of theorem 3

For each m, by definition of $\hat{t}_m$, there is some $t_1^m, \ldots, t_{m-1}^m, t_{m+1}^m, \ldots, t_M^m$ such that

$$(t_1^m, \ldots, t_{m-1}^m, \hat{t}_m, t_{m+1}^m, \ldots, t_M^m) \in \hat{\mathcal{T}}(\alpha_1, \ldots, \alpha_M). \tag{14}$$

Thus, for each $m' \neq m$, $\hat{t}_{m'} \geqslant t_{m'}^m$ by definition of $\hat{t}_{m'}$. Then

$$\hat{S}(t_1^m, \ldots, t_{m-1}^m, \hat{t}_m, t_{m+1}^m, \ldots, t_M^m) \subseteq \hat{S}(\hat{t}_1, \ldots, \hat{t}_{m-1}, \hat{t}_m, \hat{t}_{m+1}, \ldots, \hat{t}_M),$$

because $\hat{S}(t_1, \ldots, t_M)$ is a non-decreasing function of $(t_1, \ldots, t_M)$. Therefore,

$$\begin{aligned}\widehat{\mathrm{FDP}}_m(\hat{t}_1, \ldots, \hat{t}_{m-1}, \hat{t}_m, \hat{t}_{m+1}, \ldots, \hat{t}_M) &= \frac{G_m \hat{t}_m}{1 \vee |\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_{m-1}, \hat{t}_m, \hat{t}_{m+1}, \ldots, \hat{t}_M)|} \\ &\leqslant \frac{G_m \hat{t}_m}{1 \vee |\hat{S}_m(t_1^m, \ldots, t_{m-1}^m, \hat{t}_m, t_{m+1}^m, \ldots, t_M^m)|} \leqslant \alpha_m,\end{aligned}$$

where the last step holds by definition of $\hat{\mathcal{T}}(\alpha_1, \ldots, \alpha_M)$ and uses expression (14). Since this holds for all m, this proves that $(\hat{t}_1, \ldots, \hat{t}_M) \in \hat{\mathcal{T}}(\alpha_1, \ldots, \alpha_M)$ by definition of $\hat{\mathcal{T}}(\alpha_1, \ldots, \alpha_M)$.

A.2. Proof of theorem 4

Fix any partition m. Since $\Pr(P_i = 0) = 0$ for any $i \in \mathcal{H}^0$ by our assumption (12), we assume that $P_i \neq 0$ for any $i \in \mathcal{H}^0$ without further mention; this assumption then implies that, if $g \in \hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)$ for some null group $g \in \mathcal{H}^0_m$, we must have $\hat{t}_m > 0$. We then calculate

$$\begin{aligned} \mathrm{FDP}_m(\hat{t}_1, \ldots, \hat{t}_M) &= \frac{|\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M) \cap \mathcal{H}^0_m|}{1 \vee |\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)|} = \sum_{g \in \mathcal{H}^0_m} \frac{\mathbf{1}\{g \in \hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)\}}{1 \vee |\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)|} \\ &\leqslant \alpha_m \sum_{g \in \mathcal{H}^0_m} \overbrace{\mathbf{1}\{g \in \hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)\}}^{\hat{t}_m G_m}, \end{aligned} \tag{15}$$

since $\hat{t}_m G_m / \{1 \vee |\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)|\} = \widehat{\mathrm{FDP}}_m(\hat{t}_1, \ldots, \hat{t}_M) \leqslant \alpha_m$ by definition of the method. (The notation $a \vdots b$ is defined in equation (11).)

Now fix any null group $g \in \mathcal{H}^0_m$. Define $\hat{k}^m_g = \hat{k}_{\hat{t}_m}(P_{A^m_g})$, which is the number of rejections when group A^m_g is tested with the BH procedure with threshold $\hat{t}_m$. Then, by definition of $\hat{S}$, if A^m_g is rejected then we must have $\mathrm{Simes}(P_{A^m_g}) \leqslant \hat{t}_m$ and so, as argued in equation (1), A^m_g passes the BH procedure at threshold $\hat{t}_m$, i.e.

$$g \in \hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M) \Rightarrow \hat{k}^m_g > 0,$$

and this can occur only when $\hat{t}_m > 0$ since $P_i \neq 0$ for all $i \in A^m_g \subseteq \mathcal{H}^0$. Furthermore,

$$\mathbf{1}(\hat{k}^m_g > 0) = \overbrace{\mathbf{1}\{\hat{k}^m_g > 0\}}^{\hat{k}^m_g} = \overbrace{\sum_{i \in A^m_g} \mathbf{1}\{P_i \leqslant \hat{t}_m \hat{k}^m_g / |A^m_g|\}}^{\hat{k}^m_g} = \sum_{i \in A^m_g} \overbrace{\mathbf{1}\{P_i \leqslant \hat{t}_m \hat{k}^m_g / |A^m_g|\}}^{\hat{k}^m_g}.$$

Therefore, for each $g \in \mathcal{H}^0_m$, we can write

$$\overbrace{\mathbf{1}\{g \in \hat{S}_m(\hat{t}_1, \ldots, \hat{t}_M)\}}^{\hat{t}_m G_m} \leqslant \overbrace{\mathbf{1}\{\hat{k}^m_g > 0\}}^{\hat{t}_m G_m} = \frac{1}{G_m |A^m_g|} \sum_{i \in A^m_g} \overbrace{\mathbf{1}\{P_i \leqslant \hat{t}_m \hat{k}^m_g / |A^m_g|\}}^{\hat{t}_m \hat{k}^m_g / |A^m_g|}.$$

So, returning to expression (15), we conclude that

$$\mathrm{FDP}_m(\hat{t}_1, \ldots, \hat{t}_M) \leqslant \sum_{g \in \mathcal{H}^0_m} \frac{\alpha_m}{G_m |A^m_g|} \sum_{i \in A^m_g} \overbrace{\mathbf{1}\{P_i \leqslant \hat{t}_m \hat{k}^m_g / |A^m_g|\}}^{\hat{t}_m \hat{k}^m_g / |A^m_g|}.$$

Next, let $f^m_g : [0, 1]^n \to [0, 1]$ be the function that maps P to $\hat{t}_m \hat{k}^m_g / |A^m_g|$. We observe that

(a) $\hat{t}_m$ is a non-increasing function of P by definition of our procedure and
(b) $\hat{k}^m_g$ is also non-increasing in P: if P is lower, then the threshold $\hat{t}_m$ can only rise; lower p-values and a higher (less conservative) threshold can only increase the number of rejections.

Hence f^m_g is a non-increasing function of P. By lemma 1, $\mathbb{E}[\mathbf{1}\{P_i \leqslant f^m_g(P)\} \vdots f^m_g(P)] \leqslant 1$; thus

$$\mathbb{E}[\mathrm{FDP}_m(\hat{t}_1, \ldots, \hat{t}_M)] \leqslant \sum_{g \in \mathcal{H}^0_m} \frac{\alpha_m}{G_m |A^m_g|} \sum_{i \in A^m_g} (1) = \sum_{g \in \mathcal{H}^0_m} \frac{\alpha_m}{G_m} = \alpha_m \frac{|\mathcal{H}^0_m|}{G_m}.$$

A.3. Proof of theorem 5

First we introduce some notation: let $(t^{(k)}_1, \ldots, t^{(k)}_M)$ be the thresholds after the kth pass through the algorithm. We prove that $t^{(k)}_m \geqslant \hat{t}_m$ for all m and k, by induction. At initialization, $t^{(0)}_m = \alpha_m \geqslant \hat{t}_m$ for all m. Now suppose that $t^{(k-1)}_m \geqslant \hat{t}_m$ for all m; we now show that $t^{(k)}_m \geqslant \hat{t}_m$ for all m.

To do this, consider the mth layer of the kth pass through the algorithm. Before this stage, we have thresholds $t^{(k)}_1, \ldots, t^{(k)}_{m-1}, t^{(k-1)}_m, t^{(k-1)}_{m+1}, \ldots, t^{(k-1)}_M$, and we now update $t^{(k)}_m$. Applying induction also to this inner loop, assume that $t^{(k)}_{m'} \geqslant \hat{t}_{m'}$ for all $m' = 1, \ldots, m-1$. We now prove that $t^{(k)}_m \geqslant \hat{t}_m$. By definition,

$$t_m^{(k)} = \max\left\{ T : \frac{G_m T}{1 \vee |\hat{S}_m(t_1^{(k)}, \ldots, t_{m-1}^{(k)}, T, t_{m+1}^{(k-1)}, \ldots, t_M^{(k-1)})|} \leqslant \alpha_m \right\}. \tag{16}$$

Since $t_{m'}^{(k)} \geqslant \hat{t}_{m'}$ for all $m' = 1, \ldots, m-1$, and $t_{m'}^{(k-1)} \geqslant \hat{t}_{m'}$ for all $m' = m+1, \ldots, M$,

$$\frac{G_m \hat{t}_m}{1 \vee |\hat{S}_m(t_1^{(k)}, \ldots, t_{m-1}^{(k)}, \hat{t}_m, t_{m+1}^{(k-1)}, \ldots, t_M^{(k-1)})|} \leqslant \frac{G_m \hat{t}_m}{1 \vee |\hat{S}_m(\hat{t}_1, \ldots, \hat{t}_{m-1}, \hat{t}_m, \hat{t}_{m+1}, \ldots, \hat{t}_M)|}$$

which is less than or equal to α_m by definition of $(\hat{t}_1, \ldots, \hat{t}_M)$. Therefore, $\hat{t}_m$ is in the feasible set for equation (6), and so we must have $t_m^{(k)} \geqslant \hat{t}_m$. By induction this is then true for all k and m.

Now suppose that the algorithm stabilizes at thresholds $(t_1^{(k)}, \ldots, t_M^{(k)})$, after k passes through the algorithm. After completing the mth layer of the last pass through the algorithm, we have thresholds $t_1^{(k)}, \ldots, t_m^{(k)}, t_{m+1}^{(k-1)}, \ldots, t_M^{(k-1)}$; however, since the algorithm stops after the kth pass, this means that $t_{m'}^{(k-1)} = t_{m'}^{(k)}$ for all m'. By definition of $t_m^{(k)}$,

$$\frac{G_m t_m^{(k)}}{1 \vee |\hat{S}_m(t_1^{(k)}, \ldots, t_{m-1}^{(k)}, t_m^{(k)}, t_{m+1}^{(k)}, \ldots, t_M^{(k)})|} \leqslant \alpha_m.$$

This means that $(t_1^{(k)}, \ldots, t_M^{(k)}) \in \hat{\mathcal{T}}(\alpha_1, \ldots, \alpha_M)$, and so $t_m^{(k)} \leqslant \hat{t}_m$ by theorem 3. But, by the work above, we also know that $t_m^{(k)} \geqslant \hat{t}_m$; this proves theorem 5.

A.4. Proof of lemma 1

Fix any $\epsilon > 0$. Recalling that F is the cumulative distribution function of X, we define a sequence $\infty = y_0 > y_1 > y_2 > \ldots$ as follows: for each $i \geqslant 0$ define

$$y_{i+1} := \min\left\{ y : F(y) \geqslant \frac{F_-(y_i)}{1+\epsilon} \right\},$$

where $F_-(y) := \sup\{F(y') : y' < y\} = \Pr(X < y)$. Trivially, $\lim_{i \to \infty} F(y_i) = 0$, so

$$\{y \in \mathbb{R} : F(y) > 0\} = \bigcup_{i \geqslant 0} [y_{i+1}, y_i). \tag{17}$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant Y\}}{F(Y)} \right] &= \mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant Y\}}{F(Y)} \sum_{i \geqslant 0} \mathbf{1}\{y_{i+1} \leqslant Y < y_i\} \right] && \text{by equation (17)} \\
&\leqslant \sum_{i \geqslant 0} \mathbb{E}\left[\frac{\mathbf{1}\{X < y_i\}}{F(y_{i+1})} \mathbf{1}\{y_{i+1} \leqslant Y < y_i\} \right] \\
&\leqslant (1+\epsilon) \sum_{i \geqslant 0} \mathbb{E}\left[\frac{\mathbf{1}\{X < y_i\}}{F_-(y_i)} \mathbf{1}\{y_{i+1} \leqslant Y < y_i\} \right] && \text{by definition of } y_{i+1}.
\end{aligned}$$

Now define the following partial sum for any $n \geqslant m \geqslant 0$:

$$S_{m,n} = \sum_{i=m}^{n} \mathbb{E}\left[\frac{\mathbf{1}\{X < y_i\}}{F_-(y_i)} \mathbf{1}\{y_{i+1} \leqslant Y < y_i\} \right].$$

We claim that

$$S_{m,n} \leqslant \Pr(Y < y_m | X < y_m) \text{ for all } n \geqslant m \geqslant 0. \tag{18}$$

Assuming for the moment that this claim is true, we have

$$\mathbb{E}\left[\frac{\mathbf{1}\{X < Y\}}{F(Y)} \right] \leqslant (1+\epsilon) \sum_{i \geqslant 0} \mathbb{E}\left[\frac{\mathbf{1}\{X < y_i\}}{F_-(y_i)} \mathbf{1}\{y_{i+1} \leqslant Y < y_i\} \right] = (1+\epsilon) \lim_{n \to \infty} S_{0,n},$$

where the limit holds since we have an infinite sum of non-negative terms. Since $\epsilon > 0$ is arbitrarily small and claim (18) implies that $S_{0,n} \leqslant 1$, this proves that

$$\mathbb{E}\left[\frac{\mathbf{1}\{X \leqslant Y\}}{F(Y)}\right] \leqslant 1$$

as desired.

It remains to be shown that claim (18) holds for all $n \geqslant m \geqslant 0$. We prove this for each fixed n by induction over m. Starting with $m = n$, the bound is true trivially. Assuming that it is true for some $m \geqslant 1$, we next prove it with $m-1$ in place of m. We have

$$\begin{aligned}
S_{m-1,n} &= \mathbb{E}\left[\frac{\mathbf{1}\{X < y_{m-1}\}}{F_-(y_{m-1})}\,\mathbf{1}\{y_m \leqslant Y < y_{m-1}\}\right] + S_{m,n} && \text{by definition}\\
&\leqslant \mathbb{E}\left[\frac{\mathbf{1}\{X < y_{m-1}\}}{F_-(y_{m-1})}\,\mathbf{1}\{y_m \leqslant Y < y_{m-1}\}\right] + \Pr(Y < y_m | X < y_m) && \text{by claim (18)}\\
&\leqslant \mathbb{E}\left[\frac{\mathbf{1}\{X < y_{m-1}\}}{F_-(y_{m-1})}\,\mathbf{1}\{y_m \leqslant Y < y_{m-1}\}\right] + \Pr\{Y < y_m | X < y_{m-1}\} && \text{by assumption (12)}\\
&= \Pr(y_m \leqslant Y < y_{m-1} | X < y_{m-1}) + \Pr(Y < y_m | X < y_{m-1})\\
&= \Pr(Y < y_{m-1} | X < y_{m-1}),
\end{aligned}$$

proving that claim (18) holds with $m-1$ in place of m. This concludes the proof.

References

Benjamini, Y. and Bogomolov, M. (2014) Selective inference on multiple families of hypotheses. *J. R. Statist. Soc.* B, **76**, 297–318.

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 Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Chouldechova, A. (2014) False discovery rate control for spatial data. *PhD Thesis*. Stanford University, Stanford.

Gao, J. S., Huth, A. G., Lescroart, M. D. and Gallant, J. L. (2015) Pycortex: an interactive surface visualizer for fMRI. *Front. Neurinform.*, **9**, 23.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, **15**, 870–878.

Hu, J. X., Zhao, H. and Zhou, H. H. (2010) False discovery rate control with groups. *J. Am. Statist. Ass.*, **105**, 1215–1227.

Lieberman, M. and Cunningham, W. (2009) Type I and Type II error concerns in fMRI research: re-balancing the scale. *Soc. Cogn. Affect. Neursci.*, **4**, 423–428.

Meinshausen, N. (2008) Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.

Peterson, C. B., Bogomolov, M., Benjamini, Y. and Sabatti, C. (2016) Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies. *Genet. Epidem.*, **40**, 45–56.

R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Simes, J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

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.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A. and Mitchell, T. (2014) Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS One*, **9**, no. 11, article e112575.

Wehbe, L., Ramdas, A., Steorts, R. and Shalizi, C. (2015) Regularized brain reading with shrinkage and smoothing. *Ann. Appl. Statist.*, **9**, 1997–2022.

Yekutieli, D. (2008) Hierarchical false discovery rate-controlling methodology. *J. Am. Statist. Ass.*, **103**, 309–316.